

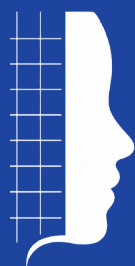
LNAI 16749

Pavel Andreev
William Van Woensel
John Holmes
Antoine Sauré (Eds.)

Artificial Intelligence in Medicine

24th International Conference, AIME 2026
Ottawa, ON, Canada, July 7–10, 2026
Proceedings, Part II

2 Part II



AIME 2026

24th International Conference
on Artificial Intelligence in Medicine
OTTAWA, CANADA

 Springer

Lecture Notes in Computer Science

Lecture Notes in Artificial Intelligence

16749

Founding Editor

Jörg Siekmann

Series Editors

Randy Goebel, *University of Alberta, Edmonton, Canada*

Wolfgang Wahlster, *DFKI, Berlin, Germany*

Zhi-Hua Zhou, *Nanjing University, Nanjing, China*

The series Lecture Notes in Artificial Intelligence (LNAI) was established in 1988 as a topical subseries of LNCS devoted to artificial intelligence.

The series publishes state-of-the-art research results at a high level. As with the LNCS mother series, the mission of the series is to serve the international R & D community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings.

Pavel Andreev · William Van Woensel ·
John Holmes · Antoine Sauré
Editors

Artificial Intelligence in Medicine

24th International Conference, AIME 2026
Ottawa, ON, Canada, July 7–10, 2026
Proceedings, Part II

Editors

Pavel Andreev 
University of Ottawa
Ottawa, ON, Canada

William Van Woensel 
University of Ottawa
Ottawa, ON, Canada

John Holmes 
University of Pennsylvania
Philadelphia, PA, USA

Antoine Sauré 
University of Ottawa
Ottawa, ON, Canada

ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Artificial Intelligence

ISBN 978-3-032-30812-2

ISBN 978-3-032-30813-9 (eBook)

<https://doi.org/10.1007/978-3-032-30813-9>

LNCS Sublibrary: SL7 – Artificial Intelligence

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2027

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

AIME 2026 arrived at a pivotal moment—as artificial intelligence in medicine was moving from a research frontier to a clinical reality, bringing extraordinary promise but also profound responsibility. It is our honor to welcome readers to the proceedings of the 24th International Conference on Artificial Intelligence in Medicine, held at the University of Ottawa, Canada, from July 7 to 10, 2026.

Since its first edition in Pavia in 1985, AIME has provided an international forum for researchers working at the intersection of artificial intelligence, biomedicine, and healthcare. Over the last four decades, the field has moved from symbolic and knowledge-based systems to machine learning, deep learning, multimodal modeling, large language models, and increasingly agentic AI systems. Yet the central concern of the AIME community has remained consistent: how to design, develop, evaluate, and responsibly use AI to better support clinical decisions, improve health systems, and provide better care for patients.

AIME 2026 took place in Ottawa—Canada’s capital city and home to federal health agencies, national research councils, research hospitals and institutes, and a vibrant technology sector. The conference was co-hosted by the Telfer School of Management and the Faculty of Medicine at the University of Ottawa—one of Canada’s leading research universities and the world’s largest bilingual university. uOttawa is a core partner of the Ottawa Academic Health Network (OAHN) linking it with affiliated hospitals and research institutes. Telfer is also one of the few business schools in North America where health research is a core strategic pillar, training professionals to navigate clinical decision making, healthcare delivery, and policy—precisely the challenges at the heart of AIME.

With 236 submissions from 53 countries, AIME 2026 became a record-setting North American edition and one of the most internationally diverse in the conference’s history. Submissions originated from Asia, Europe, North America, South America, Africa, and Oceania, demonstrating the growing global reach of the AIME community and the increasing importance of AI across health systems.

For AIME 2026, the conference themes were organized into two broad areas: *AI Foundations and Methodological Innovations* and *Applications, Systems, and Health Impact*. Many submissions targeted subthemes of *Language, Imaging, and Signal Intelligence*, followed by *Clinical AI and Decision Support Systems*; *Deployment, Evaluation, and Governance*; and *Patient-Centered and Population-Level Applications*. The volume and distribution of submissions covered the full spectrum of the field: clinical decision support, triage, and precision medicine; human-AI collaboration, trust, safety, and ethics; biomedical imaging and signal processing; and multimodal learning, generative AI, and foundation models.

To accommodate the record number of submissions spanning such a broad spectrum of topics, AIME 2026 introduced several changes to the review process. Reviewers indicated their expertise, and authors rated their papers against, the conference themes,

improving reviewer matching; Senior Program Committee (PC) members took greater responsibility in overseeing the process. We followed a single-anonymous peer review process. Most papers received three reviews from the regular PC and one review from the Senior PC, who synthesized the reviews and provided a final recommendation. Of the 221 papers sent for peer review, 30 full papers (20% acceptance rate), 21 short papers (18% acceptance rate), and 82 poster and demo papers were accepted. Authors of full and short papers were invited to present their work orally; poster and demo papers were presented across six poster sessions. To increase their visibility and accessibility, AIME 2026 implemented fully digital poster presentations, including 60-second audio summaries. A final novelty was the introduction of a Journal-First track with presentations for recent journal papers, which exposed the AIME community to state-of-the-art results while also giving the authors broader visibility.

We had the privilege of hosting two invited keynote speakers: Tal Arbel of McGill University, whose address on equitable, personalized, and image-based medicine emphasized that clinical AI must be trustworthy before it can be useful; and Muhammad Mamdani of Ontario Health and the University of Toronto, who challenged the community to close the gap between algorithmic performance and real-world impact. Additionally, a panel discussion brought together clinicians, researchers, and policymakers to address a central challenge: that the successful application of AI in medicine depends not only on model performance, but also on clinical acceptability, professional accountability, workflow integration, patient trust, and institutional governance.

The tutorial and workshop programs complemented the main conference and covered interpretable, responsible, trustworthy, and privacy-preserving AI; imaging and signal-based AI methods; AI applications in nursing, oncology, and public health; and the practical evaluation and implementation of AI. The Doctoral Consortium continued AIME's commitment to doctoral training by providing feedback to six PhD students on their research programs in an informal setting.

We sincerely thank everyone who made AIME 2026 possible: the authors, the Program Committee and Senior Program Committee members, keynote speakers, panelists, tutorial and workshop organizers, Doctoral Consortium panelists, sponsors, institutional partners, volunteers, and the local organizing team.

The question before us is no longer focused on whether AI is able to perform a task, but also how it should be evaluated, deployed, and governed, and how this can meaningfully improve care. We hope that AIME 2026 contributed to these conversations and helped move AI in medicine from promising innovation toward responsible, trustworthy, and clinically meaningful practice.

The AIME community has always believed that better science leads to better care. We left Ottawa more convinced than ever that our work matters, and that how we do it matters just as much as what we do.

May 2026

Pavel Andreev
William Van Woensel
John Holmes
Antoine Sauré

Organization

General Chair

Pavel Andreev	University of Ottawa, Canada
---------------	------------------------------

Scientific Chairs

William Van Woensel	University of Ottawa, Canada
John Holmes	University of Pennsylvania, USA

Local Organization Chair

Antoine Sauré	University of Ottawa, Canada
---------------	------------------------------

Program Committee Chairs

Enea Parimbelli (<i>Doctoral Consortium</i>)	Università di Pavia, Italy
Daniel Amyot (<i>Tutorials</i>)	University of Ottawa, Canada
Christopher Sun (<i>Workshops</i>)	University of Ottawa, Canada

Senior Program Committee

Syed Sibte Raza Abidi	Dalhousie University, Canada
Riccardo Bellazzi	Università di Pavia, Italy
Jaime Cardoso	University of Porto, Portugal
Carlo Combi	Università degli Studi di Verona, Italy
Arianna Dagliati	University of Pavia, Italy
Kerstin Denecke	Bern University of Applied Sciences, Switzerland
Barbara Di Camillo	University of Padua, Italy
Michel Dojat	INSERM, France
Peter Elkin	Ontolimatics, USA
Joseph Finkelstein	University of Utah, USA
Graciela Gonzalez-Hernandez	Cedars-Sinai Medical Center, USA

Milos Hauskrecht	University of Pittsburgh, USA
Pedro Henriques Abreu	University of Coimbra, Portugal
Zhengxing Huang	Zhejiang University, China
Jose Juarez	Universidad de Murcia, Spain
Elpida Keravnou-Papailiou	University of Cyprus, Cyprus
Nan Liu	Duke-NUS Medical School, Singapore
Peter Lucas	University of Twente, Netherlands
Mar Marcos	Universitat Jaume I, Spain
Martin Michalowski	University of Minnesota, USA
Wojtek Michalowski	University of Ottawa, Canada
Stefania Montani	University of Piemonte Orientale, Italy
Robert Moskovitch	Ben-Gurion University of the Negev, Israel
Barbara Oliboni	University of Verona, Italy
Panagiotis Papapetrou	Stockholm University, Sweden
Enea Parimbelli	University of Pavia, Italy
Silvana Quaglini	University of Pavia, Italy
Lucia Sacchi	University of Pavia, Italy
Aleksander Sadikov	University of Ljubljana, Slovenia
Brigitte Seroussi	Assistance Publique – Hôpitaux de Paris, France
Yuval Shahar	Ben-Gurion University of the Negev, Israel
Li Shen	University of Pennsylvania, USA
Myra Spiliopoulou	Otto-von-Guericke-University Magdeburg, Germany
Fabio Stella	University of Milano-Bicocca, Italy
Gregor Stiglic	University of Maribor, Slovenia
Stephen Swift	Brunel University of London, UK
Annette Ten Teije	Vrije Universiteit Amsterdam, Netherlands
Paolo Terenziani	Università del Piemonte Orientale, Italy
Allan Tucker	Brunel University of London, UK
Szymon Wilk	Poznań University of Technology, Poland
Migao Wu	Carleton University, Canada
Blaz Zupan	University of Ljubljana, Slovenia
Frank van Harmelen	Vrije Universiteit Amsterdam, Netherlands

Program Committee

Samina Abidi	Dalhousie University, Canada
Syed Sibte Raza Abidi	Dalhousie University, Canada
Giuseppe Agapito	Università Magna Graecia Catanzaro, Italy
Alexander Alekseyenko	Medical University of South Carolina, USA
Rute Almeida	CINTESIS, Portugal

Amparo Alonso-Betanzos	University of A Coruña, Spain
Sanja Avramovic	George Mason University, USA
Hamed Ayoobi	University of Groningen, Netherlands
Luís Azevedo	University of Porto, Portugal
Hajer Baazaoui	ETIS, CY Cergy Paris University, France
Arriel Benis	Holon Institute of Technology, Israel
Balu Bhasuran	Florida State University, USA
Isabelle Bichindaritz	State University of New York at Oswego, USA
Alessio Bottrighi	Università del Piemonte Orientale, Italy
Dariusz Brzeziński	Poznań University of Technology, Poland
Marcos L. P. Bueno	Radboud University, Netherlands
Jorie Butler	University of Utah, USA
Jerzy Błaszczyński	Poznań University of Technology, Poland
David Calhas	University of Ljubljana, Slovenia
Francesco Calimeri	University of Calabria, Italy
Manuel Campos	University of Murcia, Spain
Ricardo Cardoso Pereira	University of Coimbra, Portugal
Rita Casadio	University of Bologna, Italy
Chuming Chen	University of Delaware, USA
Simone Ciccolella	Università degli studi di Milano-Bicocca, Italy
Kathleen Curran	University College Dublin, Ireland
Bernardo Cánovas Segura	University of Murcia, Spain
Gayo Diallo	University of Bordeaux, France
Georg Dorffner	Medical University Vienna, Austria
Mauro Dragoni	Fondazione Bruno Kessler, Italy
Adam Dunn	University of Sydney, Australia
Inês Dutra	University of Porto, Portugal
Lauren Erdman	University of Toronto/Hospital for Sick Children, Canada
Bjoern Eskofier	FAU Erlangen, Germany
Brígida Mónica Faria	Polytechnic Institute of Porto, Portugal
Jesualdo Tomás Fernández-Breis	Universidad de Murcia, Spain
Daniela Ferreira-Santos	University of Porto, Portugal
Joseph Finkelstein	University of Arizona, USA
Pedro Furtado	University of Coimbra/CISUC, Portugal
Roberto Gatta	Università Cattolica del Sacro Cuore, Italy
Ben Glicksberg	University of California, San Francisco, USA
Natalia Grabar	STL CNRS Université Lille 3, France
Vida Groznik	University of Ljubljana, Slovenia
Bertha Guijarro-Berdiñas	University of A Coruña, Spain
Milos Hauskrecht	University of Pittsburgh, USA

Morihiro Hayashida	National Institute of Technology, Matsue College, Japan
Zhe He	Florida State University, USA
Shenda Hong	Peking University, China
Zhengxing Huang	Zhejiang University, China
Xiayuan Huang	Yale University, USA
Kyu-Baek Hwang	Soongsil University, South Korea
Nevo Itzhak	Ben-Gurion University of the Negev, Israel
Filip Jagodzinski	Western Washington University, USA
Milica M. Jankovic	University of Belgrade, Serbia
Xiaoqian Jiang	UTHealth at Houston, USA
Jose Juarez	Universidad de Murcia, Spain
Rafał Józwiak	Warsaw University of Technology, Poland
Charles Kahn	University of Pennsylvania, USA
Eleni Kaldoudi	Democritus University of Thrace, Greece
Evangelia Katsoulakis	University of South Florida, USA
Sophie Kearney	University of Pennsylvania, USA
Denisse Kim	University of Murcia, Spain
Dokyoon Kim	University of Pennsylvania, USA
Jaebum Kim	Konkuk University, South Korea
Primoz Kocbek	University of Maribor, Slovenia
Haridimos Kondylakis	Institute of Computer Science, FORTH, Greece
Leon Kopitar	University of Maribor, Slovenia
Krzysztof Krawiec	Poznań University of Technology, Poland
Craig Kuziemsky	MacEwan University, Canada
Himabindu Lakkaraju	Harvard University, USA
Pedro Larranaga	University of Madrid, Spain
Nguyen Quoc Khanh Le	Nanyang Technological University, Singapore
Carson Leung	University of Manitoba, Canada
Chung-Sheng Li	PwC, USA
Michael Liebman	IPQ Analytics. LLC, USA
Aneta Lisowska	Vrije Univeriteit Amsterdam, Poland
Mei Liu	University of Florida, USA
Hao Liu	Montclair State University, USA
Enrico Longato	University of Padua, Italy
Beatriz Lopez	University of Girona, Spain
Antonio Lopez-Martinez-Carrasco	University of Murcia, Spain
Gang Luo	University of Washington, USA
Yves Lussier	University of Utah, USA
Hiroshi Mamiya	McGill Clinical and Health Informatics, Canada
Simone Marini	University of Florida, USA
Begoña Martinez-Salvador	Universitat Jaume I, Spain

Stan Matwin	Dalhousie University, Canada
Eneida Mendonca	Cincinnati Children's Hospital Medical Center, USA
Ana Maria Mendonça	INESC TEC, Portugal
Martin Michalowski	University of Minnesota, USA
Silvia Miksch	Vienna University of Technology, Austria
Diego Molla	Macquarie University, Australia
Irina Moreira	Center for Neuroscience and Cell Biology, Portugal
Robert Moskovitch	Ben-Gurion University of the Negev, Israel
Laura Moss	University of Glasgow, UK
Fleur Mougín	AHeaD, INSERM U1219 – Université de Bordeaux, France
Kenta Nakai	University of Tokyo, Japan
Loris Nanni	University of Padua, Italy
Ardalan Naseri	University of Texas Health Science Center at Houston, USA
Goran Nenadic	University of Manchester, UK
Denis Newman-Griffis	University of Sheffield, UK
Shu-Kay Ng	Griffith University, Australia
Tram Anh Nguyen	University of Pennsylvania, USA
Giovanna Nicora	University of Pavia, Italy
Anne Torill Nordsletta	Norwegian Centre for e-health Research, Norway
Arnau Oliver	University of Girona, Spain
Daniele Pala	University of Pavia, Italy
Laura Maria Peltonen	University of Turku and Turku University Hospital, Finland
Samantha Piekos	University of Pennsylvania, USA
Maciej Piernik	Poznań University of Technology, Poland
Luca Piovesan	Università del Piemonte Orientale, Italy
Ronald Piscotty	Oakland University, USA
Jędrzej Potoniec	Poznań University of Technology, Poland
Mattia Proserpi	University of Florida, USA
Lisiane Pruinelli	University of Florida, USA
Nisha Puthiyedth	Thompson Rivers University, Canada
Arya Rahgozar	University of Ottawa, Canada
Nadav Rappoport	Ben-Gurion University of the Negev, Israel
Je-Keun Rhee	Soongsil University, South Korea
Kirk Roberts	University of Texas Health Science Center at Houston, USA
Joseph Romano	University of Pennsylvania, USA
Simona E. Rombo	University of Palermo, Italy

Abdul Roudsari	University of Victoria, Canada
Luis Rueda	University of Windsor, Canada
Elisa Salvi	Norwegian Centre for e-health Research, Norway
Leonardo Sanna	Fondazione Bruno Kessler – FBK, Italy
Joana Cristo Santos	University of Coimbra, Portugal
Neil Sarkar	Brown University, USA
Abeed Sarker	Emory University School of Medicine, USA
Isabel Sassoon	Brunel University of London, UK
Michael Ignaz Schumacher	University of Applied Sciences Western Switzerland (HES-SO), Switzerland
Floriano Scioscia	Polytechnic University of Bari, Italy
Joao Setubal	University of São Paulo, Brazil
Arash Shaban-Nejad	University of California, Berkeley, USA
Wenqi Shi	UT Southwestern Medical Center, USA
Justyna Skibińska	University of Ljubljana, Slovenia
Neil Smalheiser	University of Illinois Chicago, USA
Wenyu Song	Brigham and Women's Hospital, USA
Giltae Song	Pusan National University, South Korea
Marco Spruit	Leiden University Medical Centre, Netherlands
Manuel Striani	University of Piemonte Orientale, Italy
Madhumita Sushil	University of California, San Francisco, USA
Baris Suzek	Muğla Sıtkı Koçman University, Turkey
Sing-Hoi Sze	Texas A&M University, USA
Erica Tavazzi	University of Padua, Italy
César Teixeira	University of Coimbra, Portugal
Shubo Tian	National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, USA
Max Topaz	University of Columbia, USA
Ryan Urbanowicz	University of Pennsylvania, USA
Iacopo Vagliano	Amsterdam Universitair Medische Centra, Netherlands
Ágnes Vathy-Fogarassy	University of Pannonia, Hungary
Alfredo Vellido	Universitat Politècnica de Catalunya, Spain
Pierangelo Veltri	Università della Calabria, Italy
Maria-Esther Vidal	TIB–Leibniz Information Centre for Science and Technology, Germany
Giorgio Vinciguerra	University of Pisa, Italy
Francesca Vitali	University of Arizona, USA
Dimitrios Vogiatzis	ACG Deree, NCSR “Demokritos”, Greece
Juexin Wang	Indiana University Indianapolis, USA
Junbai Wang	Radium Hospital, Norway

Jens Weber	University of Victoria, Canada
Chunhua Weng	Columbia University, USA
Szymon Wilk	Poznań University of Technology, Poland
Huanmei Wu	Temple University, USA
Sejung Yang	Yonsei University, South Korea
Ye Ye	University of Pittsburgh, USA
Rui Zhang	University of Minnesota, USA
W. Jim Zheng	U Texas Health Science Houston, USA
Pierre Zweigenbaum	Université Paris-Saclay, CNRS, LISN, France
Paweł Śniatała	Poznań University of Technology, Poland
Martin Špendl	University of Ljubljana, Slovenia

Sponsors

We would like to thank the following organizations for sponsoring the conference:

Platinum Sponsors



Institut du Savoir Montfort (ISM)



Réseau universitaire de la santé d'Ottawa
Ottawa Academic Health Network

Ottawa Academic Health Network (OAHN)



Telfer School of Management



Faculty of Medicine

Keynote Sponsor



CarMa Global Enterprise

Silver Sponsors



UNIVERSITY OF OTTAWA
HEART INSTITUTE
INSTITUT DE CARDIOLOGIE
DE L'UNIVERSITÉ D'OTTAWA

The University of Ottawa Heart Institute

Other Sponsors and Partners

Springer



Artificial Intelligence journal



uOttawa INNOVA

Contents

Building Trustworthy Artificial Intelligence Enabled Electronic Health Record Systems	1
<i>Nabil Georges Badr and Hazar Haidar</i>	
Leadership, Cultural Intelligence, and AI-Assisted Governance: A Compliance Perspective on the Pharmaceutical and Medical Device Industries	7
<i>Ali Abou Abbas and Arwa Asiri</i>	
A PRISMA-Aligned Workflow for LLM-Assisted Title–Abstract Screening: A Methodological Case Study in Atrial Fibrillation Literature	13
<i>Olena Holub, Alex Nakonechnyi, Joshua Metcalf, and Rebecca Allen</i>	
Reasoning-Enhanced Rare-Event Prediction with Balanced Outcome Correction for Clinical Decision Support	18
<i>Vitaly Bulgakov and Alexander Turchin</i>	
Explainable AI-Driven Optimized LSTM for Heart Sound Classification and Diagnostic Support	24
<i>Faiq Ahmad Khan, Arshad Hassan, Lia Anwar, and Amine Bermak</i>	
VAIDYA: Validated Agents for Intelligent Diagnosis and Yielded Analysis	29
<i>Kalash Shah, Gautam Bhutani, Rohitaswa Sarbhangia, and J. Snehan</i>	
Redefining Cardiac Arrhythmia Detection and Monitoring in Athletes Using Artificial Intelligence: A Systematic Review	35
<i>Rayyan Tamim, Danielle Cruise, and Mirou Jaana</i>	
Within-Phenotype Socioeconomic Disparities in Cancer Detection: A Topological Data Analysis of a Population-Based Cohort	40
<i>Nelofar Kureshi, Samina Abidi, Ellen Sweeney, and Syed Sibte Raza Abidi</i>	
Synthetic Validation of Pediatric Trust Instruments Using Persona-Driven Large Language Models	45
<i>Ella Boone, Katya Loban, Elena Guadagno, and Dan Poenaru</i>	
Supporting Autonomy in Dementia Patients with a Cognitive Assistant	50
<i>Felix Ocker, Steffen Heinrich, Alexa von Bosse, Jörg Deigmöller, Pavel Smirnov, and Julian Eggert</i>	

Bidirectional Floating Feature Selection Guided by Uncertainty Quantification 56
Marcos López-De-Castro, José González-Gomariz, Alberto García-Galindo, Kewei Ni, Farnoosh Abbas-Aghababazadeh, Benjamin Haibe-Kains, and Ruben Armañanzas

WearAble AI: AI Assistant for Wearable Data Analysis and Visualization 62
Arash Atrsaei, Haneen Njoun, Albert Pla Planas, and Sven Jager

A Phyto-Ontology for Integrative Therapeutic Discovery 68
Oresta S. I. Hewryk and Joseph D. Romano

Evaluating Temporal Clinical Reasoning and Audience-Tailored Explanations in Large Language Models 73
Gianluca Apriceno, Tania Bailoni, and Mauro Dragoni

Seeing the Wood for the Trees: Rethinking AI Ethics Beyond Anthropocentrism 79
Bianca Lerma, Gianluca Apriceno, and Mauro Dragoni

Bridging Innovation and Regulation: Ethical Governance of Medical AI in Low- and Middle-Income Countries 85
Dalson Nyangoma and Gerald Mboowa

A Multimodal-Based Framework for Smart ECG Report Generation 90
Hong Duc Nguyen, Duc Tri Phan, and S. Supraja

Same Models, Better Results: Improving Multimodal LLMs’ Accuracy in Identifying Myocardial Infarction from ECG Images 96
Japp Adhikari, Keyuan Jiang, Hania Farhan, and Gordon R. Bernard

Transforming Hospitals with Artificial Intelligence: Applications in Clinical Education and Patient Care 101
Nihal Haque, Ervin Sejdic, Michael Liut, Roland Mollanji, Phil Shin, Duska Kennedy, and Ghadir Ali

Network-Based Integration of Multi-omics Data for Biomarker Discovery in Acute Coronary Syndromes 106
Elena Marinello, Shahryar Noei, Marco Chierici, Núria Amigó, Dane Cvijanović, Aleksandar Davidović, Dalibor Dragišić, Claudia Gíglione, Dimitris Kardassis, Marko Milanov, Jelena Munjas, Iva Perović-Blagojević, Aleksa Petković, Tamara Ratković, Pol Torne, Christos Tsatsanis, Sandra Vladimirov-Sopić, Luka Vukmirović, Paolo Magni, Miron Sopić, Barbara Di Camillo, and Giuseppe Jurman

Automated ICF Profiling from ICU Time Series Using Transformer-Based Physiological Prediction	111
<i>Feten Skhiri Gabbouj, Mathias Blandeau, Matei Mancas, Bernard Gosselin, and Laura Wallard</i>	
InstructFusionFormer: Instruction-Conditioned Multimodal Fusion for Esophageal Disease Diagnosis and Staging	116
<i>Zenebe Markos Lonseko, Helen Haile Hayeso, Tao Gan, and Nini Rao</i>	
Automated Behavioral Health Assessment Using Vocal Biomarkers	121
<i>Namhee Kwon, Nate Blaylock, Raymond Brueckner, Vinod Subramanian, and Henry O’Connell</i>	
Explainable Knowledge Graph Embeddings for Link Prediction and Path Attribution in Healthcare	127
<i>Vijaya Laxmi Boddu, Akhil Chaudhary, Somayeh Kafaie, and Enayat Rajabi</i>	
LLM-as-a-Judge for Evaluating the Quality of Retrieval-Augmented Generation Systems	133
<i>Leonardo Sanna, Erica Solinas, and Mauro Dragoni</i>	
Flow Builder: No-Code Conversation Design Tool for Digital Therapeutics in Psychology	139
<i>Leonardo Sanna, Mattia Franzin, Mauro Dragoni, and Claudio Eccher</i>	
MAESTRO: a Framework for Trustworthy Integration of LLMs in Psychological Digital Interventions	145
<i>Leonardo Sanna, Mattia Franzin, Simone De Carli, Marco Bolpagni, Simone Casazza, Silvia Rizzi, Claudio Eccher, and Mauro Dragoni</i>	
Inferring Chronic Treatment Onset from ePrescription Data: A Renewal Process Approach	151
<i>Pavlin G. Poličar, Dalibor Stanimirović, and Blaž Zupan</i>	
What Does It Mean for LLMs to Be Empathetic? A Co-design Study of Patient Trust in LLM-Based Chronic Care for Heart Failure	156
<i>Xinyi Celine Liu, Pedro Elkind Velmovitsky, Dina Kanaan, Raima Lohani, Vina Mohabir, David Wiljer, Joseph Cafazzo, Shaina Raza, Broderick Bellard, Shumit Saha, Myrte de Alfred, and Quynh Pham</i>	

Predicting Critical Deterioration of Patients in Emergency Units Using Administrative Health Data	161
<i>Clément Lens, Bilal Majed, Pierre Marquis, Karim Tabia, and Romain Wallon</i>	
Comparing the Effects of Persistence Barcodes Aggregation and Feature Concatenation on Medical Imaging	167
<i>Dashti A. Ali, Richard K. G. Do, William R. Jarnagin, Aras T. Asaad, and Amber L. Simpson</i>	
Electroencephalogram (EEG) Subband Features During Eye Opening and Closure Enable Classification of Amblyopes and Visually Normal Controls: A Pilot Study	172
<i>Gabriel W. Fernandes, Lachlan D. Barnes, Jerry J. Zhang, and Luke E. Hallum</i>	
A Critical Survey on the Evaluation of Large Language Models in Clinical Applications	177
<i>Yasmin Saafi, Nizar Omheni, Zayneb Mannai, and Ramzi Mahmoudi</i>	
Multiresolution Fusion Network with Attention for Brain and Breast Cancer Detection	182
<i>Vagish Kumar and Souvik Chakraborty</i>	
Fairness-Aware Time Series Decision Support for Virtual Ward Triage	187
<i>Basra Jehangir, Weizi Vicky Li, Dionne Tannetta, and Andrew Walden</i>	
When More Data is Not Better: Physiological Drift Explains Transfer Learning Failure in Rare ICU Populations	193
<i>Michael Poette, Sandrine Mouysset, Daniel Ruiz, Jean-Marc Alliot, Fanny Vardon, and Vincent Minville</i>	
Informative Missingness to Generate Irregular Clinical Time Series	198
<i>Hadi Mehdizavareh, Gabriele Santangelo, Giovanna Nicora, Simon Lebech Cichosz, Arianna Dagliati, Arijit Khan, and Riccardo Bellazzi</i>	
Logic-Guided Interpretable Hospital Readmission Risk Modeling Using Italian Administrative Healthcare Data	204
<i>Marina Andrić, Gianluca Apriceno, Kevin Gelmini, and Mauro Dragoni</i>	
DST-GAN: Dempster-Shafer Based Fusion for Multimodal Oversampling of Healthcare Data	210
<i>Kevin Fee, Mohammed Hasanuzzaman, Suneil Jain, Ross G. Murphy, and Anna Jurek-Loughrey</i>	

A Structured Residual Refinement Transformer Framework for Long-Horizon Healthcare Demand Forecasting with Causal and Seasonal Signals	216
<i>Guillem Hernández Guillamet, Joan Samper Vila, and Beatriz López Ibáñez</i>	
A Deep Learning Framework for Vision-Based Intraocular Pressure Monitoring Using Microfluidic Contact Lenses	222
<i>Justin Jacob, Tanzila Afrin, Yong Jun Lai, and Michael Greenspan</i>	
Effectiveness of Artificial Intelligence in Cardiovascular Disease Prevention: A Systematic Review	228
<i>Chetan Shirvankar, Lily Puterman-Salzman, Josep Iglesias-Grau, Hervé Tchala Vignon Zomahoun, and Samira Abbasgholizadeh Rahimi</i>	
Rehabilitation Engagement Prediction in Pediatric Patients with Neuromotor Disorders: An Adaptive Hierarchical Bayesian Approach	233
<i>Simone Costantini, Emilia Biffi, Giuseppe Andreoni, Anna Bianchi, and Fabio Storm</i>	
Explainable AI Identifies High Retrospective Prodromal Symptom Burden in Parkinson’s Disease	239
<i>Dante Trabassi, Stefano Filippo Castiglia, Roberto De Icco, Cristina Tassorelli, Alberto Ranavolo, and Mariano Serrao</i>	
A Bayesian Latent-Threshold Framework for Fall Occurrence in Parkinson’s Disease	244
<i>Dante Trabassi, Stefano Filippo Castiglia, Roberto De Icco, Cristina Tassorelli, Alberto Ranavolo, and Mariano Serrao</i>	
Ontology-Aligned Clinical NER in Emergency Medicine Using Weak Supervision	249
<i>Amila Kugic, Adrian Schiegl, Alistair Tiefenbacher, Lukas Seper, Stefan Schulz, and Markus Kreuzthaler</i>	
Biological Plausibility Assessment of Viral Sequences Generated by a Genomic Language Model	254
<i>Pablo Arozarena Donelli, Simone Rancati, Giovanna Nicora, Riccardo Bellazzi, Enea Parimbelli, and Luigi Portinale</i>	
Adaptive Reinforcement Learning for Quality Assurance in Clinical Assessment: A Multi-agent Framework	259
<i>Varun Gazala and Sayda Elmi</i>	

Artificial Intelligence Based Classification of Oral Cancer
and Pre-malignant Disorders from Electrical Impedance Spectroscopy 265
*Mary Paterson, Craig Murdoch, Dawn Walker, Helen Colley,
Malwina Matella, Rachel Furmidge, and Zi-Qiang Lang*

Vision-Based Evaluation of Standardized Balance Tasks for Clinical
and Remote Assessment 271
*Venkata Sai Pavan Raghavendra Paravatham, Sayda Elmi,
Charitha Chinnapapannagari, and Morris Bell*

Machine Learning for Early Prediction of Gestational Diabetes Mellitus
Using Routinely Collected Clinical Data 276
*Syeda Rizvi, Mais Alkhateeb, Shehel Yoosuf, Aliya Tabassum,
Arfan Ahmed, and Alaa Abd-Alrazaq*

Bridging the Pathology Communication Gap: Early Experience with Osler,
a Patient-Facing AI Chatbot 281
*Morgan Black, Will Dowon Jeong, Sherman Lin, Ali Brown,
Matthew Cecchini, and Jason K. Wasserman*

Graph Attention Networks for Donor-Recipient Matching in Kidney
Transplantation 286
*Sheida Majouni, Karthik Tennankore, Samina Abidi, Amanda Vinson,
and Syed Sibte Raza Abidi*

Multimodal Survival Prediction of Patients with Metastatic
Esophagogastric Cancer Based on Clinical Progress Reports and Tabular
Data 292
*Irene Cara, Dimitris Katsimpokis, Gijs Geleijnse,
Hanneke W. M. van Laarhoven, and Rob H. A. Verhoeven*

Clinical Deployability: A Socio-technical Construct for Evaluating
Real-World Readiness of Medical AI 297
Federico Cabitza and Mauro Dragoni

A Knowledge Acquisition Platform and Sharing Library for the TMR
Computable Guideline Model 303
Saif Latifi and Vasa Curcin

AI-Informed 30-Day MACE Risk Predictions in ED Arrivals 308
*Ben J. Gross, M.D. Radha Patel, Kai Zheng, Shamim Nemati,
and M.D. Matheus Ramsis*

Towards Predicting Sedation Depth in Endoscopy with Large Clinically Annotated EEG Data of Continuous Propofol Sedation	313
<i>Jakob Garbe, Quang Vu Nguyen, Jan W. Kantelhardt, Florian Dünninghaus, Karla Erffmeier, Katja Seeliger, and Thomas Schmid</i>	
Establishing a Data-Driven Upper Bound for CAR-T Prognosis: Quantifying the Impact of Heuristic Discretization	319
<i>Francesco Olivato, Mirko Farina, Michele Malagola, Domenico Russo, and Roberto Gatta</i>	
PreClaim-GCT: Self-supervised Graph–Transformer Pretraining for Incident Outcome Prediction from Healthcare Claims	324
<i>Rui Wang, Ming-Ching Chang, and Emily Leckman-Westin</i>	
Artificial Intelligence–Driven Wearable Technologies for Early Cardiovascular Disease Prediction: A Narrative Review	330
<i>Lariza Noleto, Amanda Guimarães, Pedro Coelho, Diandra Noleto, Matheus Moraes, and Marcos Nascimento</i>	
Robustness of Transformer-Based Fluence Map Prediction Under Clinically Realistic Perturbations	335
<i>Ujunwa Mgboh, Rafi Ibn Sultan, Joshua Kim, Kundan Thind, and Dongxiao Zhu</i>	
AI-Driven Literature Term Extraction for Optimizing Laboratory Data Capture and Prioritization	341
<i>Aída Muñoz Monjas, David Rubio Ruiz, David Perez-Rey, Kinwei Arnold Chan, and Matvey B. Palchuk</i>	
A Heuristic for Spatial Rearrangement of Large Cytology Images for Efficient Classification	346
<i>Rafael Silva Godoy, Carlos Henrique Costa Ribeiro, and Israel Tojal da Silva</i>	
OpenCDT: An Open Neuro-Symbolic Infrastructure for Executable Clinical Decision Tools	351
<i>Antonina Spalińska, Krzysztof Spaliński, Michał Chojnicki, Magdalena Jasionowska-Skop, Artur Przelaskowski, and Wojciech Glinkowski</i>	
Clinician-Supervised Large Language Models Generated Exercise Protocols for Degenerative Knee Disease: A Four-Arm Pilot Randomized Trial	357
<i>Ömer Alperen Gürses, Nazife Kapan Tunçer, and İsmail Özsoy</i>	

Scaling Healthcare Engagement with AI Voice Agents: A Real-World Evaluation Study 362
Robert Ellis, Michael Goodman, Sai Moturu, Sherene Philip, Paula Buzzard, Anmol Madan, and Sameer Berry

Conformal Prediction for Risk-Controlled Medical Entity Extraction Across Clinical Domains 367
Manil Shrestha and Edward Kim

Can LLMs Mimic Experienced Neurosurgeons in Real-World Spine Surgery Triage? 372
Mahvash Siavashpour, Erin McCabe, Andrew Nataraj, Niketh Pareek, Osmar Zaiane, and Douglas Gross

Imitation Learning for Clinical Decision Support in Pediatric ECMO 377
Fateme Golivand Darvishvand, Michael Skinner, Saurabh Mathur, Ameet Soni, Phillip Reeder, Kristian Kersting, Lakshmi Raman, and Sriraam Natarajan

Towards Multidisciplinary Summarization of Hospital Stays: Efficient Sentence-Level Clinical Section Categorization 383
Baris Karacan, Vaibhav Bhargava, Barbara Di Eugenio, Natalie Parde, Mary Khetani, Yu-Shan Tseng, Vanessa Barbosa, Julie Vignato, Lindsey Knake, Rajashree Dahal, Emily Spellman, Danielle Hitzel, Janine Petitgout, Kristi Haughey, Amanda Karstens, Brianna Clarahan, Rachel Dawson, Lauren Boyd, Mackenzie Weis, Angie Tipton, Jaewon Bae, Catherine K. Craven, Karen Dunn Lopez, and Andrew D. Boyd

Evaluating Tile-Level Feature Embeddings for Attention-Based MIL in Placental Histopathology 389
Saleh Abou-Alwan, Afsoon Khodae, Dina El Demellawy, Shannon Bainbridge, and Adrian D. C. Chan

Evaluating a Locally Deployed 20-Billion Parameter Large Language Model for Automated Abstract Screening in Systematic Reviews 395
Paulo Henrique Moreira Melo, Elena Guadagno, and Dan Poenaru

Contextualized Interpretation of Machine-Learning Predictions in Rare Disease Omics: Integrating SHAP, Biomedical Knowledge, and Language Models 399
Daniel Daza, Yorrick Jaspers, Alberto Bernardi, Luca Costabello, Christophe Gueret, Michael Cochez, Marc Engelen, Stephan Kemp, and Martijn Schut

Annotating Motivational Interviewing Transcripts Using Large Language Models	404
<i>Dinesh Rivankar, Tanmay Tulsidas Verlekar, and Abhijit Nadkarni</i>	
Forecasting Gait Dynamics with Foundation Models	410
<i>Federico Colelli Riano, Alberto Malovini, Armando Coccia, Federica Amitrano, Giovanni D'Addio, and Riccardo Bellazzi</i>	
Sequential Counterfactual Inference for Temporal Clinical Data: Addressing the Time Traveler Dilemma	415
<i>Jingya Cheng, Alaleh Azhir, Jiazi Tian, and Hossein Estiri</i>	
Explainable Case-Based Diagnosis and Risk Prediction in Oncology	421
<i>Isabelle Bichindaritz and Leszek Kotula</i>	
Multi-view Lumbar MRI for Therapy-Oriented Decision Support System in Low Back Pain	427
<i>Ruben Piperno, Luca Bacco, Giuseppe Francesco Papalia, Fabrizio Russo, Girolamo Maltese, Gianluca Vadalà, Rocco Papalia, Leandro Pecchia, and Mario Merone</i>	
A Multimodal Wearable Framework for Predicting Epileptic Seizures	433
<i>Agnese Bonfigli, Alberto Maria Di Giacinto, Flavia Giordani, Leandro Pecchia, Mario Merone, and Luca Bacco</i>	
Derived Neural Markers of Creativity: Debiased Weighted Phase Lag Index for Neurocognitive State Recognition	439
<i>Morteza Zangeneh Soroush, Daniel Massicotte, and Wei-Ping Zhu</i>	
RareGraph-AgentAI: A Multimodal Knowledge Graph and Multi-agent LLM Framework for Rare Disease Evaluation and Gene Prioritization	445
<i>Quan M. Nguyen and Kai Wang</i>	
Author Index	451



Building Trustworthy Artificial Intelligence Enabled Electronic Health Record Systems

Nabil Georges Badr¹  and Hazar Haidar² 

¹ Higher Institute for Public Health, USJ, Beirut, Lebanon
nabil@itvaluepartner.com

² Université du Québec à Rimouski, Rimouski, Quebec, Canada
hazar_haidar@uqar.ca

Abstract. This paper examines how to maintain trustworthiness when integrating artificial intelligence into electronic health record information systems (EHR-IS). By using the National Institute of Standards and Technology (NIST) AI Risk Management Framework as a conceptual lens, we synthesize literature on data-quality attributes and identify operational practices that uphold fairness, security, robustness, and governance. Our findings show that AI both amplifies upstream data weaknesses and offers mechanisms to strengthen data quality. We propose guidelines for developers and implementers in sustaining trustworthy AI-enabled EHR-IS environments.

Keywords: Artificial Intelligence · Data Quality · Electronic Health Records

1 Introduction

Electronic health record information systems (EHR-IS) capture extensive patient information, and the integration of artificial intelligence (AI) into these systems promises substantial improvements in clinical decision-making, patient outcomes, and operational efficiency. However, current EHR environments face persistent data quality (DQ) challenges, including heterogeneous data sources, inconsistent formats, duplication, missing context, and variable documentation, that undermine reliable use of AI [1]. Existing DQ frameworks do not adequately address AI-specific risks such as bias, multimodality, model drift, and lack of explainability [2]. Conversely, governance frameworks like NIST AI-RMF outline trustworthiness requirements—namely fairness, security, robustness and governance, but leaves implementers without actionable guidance [3]. Our paper bridges these gaps by proposing guidelines that incorporate AI-specific safeguards with imperatives to maintain foundational DQ attributes.

2 Background and Approach

The literature on DQ identifies key *attributes of accuracy, objectivity, credibility, concordance, completeness, timeliness, relevancy, accessibility, interpretability and accessibility* [2]. By linking NIST trustworthiness concepts to concrete DQ practices, this

framing provides a practical foundation for designing and maintaining trustworthy clinical AI (Table 1). Within this framing, the NIST AI-RMF defines fairness as detecting, measuring, and mitigating harmful bias through representative data and consistent measurement. Security focuses on protecting confidentiality, integrity, and availability via strong defenses, adversarial resistance, continuous monitoring, and resilient operations. Robustness requires maintaining valid, reliable performance across conditions by accounting for uncertainty, drift, noise, and edge-case behavior. Governance provides the oversight, structures, and documentation needed to sustain fairness, security, and robustness throughout the AI lifecycle [3]. We begin with a targeted review of the literature on AI integration within EHR-IS. We retain peer-reviewed English-language publications (2020–2025) that explicitly engage with at least one DQ attribute. The query returns 266 unique works, excluding duplicates. We then code the literature using NIST concepts as analytic categories, examining how reported practices and challenges relate to sustaining trustworthiness in these settings.

Table 1. DQ Affected - NIST Concept & Operational Practices – from our review

DQ Attributes Affected	NIST Concept & Operational Practices	Expected Outcomes
Completeness, Objectivity, Concordance	Fairness: Subgroup audits, semantic harmonization, representative sampling	Reduced inequitable outcomes
Integrity, Accessibility	Security: Encryption, federated learning, zero-trust	Protected data flows
Accuracy, Timeliness, Sequencing	Robustness: Drift monitoring, anomaly detection, multimodal consistency checks	Reliable performance
Credibility, Interpretability	Governance: Documentation standards, workflow alignment, oversight committees	Sustained accountability

3 Findings and Discussion

The integration of AI reshapes DQ by amplifying upstream weaknesses and increasing dependence on complete, structured, and well-governed inputs. Accuracy and objectivity degrade when data is incomplete, biased, or inconsistently measured, causing AI to propagate errors [4] and magnify bias [5], while credibility is undermined when data sources are unvalidated [6]. At the same time, AI can strengthen intrinsic DQ by detecting anomalies and correcting errors [7, 8] and by standardizing heterogeneous inputs [9]. Reliable inference depends on completeness, timeliness, and sequencing, requiring current data, accurate temporal ordering [10, 11], and safeguards against drift [12]. AI may enhance relevance by prioritizing clinically meaningful information and reducing subjective variation through harmonized documentation [9, 13, 14]. However, interpretability remains

fragile in the absence of semantic clarity or explanations aligned with clinical workflows [15, 16]. Concordance depend on structural consistency across multimodal data sources [17, 18], and resilient infrastructures must include secure interoperable architectures [6, 14, 19, 20].

3.1 Effective Management of Bias (Fairness)

Maintaining fairness requires proactive mechanisms to detect, measure, and mitigate harmful bias throughout the AI lifecycle. Ensuring representative and complete datasets is foundational, as incomplete or skewed data can produce inequitable outcomes [4, 9]. Fairness also depends on identifying and correcting bias arising from imbalanced or inconsistently measured inputs [5]. Documentation practices must be harmonized to reduce subjective variation that disproportionately affects underrepresented groups [13, 14], while semantic clarity and meaningful labeling help prevent misinterpretation across populations [15, 16]. Structural consistency across multimodal data sources is equally essential to avoid inequitable model behavior [17, 18]. Finally, federated learning supports fairness by enabling access to more representative datasets while limiting the exposure of sensitive subpopulations [21, 22].

3.2 Preserving Confidentiality, Integrity, and Availability (Security)

Security requires building AI systems capable of withstanding threats, vulnerabilities, and malicious interference while preserving confidentiality, integrity, and availability. Core practices include implementing secure access controls, encryption, and auditing to protect data integrity and prevent unauthorized manipulation [20]. Ensuring resilient, distributed infrastructures strengthens availability under operational stress or adversarial conditions [19]. Security also depends on federated learning, which reduces exposure risks by keeping sensitive data decentralized [21, 22]. Furthermore, the literature underscores that maintaining interoperability is essential for secure cross-institutional data exchange [14], while timely and consistent workflows help prevent integrity failures and operational vulnerabilities [6].

3.3 Ensuring Validity and Reliable Performance (Robustness)

Robustness requires ensuring that AI systems perform consistently and reliably across varying conditions, data distributions, and operational contexts. This begins with detecting anomalies and correcting errors to maintain valid outputs [7, 8]. Preventing model drift through current, complete data is essential for sustaining reliability over time [12]. Robust inference also depends on accurate sequencing metadata and temporal ordering, which preserve the integrity of event relationships [10, 11]. Standardizing heterogeneous inputs ensures consistent model behavior [15, 16], while maintaining structural consistency across multimodal data prevents degradation in edge-case or complex scenarios [17, 18]. It is also imperative to detect mismatches and inconsistencies to further support concordant, reliable data inputs [14].

3.4 Sustaining Oversight, Accountability, and Lifecycle Control (Governance)

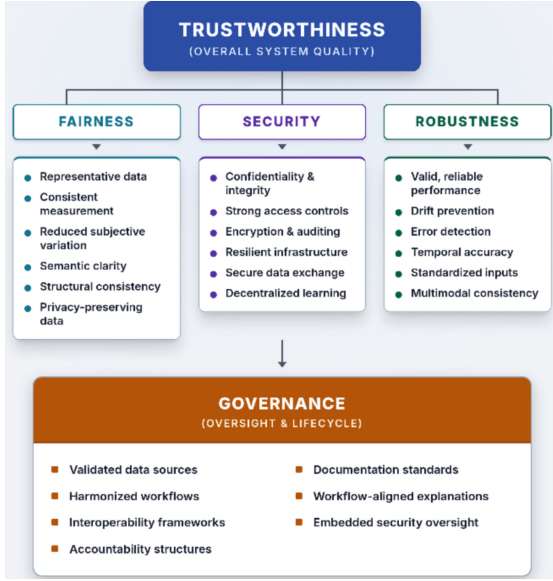


Fig. 1. Guidelines for Maintaining Trustworthiness in AI-enabled EHR-IS

Governance provides the organizational foundation that enables fairness, security, and robustness to be sustained across the AI lifecycle. Effective governance begins with the validation and verification of data sources to mitigate credibility risks [6] alongside the enforcement of standards for complete, current data and documentation [4, 9]. Harmonizing documentation practices and clinical workflows supports consistent, auditable processes across institutions [13, 14]. Governance also requires aligning system explanations with clinical workflows to promote transparency and accountable use [9]. Maintaining interoperability frameworks enables responsible cross-institutional data flows [14], while the implementation of secure access controls, encryption, and auditing mechanisms embeds security within institutional oversight [20]. Finally, governance structures, audit mechanisms, and training programs help ensure accountability across clinicians, developers, data stewards, and institutions [23].

4 Conclusion

As AI becomes embedded in routine clinical operations, maintaining trustworthiness will depend on institutional capacity to operationalize DQ-driven governance across the AI lifecycle (Fig. 1). Nevertheless, limitations to this study include the non-exhaustive, English-only literature sample, which may omit relevant evidence. Future work should validate these principles empirically and extend them across diverse clinical settings and data ecosystems.

Disclosure of Interests. The authors have no competing interests.

References

1. Badr, N.G.: Guidelines for health IT addressing the quality of data in EHR information systems. In: *HEALTHINF*, pp. 169–181 (2019)
2. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. *J. Manag. Inf. Syst.* **12**, 5–33 (1996)
3. Tabassi, E.: Artificial intelligence risk management framework (AI RMF 1.0). National Institute of Standards and Technology (U.S.). Gaithersburg, MD (2023). <https://doi.org/10.6028/NIST.AI.100-1>.
4. Bae, H., Park, S.Y., Kim, C.E.: A practical guide to implementing artificial intelligence in traditional East Asian medicine research. *Integr. Med. Res.* **13**, 101067 (2024)
5. Vaidya, A., et al.: Demographic bias in misdiagnosis by computational pathology models. *Nat. Med.* **30**, 1174–1190 (2024)
6. Lewis, D.: AI enabled EHR quality improvement: standardization, auditability, and trust. *JAMIA.* **30**, 1452–1464 (2023)
7. Bower, J.K.: Improving EHR accuracy through automated error detection. *Appl. Clin. Inform.* **8**, 924–933 (2017)
8. Chen, F., Wang, L., Hong, J., Jiang, J., Zhou, L.: Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *J. Am. Med. Inform. Assoc.* **31**, 1172–1183 (2024). <https://doi.org/10.1093/jamia/ocae060>
9. Chelliah, A.: AI driven clinical relevance filtering for decision support. *Nat. Digit. Med.* **7**, 112–125 (2024)
10. Esteva, A.: Deep learning-enabled medical computer vision. *Nat. Med.* **27**, 1322–1332 (2021)
11. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* **366**, 447–453 (2019). <https://doi.org/10.1126/science.aax2342>
12. Saxena, A.K., Dixit, R.R., Aman-Ullah, A.: An LSTM neural network approach to resource allocation in hospital management systems. *Int. J. Appl. Health Care Anal.* **7**, 1–12 (2022)
13. Adekunle, B.I., Chukwuma-Eke, E.C., Balogun, E.D., Ogunsola, K.O.: Machine learning for automation: Developing data-driven solutions for process optimization and accuracy improvement. *Mach. Learn.* **2**, 800 (2021)
14. Antunes, R.S., André da Costa, C., Küderle, A., Yari, I.A., Eskofier, B.: Federated learning for healthcare: systematic review and architecture proposal. *ACM Trans. Intell. Syst. Technol. TIST.* **13**, 1–23 (2022)
15. Liu, H., Estiri, H., Wiens, J., Goldenberg, A., Saria, S., Shah, N.: AI model development and validation (2019)
16. Musik, S., Sasin-Kurowska, J., Panczyk, M.: Bridging the past and future of clinical data management: the transformative impact of artificial intelligence. *Open Access J. Clin. Trials.* **17**, 15–33 (2025)
17. Chaabene, S., Boudaya, A., Bouaziz, B., Chaari, L.: An overview of methods and techniques in multimodal data fusion with application to healthcare. *Int. J. Data Sci. Anal.* **19**, 1–25 (2025)
18. Rotenstein, L.S., Wachter, R.M.: Are artificial intelligence-generated replies the answer to the electronic health record inbox problem? *JAMA Netw. Open.* **7**, e2438528 (2024). <https://doi.org/10.1001/jamanetworkopen.2024.38528>

19. Aminizadeh, S., et al.: Opportunities and challenges of artificial intelligence and distributed systems to improve the quality of healthcare service. *Artif. Intell. Med.* **149**, 102779 (2024)
20. Torab Miandoab, A.: Temporal reliability in AI driven clinical prediction. *Artif. Intell. Med.* **140**, 102593 (2023)
21. Beam, A.L., Kohane, I.S.: Big data and machine learning in health care. *JAMA*. **319**, 1317 (2018). <https://doi.org/10.1001/jama.2017.18391>
22. Goktas, P., Grzybowski, A.: Shaping the future of healthcare: ethical clinical challenges and pathways to trustworthy AI. *J. Clin. Med.* **14**, 1605 (2025)
23. Bouhouita-Guermech, S., Haidar, H.: Scoping review shows the dynamics and complexities inherent to the notion of “responsibility” in artificial intelligence within the healthcare context. *Asian Bioeth. Rev.* **16**, 315–344 (2024). <https://doi.org/10.1007/s41649-024-00292-7>



Leadership, Cultural Intelligence, and AI-Assisted Governance: A Compliance Perspective on the Pharmaceutical and Medical Device Industries

Ali Abou Abbas^(✉) and Arwa Asiri

Plekhanov Russian University of Economics, Dubai, UAE
ally.a.abou@gmail.com

Abstract. This article examines how transformational leadership and cultural intelligence (CQ) shape regulatory compliance in the pharmaceutical and medical device industries. Compliance, defined here as the sustained organizational capacity to meet regulatory standards through culture and conduct rather than procedural adherence alone, is increasingly challenged by the demands of Compliance 4.0—an era in which human behavioral governance and automated algorithmic oversight must operate in tandem. Despite alignment with international standards, under-reporting persists, limiting organizational learning. Using a convergent mixed-methods design—three semi-structured interviews with senior leaders overseeing MENA-region corporations and a structured questionnaire (N = 50)—findings show that compliance variation correlates with leaders' capacity to model ethical behavior and navigate culturally shaped views of authority. AI-assisted analysis supports risk detection but cannot substitute for human judgment and culturally informed leadership. The study acknowledges its exploratory scope and small qualitative sample, and positions its contribution as generating theoretically grounded hypotheses for larger-scale investigation.

1 Introduction

Compliance in pharmaceutical and medical device companies is a basic requirement for patient safety and market integrity. In a global sector valued at roughly \$1.6 trillion (Statista, 2024), standards such as GMP and ISO 13485 define the regulatory floor—yet a persistent gap between regulation and implementation consistently traces to leadership style and organizational culture rather than technical deficiency [1]. Compliance 4.0 describes the current governance era in which organizations must simultaneously manage human behavioral risks and automated algorithmic systems, recognizing that the majority of compliance failures still originate in human decisions rather than technical malfunctions [2].

The scale of the problem is measurable. FDA data show a 20% rise in data-integrity Warning Letters, signaling systemic management breakdown [3]. Punitive enforcement compounds this: up to 40% of workers in high-pressure roles suppress quality concerns to avoid consequences, normalizing risk rather than addressing it [3]. In multinational

corporations, these failures deepen when compliance expectations override local cultural norms [4]. This study therefore examines how transformational leadership and CQ together shape compliance outcomes, and where AI-assisted tools can meaningfully support—without replacing—governance.

2 Literature Review

Compliance scholarship has shifted from a legalistic to a behavioral framing of organizational integrity. Tyler [5] argues compliance is sustainable only when perceived as fair and legitimate; Weske et al. [6] link standardized procedures to patient outcomes. Bass and Riggio [7] established transformational leadership as the primary driver of “compliance citizenship”; Abou Abbas [8] applies this to pharma, finding employees follow leaders’ ethical vision over abstract mandates, while Edmondson [3] shows why silence cultures cause even robust frameworks to fail. Ang and Van Dyne’s [9] CQ framework demonstrates that regulatory adherence is frequently undermined by cultural friction. CQ’s four dimensions—metacognitive, cognitive, motivational, and behavioral—equip leaders to adapt across contexts; metacognitive CQ is especially critical in MENA pharmaceutical governance, where relational business norms can obscure the boundary between professional engagement and undue influence [8]. Leadership sets the ethical climate; CQ ensures it holds across cultural complexity. Vataman et al. [10] show AI shifts risk assessment from reactive to predictive; Geertshuis [11] notes AI-leveraging leaders make more consistent decisions. The Pfizer (\$2.3B, 2009) and Acadia (\$19.85 M, 2024) settlements illustrate how leadership disconnect translates into sanction and patient harm [12, 13].

3 Methodology

This study employs a convergent mixed-methods design, selected to triangulate statistical risk patterns with lived leadership experience.

Scope and Limitations: The sample is purposive and exploratory. Findings are not intended to generalize to the full industry but to generate theoretically grounded propositions for larger-scale testing—an appropriate scope for a mixed-methods pilot in a complex regulatory domain [14].

The qualitative component comprised three semi-structured interviews with senior leaders overseeing MENA-region corporations (5–10+ years’ experience), covering leadership dynamics, psychological safety, and AI-enabled governance; analyzed using grounded theory following Braun and Clarke [15], with codes developed inductively and verified through member checking [16]. The quantitative component surveyed 50 compliance professionals; a Composite Risk Index was constructed by multiplying mean perceived likelihood (%) by mean impact (1–4 Likert scale) per domain, following established ERM taxonomy [17]. This multiplicative approach weights domains by both frequency and severity, enabling comparative prioritization. Data were processed in SPSS; Pearson correlations assessed relationships between risk dimensions. Informed consent and anonymity were maintained throughout.

4 Results

4.1 Quantitative Risk Mapping

General Compliance (132.2) and HCP Interactions (126.5) ranked highest, followed by Promotional Activities (112.8); Data Integrity, Training, and AI-Specific Risks clustered at medium levels (Table 1). A strong Impact–Consequence correlation ($r = 0.52$, $p < 0.001$) confirms severity, not frequency, drives compliance prioritization; the non-significant Likelihood–Consequence correlation ($r = 0.18$, $p = 0.214$) reinforces this (Table 2). These moderate-to-strong correlations are consistent with an exploratory study and support the direction of the theoretical model without claiming confirmatory power (Table 3).

Table 1. Composite Risk Index by Domain (N = 50)

Risk Domain	Likelihood (%)	Impact (1–4)	Composite	Level
General Compliance	47.2	2.8	132.2	HIGH
HCP Interactions	40.8	3.1	126.5	HIGH
Promotional Activities	38.9	2.9	112.8	HIGH
Training Adequacy	36.4	2.7	98.3	MEDIUM
Data Integrity	32.6	2.9	94.5	MEDIUM
AI-Specific Risks	33.5	2.8	93.8	MEDIUM

Table 2. Correlation Analysis Between Risk Dimensions

Correlation	Pearson’s r	Sig.	Interpretation
Likelihood – Impact	0.40	$p < 0.001$	Moderate positive
Likelihood – Consequence	0.18	$p = 0.214$	Not significant
Impact – Consequence	0.52	$p < 0.001$	Strong positive

4.2 Qualitative Themes

Table 3. Participant Profiles

ID	Role	Experience	Region
IO1	Senior Compliance Officer	5+ yrs	MENA / Regional
IO2	Compliance Director	10+ yrs	Middle East & Europe (MENA)
IO3	Head of Ethics & Compliance	8+ yrs	Global (MENA oversight)

Theme 1 Leadership as Compliance Driver

IO3: “Trust and tone from the top... leadership shows all employees how compliance should be top down.”

This theme emerged from codes clustered around authority modeling, ethical signaling, and reporting norms across all three interviews. Participants uniformly identified leadership tone as the primary compliance determinant, aligning with Bass and Riggio [7] on transformational leadership and Edmondson [3] on psychological safety. The convergence across participants—despite their different regional contexts—strengthens confidence in this theme despite the small sample.

Theme 2 CQ and High-Risk Interaction Areas

IO2: “Grants are of high risk because they could be used as a kind of influencing HCPs to prescribe or use company products.”

Codes related to cultural ambiguity, relationship-based risk, and boundary management concentrated in this theme. HCP interactions, grants, and promotional activities were the most difficult domains to govern—precisely where CQ matters most. Leaders lacking metacognitive CQ are more likely to normalize ethical drift in relationship-based business contexts [8]. Tyler’s [5] procedural justice framework explains why rules lose force when perceived as culturally illegitimate rather than contextually grounded.

Theme 3 AI as Supportive Tool: Managerial Perception

IO1: “AI will play a role in analyzing data... it’s very hard to rely on AI to run a compliance program on its own, as it relies on human data entry and does not verify data authenticity.”

This theme reflects how compliance managers perceive AI, not an evaluation of implemented AI systems. Codes in this theme captured cautious optimism, data dependency concerns, and human oversight requirements. Findings are consistent with Vataman et al. [10] and Geertshuis [11]: AI enhances monitoring capability but cannot replace ethical judgment or CQ-informed interpretation. Effective AI deployment requires leaders who can configure and contextualize outputs across culturally distinct regulatory environments.

5 Discussion and Conclusion

Taken together, the quantitative and qualitative findings show that compliance risk is fundamentally a behavioral and relational problem. The Impact–Consequence correlation ($r = 0.52$) shows severity drives resource allocation more than frequency—a pattern with

direct implications for how compliance programs should be designed and prioritized. Transformational leadership and CQ operate as equally important governance pillars: leadership sets the ethical climate; CQ ensures it functions across cultural complexity. This distinction is especially consequential in MENA contexts, where hierarchical dynamics and relational norms shape how authority and risk are perceived, and where metacognitive CQ determines whether leaders recognize early-stage ethical drift before it becomes a regulatory event.

AI tools add value for risk detection but are limited by data quality and the cultural competence of those interpreting their outputs. The study's AI findings reflect managerial perception rather than system evaluation, a limitation that opens the door to future empirical work within MENA regulatory environments, where cultural specificities may expose limits in tools designed for other contexts. The combined framework of transformational leadership, CQ, and AI-assisted monitoring offers a governance model that is both evidence-based and actionable to multinational compliance challenges.

References

1. Pagliaro, S., Lo Presti, A., Barattucci, M., Giannella, V.A., Barreto, M.: On the path towards ethical behavior: how ethical culture and leadership influence employees' regulatory compliance. *Front. Psychol.* **9**, 961 (2018)
2. Deloitte: quality 4.0: the future of quality in pharma and MedTech. Deloitte Insights (2023)
3. Edmondson, A.C.: *The Fearless Organization: Creating Psychological Safety in the Workplace for Learning, Innovation, and Growth*. Wiley, Hoboken (2018)
4. Hofstede, G.: Dimensionalizing cultures: the Hofstede model in context. *Online Read. Psychol. Cult.* **2**(1), 8 (2011)
5. Tyler, T.R.: *Why People Obey the Law*. Princeton University Press, Princeton (2006)
6. Weske, U., Boselie, P., Van Rensen, E.L., Pors, H.J.: Strategic human resource management and quality of care. *J. Health Organ. Manag.* **32**(3), 465–481 (2018)
7. Bass, B.M., Riggio, R.E.: *Transformational Leadership*, 2nd edn. Psychology Press, New York (2006)
8. Abou Abbas, A.: *Marketing strategies of international pharmaceutical companies in the middle East and North Africa region*. Doctoral dissertation, Walden University (2018)
9. Ang, S., Van Dyne, L.: Cultural intelligence: its measurement and effects on cultural judgment and decision making, cultural adaptation and task performance. *Manag. Organ. Rev.* **3**(3), 335–371 (2007)
10. Vataman, C., Lungu, C., Gheorghe, I.G.: Artificial intelligence in pharmaceutical risk management: a systematic review of current trends and regulatory compliance. *J. Med. Syst.* **48**(1), 15–29 (2024)
11. Geertshuis, S.A.: Transformational leadership and the adoption of artificial intelligence. *Front. Psychol.* **13**, 1–10 (2022)
12. Department of Justice: Justice department announces largest health care fraud settlement in its history (2009)
13. Department of Justice: Acadia healthcare to pay \$19.85 million to settle allegations of improper billing (2024)
14. Patton, M.Q.: *Qualitative Research and Evaluation Methods*, 4th edn. SAGE Publications, Thousand Oaks (2015)
15. Braun, V., Clarke, V.: Using thematic analysis in psychology. *Qual. Res. Psychol.* **3**(2), 77–101 (2006)

16. King, N., Horrocks, C.: Interviews in Qualitative Research. SAGE Publications, London (2010)
17. Statista: Total global spending on medicines 2023–2027. Statista Research Department (2024)



A PRISMA-Aligned Workflow for LLM-Assisted Title–Abstract Screening: A Methodological Case Study in Atrial Fibrillation Literature

Olena Holub¹ , Alex Nakonechnyi² , Joshua Metcalf¹ , and Rebecca Allen³ 

¹ University of Cincinnati, Cincinnati 45221, OH, USA
holuboa@mail.uc.edu

² University of Rochester Medical Center, Rochester 14642, NY, USA

³ Mount St. Joseph University, Cincinnati 45233, OH, USA

Abstract. Systematic literature reviews are central to evidence synthesis but remain resource-intensive due to screening and data preparation demands. Large language models (LLMs) offer opportunities to improve efficiency while maintaining transparency, but PRISMA-aligned implementations of LLM content screening remain limited. Using a cardiovascular methodological case study, we developed a PRISMA-aligned first-pass workflow integrating automated deduplication, abstract retrieval, LLM-assisted title–abstract screening, and manual validation. This structured approach reduces manual workload in early review stages while supporting scalable and auditable systematic review workflows.

Keywords: Atrial fibrillation · Generative Artificial Intelligence · Multimodal risk stratification · Large Language Models (LLMs) · PRISMA Framework · Systematic Literature Review

1 Introduction

Systematic reviews support evidence synthesis, yet early-stage screening remains resource-intensive. Large language models (LLMs) may improve efficiency, but reproducible PRISMA-aligned implementations remain limited [1–4]. In this case study in atrial fibrillation literature, we asked: What is known about multimodal data for identifying patients at risk of persistent atrial fibrillation? Multimodal prediction remains heterogeneous with modest performance [5–8], while early identification is important [9–11]. We developed a reproducible PRISMA-aligned workflow integrating automated preparation, LLM screening, and consensus validation for academic community.

2 Methodology

2.1 Information Sources and Search Strategy

Search queries were adapted to database-specific indexing in PubMed, Scopus, and IEEE Xplore to capture atrial fibrillation, multimodal analysis, and disease progression. Full search strings are provided in the Supplementary Materials.

2.2 Deduplication and Metadata Preparation

A Python-based deduplication pipeline consolidated records sharing identical DOIs. For records without DOIs, duplicates were identified using a $\geq 90\%$ similarity threshold across normalized title, first author, and publication year, yielding 10,930 unique entries. Metadata were standardized and missing abstracts recovered through API queries (Cross-Ref, PubMed, OpenAlex, Semantic Scholar, Europe PMC), followed by cleaning and validation. The finalized dataset was exported to SQLite for screening.

2.3 Automated Title–Abstract Screening with LLM

Model Setup. Title–abstract screening was automated using the DeepSeek-R1-Distill-Qwen-14B model deployed locally through LM Studio (v0.3.14). The model was loaded in GGUF format (Q8_0 quantization, 15.7 GB, Qwen2 architecture) and accessed through an OpenAI-compatible HTTP API using a Python 3.11.5 client. Runtime parameters included a 16,384-token context length, GPU utilization, 24 CPU threads, and a batch size of 512, with deterministic inference enforced by fixing temperature to zero. Quoted evidence was validated against source text to prevent unsupported rationales, and decision fields were normalized using predefined categories and confidence bounds. Validated outputs were stored in SQLite as an auditable log. Processing 10,930 records required 30 h. Runtime scaled proportionally with dataset size under fixed hardware. Full code and prompts are provided in the Supplementary Materials.

Prompt Structure and Instructional Design. Eligibility criteria were encoded directly into the prompt, requiring studies to (1) include adults with atrial fibrillation, (2) use multimodal data defined as integration of at least two distinct sources, and (3) address progression to persistent atrial fibrillation or early treatment. The model returned structured outputs consisting of five predefined fields. Screening Outcome recorded eligibility classification (Include, Exclude, or Potentially Eligible). Coded Reason for Decision was selected from a predefined taxonomy to standardize classifications. Quoted Rationale extracted short verbatim phrases from the title or abstract as supporting evidence. Structured Rationale provided a brief explanation linking evidence to eligibility rules. Confidence Score recorded relative certainty (0–1) for logging and troubleshooting, not as a calibrated probability (Fig. 1).

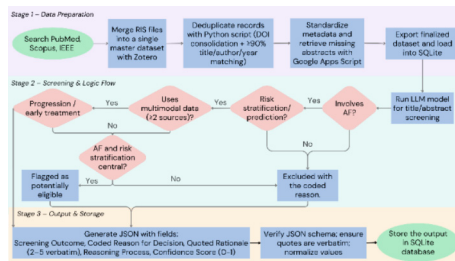


Fig. 1. Screening and Eligibility Decision Workflow

Manual Validation. To evaluate classification reliability, a stratified validation sample was drawn from the 10,930 screened records, of which 829 (7.6%) were labeled as Include or Potentially Eligible. Because of the low prevalence of included records, 100 LLM-Included and 400 LLM-Excluded records were selected using simulation-based sample sizing, with post-stratification weights applied to reflect dataset distribution. Three human reviewers independently screened sampled records in pairs, with disagreements resolved by a third reviewer serving as arbiter. Agreement between LLM and human classifications was evaluated using Cohen’s κ with bootstrap confidence intervals. Treating Include and Potentially Eligible as screening-positive classifications, the LLM achieved $\kappa = 0.835$ (95% CI: 0.769–0.896). Among 76 human-labeled inclusions, all were identified (sensitivity = 1.000), with 24 false positives and no false negatives (specificity = 0.943). Positive predictive value was 0.760, negative predictive value was 1.000, and $F1 = 0.864$, with positive and negative agreement of 0.864 and 0.971. False positives mainly reflected ambiguous descriptions of multimodal methods, with no consistent patterns among exclusions. Among 40 Potentially Eligible records, 22 (55%) were confirmed as inclusions and 18 (45%) as exclusions (Table 1).

Table 1. Confusion matrix comparing LLM classifications with consensus human labels.

	Human Include	Human Exclude	Total
LLM Include or Potentially Eligible	76	24	100
LLM Exclude	0	400	400
Total	76	424	500

3 Findings

3.1 Dataset Flow

The search strategy across PubMed, Scopus, and IEEE Xplore retrieved 14,550 records, yielding 10,930 unique entries after deduplication. Among these, 441 records lacked abstracts; 107 were recovered, while 334 remained unavailable. All records were retained, as classification could proceed using titles when abstracts were missing. Of the 10,930 screened records, 479 (4.4%) were classified as Include, 350 (3.2%) as Potentially Eligible, and 10,101 (92.4%) as Not Eligible. Most records were excluded at the abstract stage, reducing downstream manual review requirements (Fig. 2).

3.2 Practical Considerations

Large-scale screening required customized workflows, as few open-source tools support datasets of this size. Reliable metadata normalization and abstract recovery were necessary to maintain consistent input quality. Several design principles supported stable performance, including structured eligibility rules, predefined output categories,

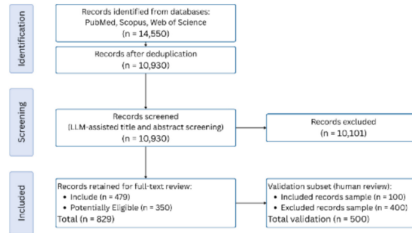


Fig. 2. PRISMA flow diagram

deterministic settings, and required evidence from source text. The introduction of a Potentially Eligible category improved efficiency by directing manual review toward borderline cases while preserving relevant studies. These principles support reproducible large-scale screening workflows and may be applied to other systematic review settings (Table 2).

Table 2. LLM pipeline features and use cases in the study.

Feature	Use Case in the Study
Short illustrative examples embedded in rules	Prompts included brief positive and negative examples to anchor decisions, such as “multimodal fusion model...” (Include) and “ventricular tachycardia...” (Exclude).
Eligibility criteria defined using fixed rule-based logic	Records were classified only when all required elements were present, following rules such as “assign if the article addresses all required conditions.”
Required direct text evidence from abstracts	The model extracted short verbatim phrases (2–5 snippets) from titles or abstracts to justify each classification decision.
Output constrained to predefined categories	Decisions were restricted to three outcomes: Include, Exclude, or Potentially Eligible.
Deterministic inference configuration	Model temperature was fixed at zero to prevent randomness and ensure identical outputs across runs.

4 Conclusion

This study demonstrates that large-scale, LLM-assisted screening of bibliographic datasets is feasible when reproducibility, transparency, and structured eligibility rules are integrated into workflow design. Within this framework, large language models function as rule-bound assistants that provide consistent, auditable first-pass classifications while preserving the central role of human reviewers. This methodological case study

was limited to title–abstract screening with the human in the loop; sensitivity testing using alternative prompts or methods was outside the scope of this work. Future work will extend this workflow to full-text screening, enabling PRISMA-aligned eligibility confirmation and structured extraction of key study characteristics and bias domains.

Supplementary Material

GitHub repository containing the database search strategies and all code used throughout the project: <https://github.com/olenaholubai/prisma-llm-af>.

References

1. Li, M., Sun, J., Tan, X.: Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. *Syst. Rev.* **13**, 219 (2024)
2. Susnjak, T.: PRISMA-DFLLM: an extension of PRISMA for systematic literature reviews using domain-specific finetuned large language models. arXiv preprint arXiv:2306.14905 (2023). <https://doi.org/10.48550/arXiv.2306.14905>
3. Lieberum, J.L., et al.: Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review. *J. Clin. Epidemiol.* **181**, 111746 (2025)
4. Galli, C., Gavrilova, A.V., Calciolari, E.: Large language models in systematic review screening: opportunities, challenges, and methodological considerations. *Information.* **16**, 378 (2025). <https://doi.org/10.3390/info16050378>
5. Wang, Q., et al.: Plasma Galectin-3 is associated with progression from paroxysmal to persistent atrial fibrillation. *BMC Cardiovasc. Disord.* **21**, 226 (2021)
6. Wan, P., et al.: Right atrial wall inflammation detected by (18)F-FDG PET/CT may be significantly associated with persistent atrial fibrillation: a prospective case-control study. *BMC Cardiovasc. Disord.* **23**, 587 (2023)
7. Atta-Fosu, T., et al.: A new machine learning approach for predicting likelihood of recurrence following ablation for atrial fibrillation from CT. *BMC Med. Imaging.* **21**, 45 (2021). <https://doi.org/10.1186/s12880-021-00578-4>
8. Yin, Y., et al.: Deep learning radiomics of left atrial appendage features for predicting atrial fibrillation recurrence. *BMC Med. Imaging.* **25**, 186 (2025)
9. Kirchhof, P., et al.: Early rhythm-control therapy in patients with atrial fibrillation. *N. Engl. J. Med.* **383**, 1305–1316 (2020)
10. Camm, A.J., et al.: The increasing role of rhythm control in patients with atrial fibrillation: JACC state-of-the-art review. *J. Am. Coll. Cardiol.* **79**, 1932–1948 (2022). <https://doi.org/10.1016/j.jacc.2022.03.337>
11. Andrade, J.G., et al.: Progression of atrial fibrillation after cryoablation or drug therapy. *N. Engl. J. Med.* **388**, 105–116 (2023). <https://doi.org/10.1056/NEJMoa2212540>



Reasoning-Enhanced Rare-Event Prediction with Balanced Outcome Correction for Clinical Decision Support

Vitaly Bulgakov^{1,2(✉)} and Alexander Turchin¹

¹ Mass General Brigham, Boston, MA, USA
vbulgakov@bwh.harvard.edu

² Profitaya LLC, Boston, MA, USA

Abstract. Rare but critical clinical events such as in-hospital cardiac arrest (IHCA), unexpected ICU transfer, rapid clinical deterioration, or hospital readmission remain challenging targets for machine-learning models due to extreme class imbalance and high false-positive costs. Recent advances in large language models (LLMs) enable reasoning over unstructured clinical narratives, but their direct use for prediction is limited by bias and lack of calibration.

We present an ongoing research framework inspired by the Low-Prevalence CORrector for Prediction method (LPCORP, provisional patent filed) that combines LLM-generated clinical reasoning with a lightweight correction model to improve rare-event prediction without resampling. We describe the methodology, provide results on rare-event medical prediction tasks, and discuss implications for interpretability, cost-aware decision support, and safe deployment in clinical settings. We provide a cost reduction analysis that quantifies the return from preventive measures compared to post-factum damage control.

Keywords: Rare-event prediction · Clinical decision support · Large language models · Cost-sensitive learning · Machine Learning · Artificial Intelligence · Reasoning · Clinical Notes

1 Introduction

Rare but critical clinical events are low-prevalence outcomes with high clinical and operational impact. Extreme class imbalance and high false-positive costs limit existing prediction models and contribute to alert fatigue.

This work focuses on unstructured clinical notes, which capture contextual and anticipatory signals often missed by structured data but are difficult to exploit in low-prevalence settings. We evaluate the proposed method across multiple domains, including in-hospital cardiac arrest (IHCA), a rare event affecting only 1–4% of hospitalized patients, where early intervention can substantially reduce morbidity, mortality, and cost, as well as a hospital readmission prediction within 30 days after discharge.

We propose a two-stage reasoning-enhanced framework that leverages large language models (LLMs) to analyze clinical narratives and applies a balanced outcome correction stage to mitigate bias in rare-event prediction. In the second stage we use and compare Logistic Regression (LR) and a simple Multi-Layer Perceptron (MLP). Our evaluation of standalone LR, MLP, and LLM predictors showed, that all baselines consistently underperformed LPCORP across metrics. Standalone LR and MLP models trained directly on the task achieved lower performance, indicating that the LLM contributes additional predictive signal. We evaluated several open-weight reasoning LLMs with $\sim 7\text{--}10\text{B}$ parameters and report results using the best-performing model. Its relatively small size enables deployment of LPCORP in a local environment.

Performance is assessed using cost-aware analysis to estimate the net benefit of prediction-guided interventions, aligning evaluation with clinical decision-making priorities.

2 Related Work

Rare-event prediction under extreme class imbalance is known to make standard accuracy metrics misleading and precision highly sensitive to false positives [1], motivating cost-sensitive evaluation aligned with decision impact [2]. In health-care, models for events such as in-hospital cardiac arrest (IHCA) often rely on structured data and early warning systems, achieving high sensitivity but producing excessive false positives and alert fatigue [3].

Unstructured clinical notes provide additional contextual signal beyond structured data, but note-based approaches remain difficult to calibrate in low-prevalence settings. State-of-the-art large language models (LLM) offer strong reasoning capabilities over such narratives, yet their direct use for prediction is limited by bias and lack of calibration.

This work builds on our prior framework for reasoning-enhanced rare-event prediction with balanced outcome correction [4], which reframes prediction as correctness assessment. We adapt it to clinical decision support using unstructured notes and evaluate it through cost-aware analysis. Further discussion of rare-event prediction, challenges associated with highly imbalanced data, cost evaluation, and applications in the medical domain can be found in [4–12].

3 Method

We propose LPCORP, a two-stage framework addressing low-prevalence prediction challenges, such as majority-class dominance, and enhancing predictive capacity using features derived from LLM reasoning over clinical notes toward the target outcome.

Stage 1: Reasoning. Given input x , an LLM produces reasoning $R(x)$ and a tentative label $C(x)$ (conclusion) using structured prompting. The reasoning trace provides enriched representation without fine-tuning.

Stage 2: Correction. Let $G \in \{0, 1\}$ denote the ground-truth label and define the correctness indicator

$$S = 1\{C(x) = G\}.$$

A logistic regression (LR) or another conventional machine learning model is trained to predict S using features derived from the combined text of clinical notes x and the corresponding reasoning $R(x)$, and is applied to the same combined representation at inference. Thus, LR estimates the correctness of the LLM output rather than the outcome itself.

Let $\hat{S} \in [0, 1]$ denote the predicted probability of correctness. Applying a threshold p , the final prediction is

$$\hat{y} = C(x) \mathbf{1}\{\hat{S} \geq p\} + (1 - C(x)) \mathbf{1}\{\hat{S} < p\}$$

We define $\text{TPR}(p) = \Pr(\hat{y} = 1 \mid y = 1)$; $\text{TNR}(p) = \Pr(\hat{y} = 0 \mid y = 0)$, and select p by minimizing a cost function:

$$p^* = \arg \min_p \mathcal{C}(\text{TPR}(p), \text{TNR}(p)).$$

4 Prediction-Driven Cost Reduction

As this study addresses rare-event prediction, where post-event damage control is substantially more expensive than preventive intervention, the objective is cost reduction rather than conventional prediction metrics. The proposed formulation does not require absolute cost values and depends only on the cost ratio $k = C_i/C_e$, where C_i is the intervention cost and C_e is the event cost.

Starting from the return definition

$$\text{Return} = (C_b - \text{Cost})/C_b$$

where C_b denotes the baseline cost and Cost the cost under prediction-guided intervention, and omitting intermediate derivation for brevity, we obtain

$$\text{Return} = R_e \left(E - \frac{k}{Pr} \right),$$

where R_e is recall, Pr is precision, and E is the efficacy of prevention.

5 Experiments

We conducted an extensive numerical study to evaluate the performance of LPCORP on multiple rare-event prediction tasks. Due to space constraints, we present here results for two hospitalization-related tasks using data from large clinical datasets: prediction of 30-day patient readmission (PR30) and prediction of in-hospital cardiac arrest (IHCA) within next 6 h. The data for PR30 were collected from the MIMIC-IV dataset discharge notes and for IHCA from

MIMIC-III notes and procedures tables, both available through PhysioNet. In both cases, 10,000 admissions were randomly sampled from the database. After the LLM reasoning stage, this number was slightly reduced due to some unsuccessful reasoning outputs. In the correction stage, the data were split into training and test sets in an 80/20 ratio, and all evaluations were performed on the test set. In the correction stage we compared performance of two “traditional” models - Logistic Regression (LR) and a simple Multi-Layer Perceptron (MLP) as a representative of the neural network. In the IHCA case the pseudo-IHCA times were sampled from the empirical time-to-IHCA distribution to ensure temporal alignment across cases.

For cost-reduction evaluation, we used the Return formulation introduced above with $k = 0.02$, corresponding to an intervention cost equal to 2% of the event cost, and set $E = 1$ for simplicity. In both tasks, we evaluated metrics across different probability thresholds P ranging from 0.5 to 0.9. In the tables presented, we report results corresponding to the optimal value of P in each case.

Figure 1 and Fig. 2 show that logistic regression (LR) outperformed the MLP in terms of return in the correction stage, suggesting that a simpler linear model was sufficient for this step. In the PR30 task, this resulted in up to 42% cost reduction when preventive measures were applied based on the predictions. Figure 3 presents the results for the IHCA case. The improvement over the baseline LLM predictions after applying the correction step is smaller than in the previous setting, but still substantial. The optimal threshold P is 0.9, which is substantially higher than in the earlier example. This can be explained by the very low prevalence (0.02). In such highly imbalanced settings, the data are overwhelmingly composed of negative cases, so even a small false positive rate leads to a large number of false positives. Increasing the threshold mitigates this effect, improving precision and overall performance.

Settings	Num Samples	Accuracy	Precision	Recall	Return
LLM Reasoning	1842	0.727	0.316	0.16	14%
Confident cases	1842	0.685	0.356	0.453	42%
Final corrected	1842	0.685	0.356	0.453	42%

Fig. 1. Prediction metrics and Return for PR30 at threshold $P = 0.5$ with LR corrector. Prevalence = 0.2

Settings	Num Samples	Accuracy	Precision	Recall	Return
LLM Reasoning	1842	0.727	0.316	0.16	14%
Confident cases	1842	0.739	0.4	0.269	26%
Final corrected	1842	0.739	0.4	0.269	26%

Fig. 2. Prediction metrics and Return for PR30 at threshold $P = 0.5$ with MLP corrector. Prevalence = 0.2

Settings	Num Samples	Accuracy	Precision	Recall	Return
LLM Reasoning	1787	0.741	0.037	0.425	19%
Confident cases	654	0.977	0.091	0.167	13%
Final corrected	1787	0.822	0.049	0.375	22%

Fig. 3. Prediction metrics and Return for IHCA at threshold $P = 0.9$ with LR corrector. Prevalence = 0.02

6 Conclusion

LPCORP combines LLM-based reasoning with a correction stage to improve rare-event prediction. The results demonstrate consistent performance gains and substantial cost reduction compared to the baseline cost. Despite relatively low precision and recall, the method achieves significant cost savings in low-prevalence settings, reflecting the strong asymmetry between event and intervention costs. In such scenarios, cost-aware evaluation provides a more appropriate measure of utility than standard classification metrics. The proposed method offers a natural approach to addressing the challenges of highly imbalanced data. The first stage, based on LLM reasoning, does not require training and is therefore largely insensitive to class imbalance. The second stage, which involves training, addresses a different and more balanced task. These results confirm that the performance gains arise from the integration of LLM reasoning with supervised correction rather than from either component alone. The method can be used in local settings and does not require cloud-based or centralized infrastructure.

References

1. Davis, J., Goadrich, M.: The relationship between Precision–Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning (ICML) (2006)
2. Elkan, C.: The foundations of cost-sensitive learning. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI) (2001)
3. Johnson, A.E.W., et al.: Machine learning and decision support in critical care. Proceedings of the IEEE (2016)
4. Bulgakov, V., et al.: Reasoning-enhanced rare-event prediction with balanced outcome correction. arXiv preprint [arXiv:2601.16406](https://arxiv.org/abs/2601.16406) (2026)

5. Brownlee, J.: Cost-Sensitive Learning for Imbalanced Classification (2020). <https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification>
6. Hand, D.J.: Evaluating diagnostic tests: the area under the ROC curve and the balance of errors. *Stat. Med.* **29**(14), 1502–1510 (2010)
7. Andersen, L.W., et al.: In-hospital cardiac arrest: a review. *JAMA – The Journal of the American Medical Association* (2019)
8. Lee, H.-Y., et al.: Prediction of in-hospital cardiac arrest in the intensive care unit using a machine-learning multimodal approach. *JMIR Med. Inf.* (2024)
9. He, H., Garcia, E.A.: Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* (2009)
10. Baaĵ, I., et al.: Synergies between machine learning and reasoning. *Int. J. Approximate Reasoning* **171**, 109206 (2024)
11. McDermott, M., et al.: Reproducibility in machine learning for health research: Still a ways to go. *Sci. Transl. Med.* **13**(586), eabb1655 (2021)
12. Bonafide, C.P., et al.: Cost-benefit analysis of a medical emergency team in a children’s hospital. *Pediatrics* **134**(2), 235–241 (2014)



Explainable AI-Driven Optimized LSTM for Heart Sound Classification and Diagnostic Support

Faiq Ahmad Khan¹, Arshad Hassan^{2(✉)}, Lia Anwar³, and Amine Bermak²

¹ Department of Electrical and Computer Engineering, The University of Alabama, Tuscaloosa, AL, USA

² College of Science and Engineering, Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar

mukhan@hbku.edu.qa

³ Rehman Medical College, Peshawar, Pakistan

Abstract. Heart disease mortality necessitates accurate and interpretable diagnostic tools. This study optimizes a Long Short-Term Memory (LSTM) network for heart sound classification via GridSearchCV and Bayesian optimization. GridSearchCV achieved a balanced clinical profile (89.04% accuracy, 84.09% sensitivity), while Bayesian optimization maximized specificity (92.68%) at the expense of sensitivity (71.59%). To ensure clinical transparency, SHAP and LIME were integrated. SHAP identified global critical features (S1 and S2 means), while LIME provided patient-specific, local interpretability. The integration of rigorous optimization with Explainable AI validates network behavior and enhances clinical trust in automated diagnostic support systems.

Keywords: Explainable AI (XAI) · Hyperparameter Optimization · Heart Sound Classification · Phonocardiogram (PCG) · Cardiac Diagnostics

1 Introduction

Heart diseases are a primary cause of global mortality [1], making early detection through heart sound analysis crucial. However, distinguishing normal from abnormal phonocardiogram (PCG) signals is highly complex [2]. While manual auscultation remains the gold standard, its inherent subjectivity demands extensive clinical experience. Consequently, machine learning (ML) is increasingly utilized to provide robust, objective automated alternatives for classification [3].

Recent studies have successfully applied 1-dimensional Convolutional Neural Networks (1D-CNN) [4] and Artificial Neural Networks (ANN) [5] to classify these sounds using features like Mel-Frequency Cepstral Coefficients (MFCC). Despite their high accuracy, these models frequently suffer from limited transparency, hindering clinical trust. To address temporal dependencies, Long Short-Term Memory (LSTM) networks have proven highly effective for sequential PCG analysis [6], particularly when leveraging time- and frequency-domain features like the Discrete Wavelet Transform (DWT)

[7]. Maximizing LSTM performance requires rigorous hyperparameter optimization to avoid overfitting and ensure generalization.

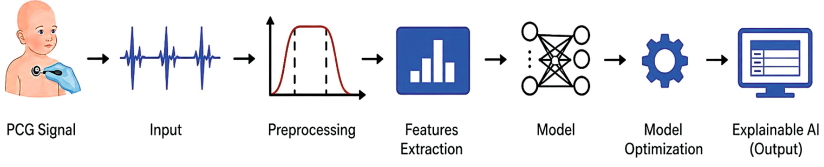


Fig. 1. Block diagram of the proposed methodology

Therefore, this study proposes an optimized LSTM network for robust heart sound classification, driven by two main contributions. First, GridSearchCV and Bayesian approaches are rigorously applied to the LSTM architecture to achieve significant performance improvements over baseline models. Second, the study integrates SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) to interpret the model’s decision-making process. While similar integrations of SHAP and LIME exist in prior literature, this work provides a uniquely structured, comparative analysis of their outputs specifically applied to hyperparameter-optimized LSTMs for cardiac diagnostic support. These contributions demonstrate the critical interaction between advanced optimization techniques and explainable AI in building reliable clinical tools. The block diagram of the complete model is shown in Fig. 1.

2 Methodology

2.1 Dataset, Segmentation, and Feature Extraction

This study utilizes the 2016 PhysioNet/CinC Challenge [8] dataset, comprising 3,126 recordings. The dataset’s inherent class imbalance is mitigated via robust evaluation metrics rather than explicit data resampling. Signals were denoised using a third-order Butterworth band-pass filter (25–200 Hz). Following the approach established by Springer et al. [9], an extended hidden semi-Markov model (HSM) enhanced with logistic regression was implemented to accurately segment the first and second heart sounds. Finally, features were extracted by applying the DWT to each RR interval (the time elapsed between consecutive R-waves) to capture mean and standard deviation coefficients, alongside MFCCs to capture spectral variability.

2.2 LSTM Architecture and Optimization Strategy

This study utilizes a Long Short-Term Memory (LSTM) neural network to effectively model the complex temporal dependencies within the extracted PCG signal features [10]. Rather than relying on static, empirical configurations, the network’s hyperparameters were systematically tuned to maximize diagnostic support capabilities. Two distinct frameworks were employed: GridSearchCV, which executed an exhaustive search over a discrete predefined parameter space to ensure a balanced diagnostic profile,

and Bayesian Optimization, which utilized a probabilistically guided sequential search designed specifically to maximize the model’s F1-score. This dual-optimization strategy enables a rigorous comparison of performance trade-offs.

Table 1. Benchmark with previous studies in (%).

Author	Dataset	Accuracy	Recall	Specificity	F1-Score
Bozkurt <i>et al</i> [13]	PhysioNet [8]	81.5	84.5	78.5	
Potes <i>et al</i> [14]	PhysioNet [8]	–	94.24	77.81	
Alkhodari <i>et al</i> [15]	PhysioNet [8]	–	92.78	79.48	
Proposed Exp 1.	PhysioNet [8]	86.88	90.91	86.25	65.31
Proposed Exp 2.	PhysioNet [8]	89.04	84.09	89.82	67.58
Proposed Exp 3.	PhysioNet [8]	89.81	71.59	92.68	65.62

2.3 Explainable AI (XAI) and Evaluation Metrics

To ensure clinical interpretability, this study employs LIME and SHAP [11, 12]. SHAP provides global feature importance based on cooperative game theory, ensuring additive consistency across the dataset. Conversely, LIME approximates the model’s behavior locally using simpler surrogate models around specific predictions. These established frameworks are applied to the optimized LSTM to provide both population-level validation and patient-specific diagnostic explanations. Furthermore, to address the dataset’s inherent class imbalance without explicit data resampling, model performance is strictly evaluated using the F1-score and balanced accuracy, ensuring reported capabilities accurately reflect true clinical proficiency.

3 Results and Discussion

Optimization Results and Clinical Trade-Offs: The baseline LSTM (Exp. 1) yielded 86.88% accuracy and 90.91% sensitivity but low precision (50.96%). GridSearchCV (Exp. 2; 32 units, 0.001 LR, batch 64) produced a highly balanced profile (89.04% accuracy, 84.09% sensitivity, 67.58% F1-score). Bayesian Optimization (Exp. 3; ~49 units, 0.0039 LR) achieved the highest accuracy (89.81%) and specificity (92.68%), but sensitivity dropped to 71.59%. Because sensitivity is paramount for anomaly detection, GridSearchCV provides a safer primary screening profile. The Bayesian model is better suited as a secondary, confirmatory tool to rule out false positives.

Table 1 benchmarks these results against three baselines [13–15] evaluated on the exact PhysioNet dataset. Direct comparison is limited as prior studies inconsistently report metrics (frequently omitting F1-scores) and heavily prioritize recall over specificity. Despite this, Exp. 2 achieves a highly competitive, balanced clinical profile. For example, while ensemble methods [14] reach 94.24% recall, they suffer from poor specificity (77.81%). Conversely, Exp. 2 delivers superior balanced diagnostic support, improving specificity to 89.82% while maintaining a robust F1-score of 67.58%.

Explainability Analysis: To ensure both individualized and population-level transparency, the complementary outputs of LIME and SHAP were compared. LIME provides local, patient-specific interpretability; its output confirms that specific feature thresholds (e.g., $S2_mean_dM04 > 0.74$) drive an abnormal prediction for an individual patient (Fig. 2a). Conversely, SHAP provides a global perspective, validating the network’s overall logic across the entire dataset (Fig. 2b). A comparative analysis of their insights reveals a strong physiological consensus: the specific features LIME isolates as critical for individual abnormal cases are the exact same features SHAP validates as universal markers for anomaly detection. Ultimately, LIME serves as an individualized clinical explanation tool, whereas SHAP provides broad model validation.

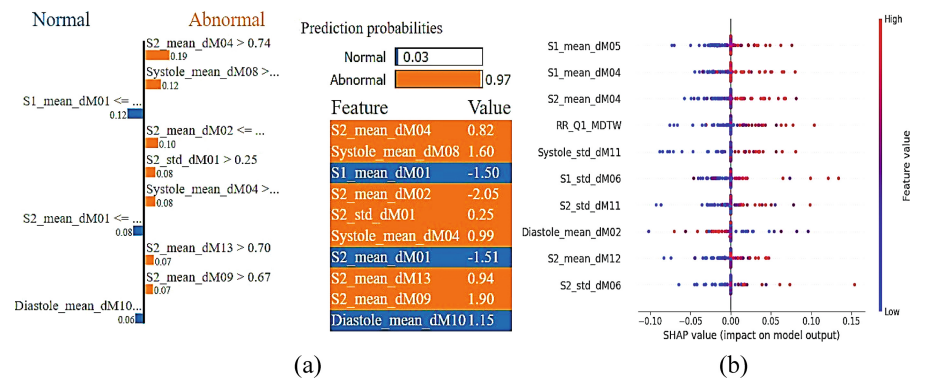


Fig. 2. (a) Lime explanations for Abnormal case. (b) SHAP summary plot showing global feature contributions.

Strengths and Limitations: The primary strength is integrating rigorous hyperparameter optimization with dual XAI frameworks, bridging predictive accuracy with clinical transparency. However, limitations exist: (1) evaluation is constrained to the imbalanced PhysioNet dataset, requiring future cross-dataset validation; (2) the Bayesian model’s severe sensitivity drop restricts it to secondary screening; and (3) the feature reduction suggested by SHAP requires empirical validation through targeted retraining on the reduced subset.

4 Conclusion

This study optimized an LSTM network for heart sound classification, achieving over 89% accuracy and demonstrating robust diagnostic support capabilities. GridSearchCV yielded a balanced clinical profile, whereas Bayesian optimization maximized specificity at the expense of sensitivity. Furthermore, SHAP and LIME frameworks successfully enhanced model transparency by isolating critical physiological markers driving the predictions. Future work includes validating computational savings via SHAP-reduced feature retraining and model compression for real-time edge deployment.

Acknowledgments. We acknowledge support provided by the QRDI Council under Grant ARG01-0522-230274 and HBKU Flagship Signature Research Grant (3rd Cycle) Program.

Disclosure of Interests The authors have no competing interests.

References

1. Finegold, J.A., Asaria, P., Francis, D.P.: Mortality from ischaemic heart disease by country, region, and age: statistics from world health organisation and united nations. *Int. J. Cardiol.* **168**(2), 934–945 (2013)
2. Clifford, G.D., et al.: Recent advances in heart sound analysis. *Physiol. Meas.* **38**, 10–25 (2017)
3. Zeinali, Y., Niaki, S.T.A.: Heart sound classification using signal processing and machine learning algorithms. *Mach. Learn. Appl.* **7**, 100206 (2022)
4. Ren, Z., et al.: Deep attention-based neural networks for explainable heart sound classification. *Mach. Learn. Appl.* **9**, 100322 (2022)
5. Khan, M.U., et al.: Artificial neural network-based cardiovascular disease prediction using spectral features. *Comput. Electr. Eng.* **101**, 108094 (2022)
6. Kamepalli, S., Rao, B.S., Kishore, K.V.K.: Multi-class classification and prediction of heart sounds using stacked LSTM to detect heart sound abnormalities. In: 2022 3rd International Conference for Emerging Technology (INCET), pp. 1–6. IEEE (2022)
7. Ahmad, B., Khan, F.A., Khan, K.N., Khan, M.S.: Automatic classification of heart sounds using long short-term memory. In: 2021 15th International Conference on Open Source Systems and Technologies (ICOSST), pp. 1–6 (2021)
8. Liu, C., et al.: An open access database for the evaluation of heart sound algorithms. *Physiol. Meas.* **37**(12), 2181–2213 (2016)
9. Springer, D.B., Tarassenko, L., Clifford, G.D.: Logistic regression-HSMM-based heart sound segmentation. *IEEE Trans. Biomed. Eng.* **63**(4), 822–832 (2015)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
11. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. *ACM comput. surv. (CSUR)*. **51**(5), 1–42 (2018)
12. Lundberg, Scott, M., Lee, S.I.: A unified approach to interpreting model predictions. *Adv. Neural Inf. Proc. Syst.* **30** (2017)
13. Bozkurt, B., Germanakis, I., Stylianou, Y.: A study of time-frequency features for CNN-based automatic heart sound classification for pathology detection. *Comput. Biol. Med.* **100**, 132–143 (2018)
14. Potes, C., Parvaneh, S., Rahman, A., Conroy, B.: Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. In: 2016 Computing in Cardiology Conference (CinC), pp. 621–624 (2016)
15. Alkhodari, M., Fraiwan, L.: Convolutional and recurrent neural networks for the detection of valvular heart diseases in phonocardiogram recordings. *Comput. Methods Prog. Biomed.* **200**, 105940 (2021)



VAIDYA: Validated Agents for Intelligent Diagnosis and Yielded Analysis

Kalash Shah^(✉), Gautam Bhutani, Rohitaswa Sarbhangia,
and J. Snehan

Fractal AI Research, Mumbai 400063, India
`kalash.shah@fractal.ai`

Abstract. Recent advances in large language models (LLMs) show strong medical reasoning, yet evaluations rely on static cases and multiple-choice questions that miss the complexity and iterative nature of clinical decisions. We propose **DiagBench**, a benchmark where models interact with an LLM-based Patient Simulator to gather clinical details and form diagnoses. We also introduce **MedConvBench**, a diagnostic conversation benchmark to assess the relevance and quality of model reasoning. Finally, we present **VAIDYA**, a modular, medically grounded framework that mirrors a physician’s stepwise diagnostic process.

1 Introduction

Recent LLM advances have improved multi-turn reasoning and medical QA performance, but differential diagnosis (DDx)—which requires iterative hypothesis refinement through targeted questioning—remains underexplored. Effective DDx depends not only on final accuracy but also on asking relevant follow-up questions and adapting reasoning over time, motivating evaluation in realistic conversational settings. Current approaches face three key limitations: reliance on domain-specific training data that risks overfitting, lack of standardized benchmarks for conversational diagnosis, and poor alignment with established medical reasoning processes, limiting interpretability and clinician trust.

To address these gaps, we propose **VAIDYA**, a modular, medically grounded agentic framework that emulates stepwise clinical reasoning. VAIDYA consists of a History Taker for structured information elicitation, a Diagnosis Maker for hypothesis refinement through symptom and etiological analysis, and a Diagnosis Termination module that synthesizes evidence into a final diagnosis. This design enables LLMs to iteratively query patient information and progressively narrow diagnostic possibilities. We also introduce two benchmarks for evaluation. **DiagBench** simulates realistic multi-turn diagnostic scenarios using curated clinical cases and assesses the final diagnostic accuracy. **MedConvBench** evaluates the quality and completeness of model-generated diagnostic questions against clinician-identified gold standards. Empirical results show that VAIDYA significantly outperforms baseline LLMs on DiagBench and MedConvBench.

K. Shah and G. Bhutani—Equal Contribution.

Our main contributions are as follows:

1. We propose VAIDYA, a medically grounded and modular framework that substantially enhances the clinical-reasoning capabilities of LLMs, consistently outperforming the baseline systems.
2. We introduce the Diagnosis Benchmark (DiagBench), comprising of clinically verified diagnostic scenarios designed to assess the diagnostic accuracy.
3. We introduce the Medical Conversation Benchmark (MedConvBench), a clinically validated dialogue-based evaluation suite designed to assess the quality and relevance of questions asked during conversational diagnosis.

2 Methodology

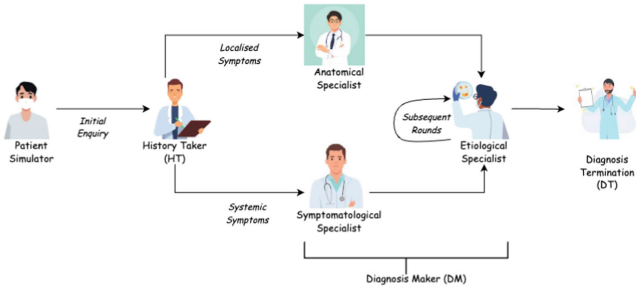


Fig. 1. Overall flow of VAIDYA.

2.1 VAIDYA

Each clinical role in the workflow (Fig. 1) is implemented as an independent LLM *module* with a role-specific system prompt, keeping VAIDYA model-agnostic. The framework consists of three components:

History Taker (HT). The first round gathers a detailed understanding of the patient’s primary symptoms, including onset, severity, and triggers/relievers [6]. Based on the responses, symptoms are classified as *systemic* or *localized*.

Diagnosis Maker (DM). This module performs the differential diagnosis across multiple rounds using three specialists:

1. *Anatomical Specialist:* Localizes the anatomical source of symptoms.
2. *Symptomatological Specialist:* Identifies associated or co-existing symptoms.
3. *Etiological Specialist:* Conducts stepwise reasoning to determine the most plausible underlying cause (e.g., infectious, inflammatory, metabolic).

In the second round, the framework selects the Anatomical or Symptomatological specialist based on symptom classification. All subsequent rounds are handled by the Etiological Specialist. After a total of five rounds, it either concludes the diagnosis or requests further clarification.

Diagnosis Termination (DT). Using the full dialogue history, this module produces the final diagnostic reasoning in a single round.

2.2 Diagnostic Assessment Setup

DiagBench. DiagBench contains 1,000 clinically validated scenarios designed to simulate real-world clinical decision-making. Each case includes a patient profile (age, gender), chief complaint, detailed case study (e.g., co-existing conditions, past history), and the expected diagnosis. During evaluation, the model first sees only the chief complaint, then interacts with the patient through follow-up questions to reach a diagnosis.

We construct DiagBench from three sources: (a) 167 curated patient-physician dialogue cases from Fareez et al. [3]; (b) 500 cases from DDXPlus [2], preserving physician questions and patient responses; and (c) 333 reformulated diagnostic questions from MedQA [5]. All cases were reviewed by 20 licensed physicians (MDs, ≥ 3 years of practice) to remove ambiguity and correct errors.

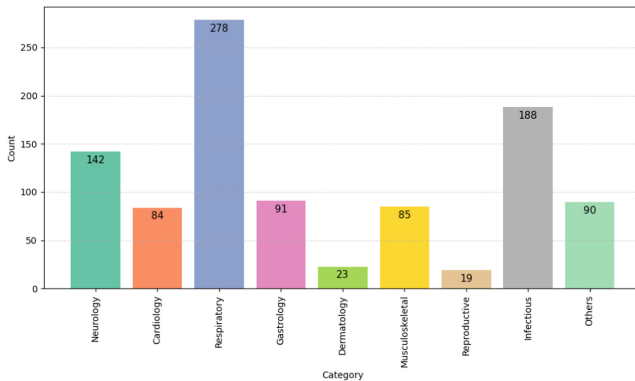


Fig. 2. The distribution of cases across medical domains for DiagBench

Figure 2 shows the distribution of cases across medical domains, highlighting meaningful representation in neurology, gastroenterology, dermatology, and reproductive health.

MedConvBench. MedConvBench comprises the first 500 cases from DiagBench and evaluates whether a model asks the clinically relevant questions

needed to reach the correct diagnosis. For each case, a set of expected questions is defined along with corresponding importance scores.

The cases are drawn from Fareez et al. and DDXPlus, where existing physician questions are treated as *expected questions*. Each case was independently reviewed by three domain-expert clinicians to ensure clarity and clinical validity. Clinicians could also add important questions not present in the original set. Questions were scored as follows:

1. *Score 1*: Critical for establishing or ruling out the diagnosis or determining treatment; omitting it would significantly impair the assessment.
2. *Score 0.5*: Supportive but not strictly necessary for diagnosis or treatment planning.
3. *Score 0*: Irrelevant to diagnosis or treatment; such questions are removed from the benchmark.

3 Experiments

We use an LLM-based patient simulator built from DiagBench vignettes. The simulator provides concise, vignette-faithful responses to model questions. GPT-4o is used to implement the simulator.

3.1 Experimental Settings

Models/Agents. We compare the following two frameworks in our evaluation. (1) **Vanilla LLM**: To ensure a fair comparison with VAIDYA, we implement a vanilla LLM agent that interacts with the Patient Simulator over multiple turns, maintains the full conversation history, and produces a diagnosis based on this context. (2) **VAIDYA**: A model-agnostic framework that follows a multi-specialist architecture as described in Sect. 2.1. Both Vanilla LLM and VAIDYA are benchmarked on **six** off-the-shelf LLMs: Qwen3 8B, Qwen2.5-72B [8], MedGemma 27B [4], Deepseek V3.1 [1], gpt-5, o4-mini [7]

Datasets. We evaluate on two datasets. **DiagBench** assesses final diagnostic accuracy using GPT-4.1: a score of 1 is assigned if the predicted diagnosis is clinically identical to the ground truth or if the core disease is correctly identified with only minor secondary omissions that do not affect prognosis or treatment, while a score of 0 is assigned for incorrect diagnoses or major etiological errors. **MedConvBench**: Let $\mathcal{E} = \{e_1, \dots, e_N\}$ denote the set of expected questions for a case, and let $w_i \in \{0.5, 1\}$ be the weight assigned to e_i . Let $r_i \in \{0, 1\}$ indicate whether a semantically matching agent question is identified for e_i (as evaluated by GPT-4.1). The normalized case score (S) is computed as

$$S = \frac{\sum_{i=1}^N w_i r_i}{\sum_{i=1}^N w_i}.$$

3.2 Results

DiagBench. Table 1 presents a comparison of the diagnostic performance of various LLMs against VAIDYA. VAIDYA (*gpt-5*) achieves the highest diagnostic accuracy at 72.5% showcasing a **21.7%** improvement, outperforming all other models. The effect is pronounced among open-source models as well: Qwen3 8B and MedGemma 27B yield an absolute improvement of 9% and 15.9% respectively. When comparing across frameworks

Table 1. DiagBench Scores

Model	Vanilla	VAIDYA
Qwen3 8B	38.4	47.4
MedGemma 27B	37.5	53.4
Qwen2.5 72B	45.1	51.5
Deepseek V3.1	48.6	52.8
o4-mini	53.8	62.5
gpt-5	50.8	72.5

Table 2. MedConvBench Scores

Model	Vanilla	VAIDYA
Qwen3 8B	0.410	0.512
MedGemma 27B	0.403	0.509
Qwen2.5 72B	0.386	0.557
Deepseek V3.1	0.393	0.582
o4-mini	0.519	0.602
gpt-5	0.526	0.75

MedConvBench. Here the score is averaged over 500 samples and is reported out of 1. Table 2 presents the performance of Vanilla LLMs against VAIDYA. As with DiagBench, VAIDYA (*gpt-5*) achieves the highest score of 0.81, giving a **23.4%** improvement. As evident from Table 2, VAIDYA was able to uplift the performance of all the LLMs with a staggering 21.9% increase for Deepseek V3.1 and 18.9% observed in the case of Qwen 2.5 72B.

4 Conclusion

We presented VAIDYA, a modular framework that enhances clinical reasoning in LLMs by emulating physicians’ step-wise diagnostic process. We introduce two new benchmarks: DiagBench and MedConvBench, which measure the accuracy and clinical alignment of the differential diagnosis produced by conversational medical agents. With the rise of RL-based training methods, these benchmarks offer natural opportunities for outcome and process-based reward signals respectively.

References

1. DeepSeek: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025). <https://arxiv.org/abs/2501.12948>
2. Fansi Tchango, A., Goel, R., Wen, Z., Martel, J., Ghosn, J.: Ddxplus: a new dataset for automatic medical diagnosis. Adv. Neural. Inf. Process. Syst. **35**, 31306–31318 (2022)

3. Fareez, F., Parikh, T., Wavell, C., Shahab, S., Chevalier, M., Good, S., De Blasi, I., Rhouma, R., McMahon, C., Lam, J.P., et al.: A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Sci. Data* **9**(1), 313 (2022)
4. Google-Deepmind: Medgemma technical report (2025). <https://arxiv.org/abs/2507.05201>
5. Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H., Szolovits, P.: What disease does this patient have? a large-scale open domain question answering dataset from medical exams (2020). <https://arxiv.org/abs/2009.13081>
6. Kuriakose, T., Kuriakose, T.: History taking: The most important clinical test. *Clinical Insights and Examination Techniques in Ophthalmology* pp. 21–29 (2020)
7. OpenAI: Gpt-4 technical report (2024). <https://arxiv.org/abs/2303.08774>
8. Qwen: Qwen2.5 technical report (2025). <https://arxiv.org/abs/2412.15115>



Redefining Cardiac Arrhythmia Detection and Monitoring in Athletes Using Artificial Intelligence: A Systematic Review

Rayyan Tamim¹ , Danielle Cruise² , and Mirou Jaana²  

¹ Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, Sherbrooke, QC, Canada

² Telfer School of Management, University of Ottawa, ON, Ottawa, ON, Canada
jaana@telfer.uottawa.ca

Abstract. Given the limited understanding of the clinical validity and impact of artificial intelligence (AI)-enabled tools in sports cardiology, this research presents a systematic review that synthesizes the evidence on AI-based detection / monitoring of arrhythmias in athletes. Following the PRISMA guidelines, we searched five databases (Embase, Medline, CINAHL, Scopus, SportDiscus) using terms related to AI, athletes, and arrhythmias. English-studies, assessing AI for arrhythmias detection/monitoring in athletes were included. Two researchers independently extracted the data (89% agreement); discrepancies were resolved by consensus. Twelve studies (2015–2025) were identified, collectively reflecting relatively low evidence level (8/12 non-randomized, non-comparative designs). Sample sizes ranged from 6 to 535 with ages 7–60 years. All studies assessed arrhythmias using electrocardiograms. The studies focused primarily on endurance athletes, with limited reporting on training intensity, exercise duration, and comorbidity data. Machine learning (5 studies) and deep learning (4 studies) models were mostly used; 66% did not describe model training/validation. 10 studies (83%) reported *improved diagnostic performance* (e.g. higher detection accuracy), and 2 studies (17%) indicated *workflow benefits* (e.g., less time consuming, remote data view), but none reported statistical significance. Evidence on AI-enabled arrhythmias detection in athletes remains early-stage and heterogeneous, limited by non-comparative designs and incomplete reporting of participants and models development/validation. Future studies should employ rigorous designs and assess clinical/statistical significance, patient-centered outcomes, and real-world implementation.

Keywords: Artificial intelligence · Arrhythmias · Athletes · Systematic Review

1 Introduction

While exercise improves cardiovascular health, athletes may be at higher risk for arrhythmias as strenuous exercise can produce this condition [1]. Arrhythmia is defined as an abnormality in heart electrical activity, resulting in tachycardia or bradycardia [2]. It

may present in different forms which vary in causes/severity [1–3]. Failing to detect it can be life-threatening, and misdiagnosing it can lead to unnecessary treatment [1, 2]. As most athletes are often asymptomatic, screening becomes critically important [2]. Existing methods for screening athletes for arrhythmias are limited in terms of their sensitivity/specificity [4]. Internationally, the 12-lead electrocardiogram (ECG) is the most effective screening tool [1, 4]. More recently, artificial intelligence (AI) has been proposed as a method to support the detection/monitoring of arrhythmias given its potential to enable large scale population-based screening, and improve diagnostic accuracy and early detection [5]. A growing body of literature has examined the effectiveness of AI for detecting/monitoring arrhythmias in the general population [5]. Nevertheless, there has been less focus on athletes, despite its potential benefits for this group. Hence, we present the preliminary results of a systematic review that aims to evaluate and synthesize the evidence on AI use for the detection and monitoring of arrhythmias in athletes and addresses the following question: “What is the effectiveness of AI in the detection and monitoring of arrhythmias in athletic populations?”.

2 Methodology

Following the Preferred Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines [6], we developed a search strategy, in consultation with a librarian, covering the domains of AI, athletes and cardiac arrhythmias. We applied it across five databases (Embase, Medline, CINAHL, Scopus, SportDiscus), from journal inception to April 7, 2025. Examples of search terms included “arrhythmias”, “artificial intelligence”, and “athletes”, using Boolean operators (e.g., and/or). Following the Population, Intervention, Comparison, and Outcome (PICO) criteria, eligibility criteria were established: full-text articles, in English, evaluating AI for arrhythmias detection and monitoring in athletes (all ages and performance levels). We excluded articles on participants not performing organized sports/physical activity, not using AI, nor focusing on arrhythmias detection. The search yielded 2,288 articles; 529 duplicates were removed. Titles and abstracts screening produced 46 articles for full text review. Of these, 37 were excluded (16 and 17 not reporting on AI and athletes, respectively). Reference lists of all eligible articles were screened for additional relevant studies; three were added, yielding a total of 12 studies included in this review (Fig. 1). A coding scheme was used to extract relevant information from the studies (e.g., population characteristics, AI use, Impacts of AI models). Two researchers independently extracted the data (89% agreement). Disagreements were resolved by consensus.

3 Results

Most studies used non-randomized intervention designs (no control groups) [3, 4, 7–12], with a few cohort [1, 13], case report [14], and AI-model training/validation studies [2]. Participants’ characteristics varied (e.g., sample sizes: [6–535], participants; mean age: [7–60.4] years). Limited information was presented on comorbidities, sport type/level, exercise duration, and arrhythmia type. Two types of AI models were applied: machine learning (ML) (5 studies), and deep learning (DL) (4 studies); three studies did not

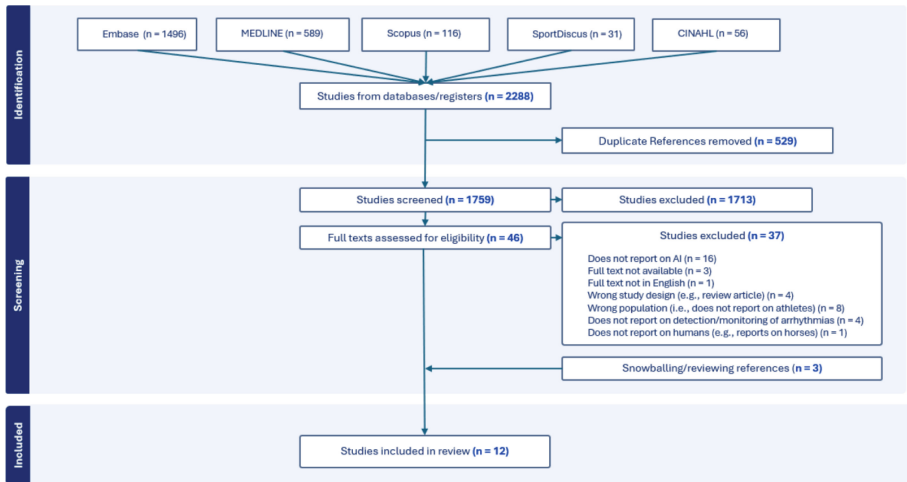


Fig. 1. PRISMA Flow Chart

specify a model type. Of the studies employing ML, nine algorithms were used (e.g., support vector machine (SVM), decision tree, K-nearest neighbours, probabilistic neural network, random forest, and supervised learning). For DL, generative adversarial network (GAN) and artificial neural network (ANN) models were used. The AI models were trained using cross validation techniques [3, 12], dataset split [2, 13], and external validation [12]; 66% did not describe the training used. Comparison across AI models was performed in only two studies [2, 3]. Multiple ML model types were compared, with SVM performing best (e.g., correctly classified 210/246 instances of premature ventricular beats) [3]. GAN performed slightly better than ANN models (i.e., 97.02% vs. 96.10% accurate arrhythmias classification) [2].

Most studies reported *improvements in diagnostic performance* (e.g., improved detection/diagnostic accuracy of arrhythmias in 10 studies [1–4, 7, 9, 10, 12–14]); however, none assessed statistical significance. To confirm diagnostic accuracy, most studies had a physician review the participants' ECG results [1, 3, 4, 13, 14] or assessed accuracy by measuring sensitivity/specificity [2, 10, 12, 13]. Two studies discussed potential false positive results for sports that require upper body movement (e.g., due to motion artefacts) or when athletes had higher heart rates [4, 9]. Only 2 studies reported positive *impacts of AI on clinical workflow*, but remained anecdotal (i.e., reduced time burden and facilitation of remote data access for physicians) [4, 10].

4 Discussion

Research on AI use for detecting/monitoring of arrhythmias in athletes is early stage with varying level of evidence. In line with the literature on the general population [5, 15], our findings show that AI is promising for *improving arrhythmia detection* in athletes and *supporting clinical workflows*. Ten studies suggest positive impacts on detecting/diagnosing arrhythmia in athletes, but statistical and clinical significance lacks. We

recommend that future research apply statistical methods to assess ECG accuracy and validity/reliability of the differences in diagnostic performance between AI-models and human interpreters. Beyond accuracy, assessment of patient outcomes (e.g., hospitalization, mortality) with AI-enabled arrhythmias detection/monitoring can inform planning and preparedness in large sports events and resources optimization.

There was little consistency in model selection/training/validation strategies and reporting standards. Guidelines exist to guide the reporting of AI research (e.g., SPIRIT-AI for intervention trials), which recommend reporting on components like clinical rationale, model training/validation, and ethics and reproducibility [16]. Future studies should follow these guidelines to ensure research quality, validity and reproducibility. We call on peer-reviewed journals to require transparent and standardized reporting for rigor, reproducibility, and research advancement in this area. Last, the wide range of ML algorithms applied suggests lack of consensus on the optimal analytical approaches for the athletic population, underscoring the need for comparative studies to identify the most effective algorithms.

5 Conclusion

While AI is promising for arrhythmias detecting/monitoring in athletes, research limitations persist. To support AI integration into athlete care, future research should prioritize rigorously designed studies assessing clinically meaningful outcomes.

Disclosure of Interests. The authors declare no competing interests relevant to this paper.

References

1. Apelland, T., et al.: Cardiac arrhythmia assessment with patch electrocardiogram versus insertable cardiac monitor: a cohort study in endurance athletes with atrial fibrillation. *BMJ Open*. **15**, e093250 (2025)
2. Zhuang, J., Sun, J., Yuan, G.: Arrhythmia diagnosis of young martial arts athletes based on deep learning for smart medical care. *Neural Comput. Applic.* **35**, 14641–14652 (2021)
3. Varriano, G., et al.: Radiomic analysis to predict arrhythmias in athletes by echocardiography: artificial intelligence vs. formal methods. *Meta-Radiol.* **3**, 100174 (2025)
4. Ausland, Å., Sandberg, E.L., Jortveit, J., Seiler, S.: Heart rhythm assessment in elite endurance athletes: a better method? *Front. Sports Act. Living.* **4**, 937525 (2022)
5. Popat, A., Yadav, S., Obholz, J., Hwang, E.A., Rehman, A.U., Sharma, P.: The efficacy of artificial intelligence in the detection and management of atrial fibrillation. *Cureus* **17** (2025)
6. Page, M.J., et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. **372**, n71 (2021)
7. Castillo-Atoche, A., et al.: Energy efficient framework for a AIoT cardiac arrhythmia detection system wearable during sport. *Appl. Sci.* **12**, 2716 (2022)
8. Chen, W., et al.: Multicenter evaluation of machine-learning continuous pulse rate algorithm on wrist-worn device. *Digit Biomark.* **8**, 218–228 (2024)
9. Gaibazzi, N., Siniscalchi, C., Reverberi, C.: The heart sentinel™ app for detection and automatic alerting in cardiac arrest during outdoor sports: field tests and ventricular fibrillation simulation results. *Int. J. Cardiol.* **269**, 133–138 (2018)

10. Haberman, Z.C., et al.: Wireless smartphone ECG enables large-scale screening in diverse populations. *J. Cardiovasc. Electrophysiol.* **26**, 520–526 (2015)
11. Xiao, Z., Sun, C., Bai, J., Li, X.: Study on cardiovascular monitoring and diagnosis of athletes. *Rev. Bras. Med. Esporte.* **27**, 822–825 (2021)
12. Nechita, L.C., et al.: A resting ECG screening protocol improved with artificial intelligence for the early detection of cardiovascular risk in athletes. *Diagnostics.* **15**, 477 (2025)
13. Munoz-Macho, A.A., Dominguez-Morales, M.J., Sevillano-Ramos, J.L.: Analyzing ECG signals in professional football players using machine learning techniques. *Heliyon.* **10**, e26789 (2024)
14. Peritz, D.C., Howard, A., Ciocca, M., Chung, E.H.: Smartphone ECG aids real time diagnosis of palpitations in the competitive college athlete. *J. Electrocardiol.* **48**, 896–899 (2015)
15. Salvi, M., et al.: Artificial intelligence for atrial fibrillation detection, prediction, and treatment: a systematic review of the last decade (2013–2023). *WIREs Data Min. Knowl. Discov.* **14**, e1530 (2024)
16. Kolbinger, F.R., Veldhuizen, G.P., Zhu, J., Truhn, D., Kather, J.N.: Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. *Commun. Med.* **4**, 71 (2024)



Within-Phenotype Socioeconomic Disparities in Cancer Detection: A Topological Data Analysis of a Population-Based Cohort

Nelofar Kureshi¹ (✉) , Samina Abidi² , Ellen Sweeney³ ,
and Syed Sibte Raza Abidi¹

¹ NICHE Research Group, Faculty of Computer Science, Dalhousie University, Halifax, Canada
nelofar.kureshi@dal.ca

² Department of Community Health and Epidemiology, Dalhousie University, Halifax, NS,
Canada

³ Atlantic PATH, Faculty of Medicine, Dalhousie University, Halifax, Canada

Abstract. Traditional cancer risk prediction models assume uniform associations between risk factors and outcomes, potentially obscuring subgroup heterogeneity. We applied Topological Data Analysis (TDA) to identify latent patient phenotypes and examine whether associations with cancer diagnosis vary across subgroups. Using 47 demographic, socioeconomic (SES), behavioural, health status, and comorbidity features from 30,760 Atlantic PATH cohort participants, TDA identified three phenotypes with distinct cancer prevalence gradients: *Lower SES-High Prevalence* (16.4%), *Moderate SES-Intermediate Prevalence* (12.9%), and *High SES-Low Prevalence* (10.9%). Within-community Firth logistic regression revealed two distinct modes of inequitable detection: in the lower-SES phenotype, detection was concentrated among participants with poor health suggests delayed symptomatic presentation rather than preventive screening. In the high-SES phenotype, lack of formal education was independently associated with reduced detection odds (OR = 0.39 vs. graduate degree), indicating educational disparities persist even where access barriers are less likely to explain them. Population-level models would obscure both patterns.

Keywords: Topological Data Analysis · Clustering · Cancer Risk Screening · Phenotyping · Health Equity · Population Health

1 Introduction

Understanding how demographic, behavioral, and socioeconomic factors relate to cancer outcomes is a central challenge in population health research. Cancer risk factor surveillance increasingly relies on regression-based approaches that assume homogeneous associations within broad population strata, potentially obscuring meaningful subgroups where detection patterns diverge [1].

Topological Data Analysis (TDA) provides a data-driven framework for characterizing population structure without imposing a priori partitions [2]. Unlike conventional

clustering, TDA preserves local similarity and connectivity in high-dimensional data, enabling discovery of latent subgroups and transitional phenotypes [3]. TDA has been applied in population health contexts to uncover clinically meaningful heterogeneity not captured by traditional approaches [2, 4]. We developed a TDA pipeline to identify phenotypically distinct communities within the Atlantic PATH cohort and examine how associations with cancer diagnosis vary across subgroups.

2 Methods

2.1 Data Source and Study Population

Atlantic PATH recruited 35,000+ adults aged 30–74 across four Canadian provinces (2009–2015) using standardized questionnaires covering demographics, health behaviors, medical history, and social determinants [5]. The present analysis used baseline data. The dataset was partitioned into a discovery set (80%) and independent holdout (“lockbox”) set (20%), stratified on cancer diagnosis. All community detection and model development were performed exclusively on the discovery set.

The analytic features comprised 47 predictors spanning demographic, behavioral, health status, and comorbidity domains. All preprocessing was performed on the discovery set.

2.2 Topological Data Analysis and Community Assignment

We applied a TDA pipeline to the discovery set using the Mapper algorithm with a two-dimensional lens constructed from Isolation Forest anomaly scores and L2 norms. Mapper divided the lens space into overlapping hypercubes (grid search: $n_cubes \in \{10–18\}$, $overlap \in \{0.25–0.50\}$), applied DBSCAN within each hypercube, and connected nodes sharing data points. Louvain community detection ($\gamma \in \{0.20–2.00\}$) was applied to the resulting graph. Cancer outcome was excluded from TDA to ensure community structure reflected phenotypic similarity rather than disease status.

Community labels were assigned to discovery set participants based on Mapper node membership. Community centroids were computed as mean feature vectors and frozen. Lockbox participants were assigned to the nearest centroid using Euclidean distance, ensuring community definitions remained fixed with no re-clustering on the holdout set. To examine predictors of cancer prevalence within each phenotype, logistic regression models were fitted separately for each community in the discovery set. The outcome was binary cancer diagnosis (ever diagnosed with cancer: yes/no). Models incorporated regularization and feature selection. Final model performance and coefficient stability were evaluated in the independent lockbox set. Cancer prevalence was summarized across phenotypes using proportions with 95% confidence intervals.

3 Results

3.1 Mapper Graph and Community Structure

Topological analysis of the discovery set (Fig. 1a) identified three distinct communities after grid search (selected configuration: $n_cubes = 18$, $overlap = 0.50$; 173 nodes, 608 edges, $Q = 0.843$). Discovery set: $n = 27,246$; lockbox: $n = 6,821$.

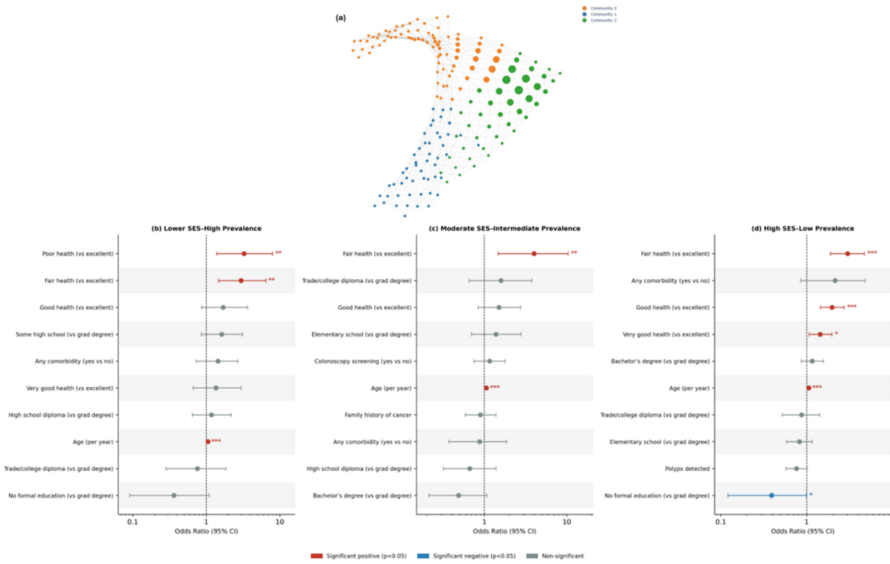


Fig. 1. (a) Mapper graph colored by community. Within-community Firth penalized logistic regression odds ratios (OR, 95% CI, lockbox set) for (a) *Lower SES-High Prevalence*, (b) *Moderate SES-Intermediate Prevalence*, and (c) *High SES-Low Prevalence*. Red = significant positive; blue = significant negative ($p < 0.05$). Reference categories: education = graduate degree; income = $< \$10$ k; health = excellent.

3.2 Phenotype Feature Profiles

Three phenotypes were identified in the holdout set ($N = 6,821$): *Low SES-High Prevalence* (171/1,040 cases; 16.4%, 95% CI: 14.3–18.8%), *Moderate SES-Intermediate Prevalence* (115/892; 12.9%, 95% CI: 10.9–15.3%), and *High SES-Low Prevalence* (534/4,889; 10.9%, 95% CI: 10.1–11.8%).

Within-phenotype logistic regression models revealed distinct predictor profiles across phenotypes (Fig. 1b, c and d). Age and general health status emerged as consistent positive predictors across all communities, with age coefficients ranging from 0.38 (*Moderate SES*) to 0.60 (*High SES*).

In the *Lower SES-High Prevalence Phenotype*, cancer detection was concentrated among participants reporting poor or fair health (poor health OR = 3.27 [1.39, 7.99], fair health OR = 2.97 [1.48, 6.47]). Age was positively associated (OR = 1.05 [1.03, 1.07]) with cancer detection. No formal education showed a negative trend (OR = 0.36 [0.09, 1.10], $p = 0.07$), consistent with under-ascertainment within this already-burdened group.

In the *Moderate SES-Intermediate Prevalence Phenotype*, fair health (OR = 4.01 [1.48, 10.38]) and age (OR = 1.06 [1.03, 1.09]; Fig. 1c) were associated with higher detection odds. Income and screening participation showed positive trends but did not reach significance.

In the *High SES-Low Prevalence Phenotype*, a steep graded health status effect was observed across all levels vs. excellent health (poor OR = 6.04 [1.93, 16.61], fair OR =

3.00 [1.90, 4.70], good OR = 1.98 [1.45, 2.73], very good OR = 1.44 [1.07, 1.97]). No formal education was independently associated with significantly lower detection odds (OR = 0.39 [0.12, 0.99], $p = 0.048$ vs. graduate degree).

4 Discussion

The two communities with the most divergent SES profiles tell a coherent story of inadequate preventive detection. In the *Lower SES-High Prevalence* phenotype, detection concentrated among participants with poor health is consistent with barriers that delay diagnosis until symptom burden demands clinical contact. In the *High SES-Low Prevalence* phenotype, where access barriers are less likely to explain the pattern, educational disparities in detection emerge as an independent finding, suggesting socioeconomic exclusion from screening pathways operates through mechanisms beyond access alone, including health literacy, provider communication, and cultural factors [4, 6]. TDA surfaces both patterns simultaneously by allowing phenotype-specific associations to emerge from population structure rather than imposed boundaries [3].

The most pronounced socioeconomic gradient emerged within the *High SES-Low Prevalence Phenotype*: no formal education was independently associated with significantly reduced detection odds (OR = 0.39), indicating that educational disparities in detection persist even within phenotypes where aggregate SES would suggest otherwise. This finding would be obscured by population-level SES stratification, where this group's overall affluence masks an internal educational disparity in cancer detection.

Strengths of this study include unsupervised phenotype discovery without predetermined clinical categories and out-of-sample validation via nearest-centroid assignment. Within-community regression used Firth penalized estimation with L1 feature selection on the discovery set to ensure stable, unbiased lockbox evaluation. Limitations include cross-sectional community definitions restricting causal inference; negative SES-detection associations most plausibly reflect under-ascertainment, though unmeasured confounding cannot be excluded.

5 Conclusions

TDA revealed two distinct modes of inequitable cancer detection invisible to population-level models: delayed symptomatic presentation in the lower-SES phenotype, and educational exclusion from preventive pathways in the higher-SES phenotype. The steepest socioeconomic gradient—educational attainment independently associated with reduced detection—emerged precisely where aggregate indicators would suggest adequate access. Incorporating phenotypic context into screening prioritization may support more equitable cancer detection strategies.

Acknowledgments. This research is funded by a CIHR catalyst grant. This research has been conducted using Atlantic PATH data under application 2022-103. The data used in this research were made available by the Atlantic Partnership for Tomorrow's Health (Atlantic PATH) study, which is the Atlantic Canada regional component of the Canadian Partnership for Tomorrow's Health funded by the Canadian Partnership Against Cancer and Health Canada. The views expressed

herein represent the views of the authors and do not necessarily represent the views of Health Canada.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Gao, X., Shen, W., Ning, J., Feng, Z., Hu, J.: Addressing patient heterogeneity in disease predictive model development. *Biometrics*. **78**(3), 1045–1055 (2022)
2. Almgren, K., Kim, M., Lee, J.: Extracting knowledge from the geometric shape of social network data using topological data analysis. *Entropy*. **19**(7), 360 (2017)
3. Skaf, Y.: Development and applications of topological data analysis for biomedicine. University of Florida (2023)
4. Kureshi, N., et al.: Utilizing topological clustering on a traumatic brain injury cohort: the association of neighborhood socioeconomic deprivation profiles with injury mortality. In: *Proceedings of the 2024 Australasian Computer Science Week* (2024)
5. Sweeney, E., et al.: Cohort profile: the Atlantic partnership for tomorrow's health (Atlantic PATH) study. *Int. J. Epidemiol.* **46**(6), 1762–1763i (2017)
6. Giurini, L., et al.: Lower neighborhood-level socioeconomic status is associated with lower colorectal cancer screening uptake in the southern community cohort study. *Cancer Prev. Res.* **18**(6), 355–363 (2025)



Synthetic Validation of Pediatric Trust Instruments Using Persona-Driven Large Language Models

Ella Boone¹ , Katya Loban¹ , Elena Guadagno² , and Dan Poenaru^{1,2} 

¹ Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada
ella.boone@mail.mcgill.ca

² Harvey E. Beardmore Division of Pediatric Surgery, The Montreal Children's Hospital,
Montreal, QC, Canada

Abstract. Trust is foundational to patient-physician relationships, yet validated instruments for pediatric emergency and surgical care are lacking and traditional development is time- and resource-intensive. Large language models (LLMs) offer a scalable solution by augmenting expert validation through synthetic panels. We developed two new trust assessment instruments: one for patient families and one for physicians. A distinct two-round instrument content validation was conducted using parallel synthetic and human expert panels. Synthetic panels consisted of persona-prompted LLMs (Claude Sonnet 4, GPT-5, Grok4), human panels served as traditional comparators. Scale-Content Validity Index (S-CVI) and Fleiss' kappa (κ) acceptance thresholds were set at ≥ 0.80 . Following both validation rounds, all instruments exceeded S-CVI ≥ 0.80 . Combined human-synthetic panels revealed substantial inter-rater agreement; κ values for dimensional and contextual validation respectively were: patient-family 0.84 (95% CI [0.72, 0.96]); 0.83 (95% CI [0.73, 0.93]), and physician 0.87 (95% CI [0.72, 1.00]); 0.88 (95% CI [0.80, 0.96]). Overall, persona-prompted LLMs demonstrated comparable validity outcomes to human experts, while accelerating timelines. This synthetic instrument validation offers a scalable blueprint for healthcare measurement development, enabling feasible creation of instruments to support evidence-based patient care.

Keywords: Large Language Models · Synthetic Experts · Pediatrics · Content Validation

1 Introduction

Trust, defined as the patient's acceptance of vulnerability and confidence in physician care, forms the foundation of the patient-physician relationship. It is based on care expectations and the ability to be vulnerable [1]. When established in primary care settings, increased trust is associated with greater care satisfaction, treatment adherence, continuity of care, and willingness to disclose information [2, 3].

In brief settings such as emergency and surgical care trust differs significantly from primary care contexts [4, 5]. While prior studies have adapted adult primary care scales to pediatric settings, none have evaluated trust-building nuances specific to emergency and surgical contexts, representing a critical gap in the literature [1, 4].

Designing validated healthcare instruments is complex and time-intensive, often requiring months to years [6, 7]. Recruiting expert panels for instrument validation in specialized pediatric fields presents additional challenges. Large language models (LLMs) offer a potential solution in this domain where synthetic data are increasingly acceptable for validation, and LLMs can serve as scalable expert panel surrogates through persona prompting [8, 9].

This study develops a methodology for preliminary content validation of trust instruments in pediatric emergency and surgical settings using synthetically generated experts in parallel with human panels. We termed this approach *Synthetic Instrument Validation* (SIV), positioning LLMs as complementary to human experts rather than attempting to replace them, by leveraging synergistic human-AI capabilities [10].

2 Methods

We developed two instruments, one for patient-families and one for physicians, each comprising an experience/perception scale (TRUST-E-F; TRUST-P-MD) and a building blocks scale (TRUST-BB-F; TRUST-BB-MD), measuring trust across four dimensions (competence, caring, communication, and dependability) [1]. Full scales are provided in Supplementary Materials. Content validation was conducted across two rounds: dimensional validation assessing item-dimension alignment, and contextual validation assessing relevance to pediatric emergency and surgical settings using parallel synthetic and human expert panels.

2.1 Instrument Design

TRUST-E-F was adapted from the *Wake Forest Trust in Physician Scale* [11], with TRUST-P-MD developed as its physician mirror; TRUST-BB-F and TRUST-BB-MD were designed to assess essential trust-forming elements.

A multidisciplinary panel (pediatric surgeon, emergency physicians, medical resident, and quantitative researcher) classified each item into one of four trust dimensions per Consensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidelines [12].

2.2 Synthetic Content Validation Rounds

Technical Framework. Three LLMs (Claude Sonnet 4, GPT-5, Grok 4) were persona-prompted twice per expert role (pediatric emergency physician, pediatric surgeon, health services researcher, patient-family representative) [13–15]. Full prompting strategies, and model details are in Supplementary Materials. Item-Content Validity Indices (I-CVI), the proportion of experts rating an item as relevant, was averaged across items to determine Scale-Content Validity Index (S-CVI) (threshold ≥ 0.80), [16] and inter-rater reliability computed via Fleiss' kappa [17, 18].

LLM Selection. Welch's t-tests compared each model's Round 1 I-CVI scores to the human panel; the most concordant model was selected for Round 2. A single calibrated model sufficed in Round 2 given the distinct validation scope.

Dimensional Validation (Round 1). Three human and 30 synthetic experts (10 per LLM) independently and simultaneously rated item-dimension alignment on a 9-point scale (1 = not relevant; 9 = extremely relevant), generating I-CVIs; human experts were blinded to synthetic outputs.

Contextual Validation (Round 2). A distinct round, items were rated on a 9-point scale for contextual relevance to the clinical context by human (7 patient-family, 5 physician) and 16 synthetic experts (10 and 6, respectively), generating I-CVI scores.

3 Results

3.1 Instrument Design

TRUST-E-F contained 12 items (2 caring, 3 competence, 3 communication, 3 dependability, 1 general), TRUST-P-MD 8 (2 per dimension), and both building block instruments, 10 items each (4 caring, 2 per remaining dimension). Review consensus confirmed item alignment across all trust dimensions.

3.2 Synthetic Content Validation Rounds

The synthetic content validation was completed in 4 weeks total, 2 weeks per round where synthetic expert evaluations were completed in 2 h.

S-CVI scores across both validation rounds are presented in Fig. 1. Dimensional and contextual scores respectively were: TRUST-E-F 0.98 (SD = 0.02); 0.94 (SD = 0.09), TRUST-BB-F 0.91 (SD = 0.15); 0.97 (SD = 0.05), TRUST-P-MD 1.00 (SD = 0.00); 0.98 (SD = 0.04), and TRUST-BB-MD 0.91 (SD = 0.17); 0.96 (SD = 0.05). Based on Welch's t-tests, Claude Sonnet 4 showed the lowest variance to human panels across three of four instruments and was selected for contextual validation.

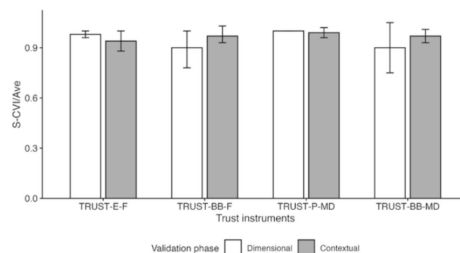


Fig. 1. Scale-Content Validity Index (S-CVI/Ave) of the human-synthetic expert panel across scales in contextual and dimensional validation rounds. Error bars represent standard deviation of Item-Content Validity Index (I-CVI) averages across items.

Substantial agreement was observed for patient-family ($n = 33, 17$; $\kappa = 0.84$ [0.72, 0.96], 0.83 [0.73, 0.93]) and physician ($n = 21, 11$; $\kappa = 0.87$ [0.72, 1.00], 0.88 [0.80, 0.96]) instruments across dimensional and contextual validation respectively.

4 Discussion

This study validated the content of trust instruments for pediatric emergency and surgical contexts using parallel human and synthetic panels through a new approach termed SIV. We have shown that persona-prompted LLMs show reliability as synthetic experts, while accelerating validation timelines from several months to four weeks.

All instruments exceeded $S-CVI \geq 0.80$, achieving above 0.90 across both validation rounds, surpassing content validity benchmarks for panels of five or more experts [16], suggesting appropriate item-dimension alignment and contextual relevance to the settings. Inter-rater reliability was substantial across patient-family ($\kappa = 0.83$ – 0.84) and physician ($\kappa = 0.87$ – 0.88) instruments, consistently exceeding the 0.80 target across diverse experts and validation.

SIV represents a paradigm shift in healthcare instrument development. By positioning synthetic expertise as complementary to human judgment, this methodology preserves critical human oversight while increasing scalability [10]. This approach is grounded in three principles: human-synthetic complementarity, parallel evaluation, and human oversight. The selection of the most human-concordant model for Round 2 reflects this complementary philosophy directly- SIV prioritizes human judgment, positioning synthetic expertise to augment rather than replace expert validation.

Limitations include the rapid evolution of LLM capabilities, meaning evaluated models (September 2025) may not reflect future performance; the untested effectiveness of SIV beyond content validation; single-context design limiting generalizability; small human panel sizes (3–7 experts) that future studies should expand; and reliance on traditional metrics, where alternative thresholds for hybrid panels merit exploration.

5 Conclusions

SIV offers a scalable, rigorous blueprint for healthcare instrument development, demonstrating that human-LLM collaboration can significantly accelerate content validation without compromising methodological standards. Further psychometric evaluation remains ongoing, and its adaptability positions SIV as a broadly applicable framework across clinical contexts.

Notes. 1. Supplementary materials: <https://tinyurl.com/btwk782>

Acknowledgments. The authors thank all participating on their evaluations. Perplexity AI was used for readability editing and figure code debugging; all AI-generated content was reviewed and edited by the authors, who take full responsibility for the manuscript.

References

1. Serhan, O., Moise, A., Guadagno, E., et al.: Exploring trust development in families of children towards surgical and emergency care providers: a scoping review of the literature. *J. Pediatr. Surg.* **58**, 871–878 (2023). <https://doi.org/10.1016/j.jpedsurg.2023.01.022>
2. Wolfensberger, M., Wrigley, A.: *Trust in Medicine: Its Nature, Justification, Significance, and Decline*, 1st edn. Cambridge University Press (2019)
3. Thom, D.H.: Physician behaviors that predict patient trust. *J. Fam. Pract.* **50**, 323–328 (2001)
4. Byczkowski, T.L., Kennebeck, S.S., Grupp-Phelan, J.: A pilot study of a measure of parental trust in pediatric emergency department care teams. *Acad. Pediatr.* **21**, 504–512 (2021). <https://doi.org/10.1016/j.acap.2020.09.005>
5. Regenbogen, S.E., Veenstra, C.M., Hawley, S.T., et al.: The effect of complications on the patient–surgeon relationship after colorectal cancer surgery. *Surgery*. **155**, 841–850 (2014). <https://doi.org/10.1016/j.surg.2013.12.011>
6. Swan, K., Speyer, R., Scharitzer, M., et al.: Measuring what matters in healthcare: a practical guide to psychometric principles and instrument development. *Front. Psychol.* **14** (2023). <https://doi.org/10.3389/fpsyg.2023.1225850>
7. Zamanzadeh, V., Ghahramanian, A., Rassouli, M., et al.: Design and implementation content validity study: development of an instrument for measuring patient-centered communication. *J. Caring Sci.* **4**, 165–178 (2015). <https://doi.org/10.15171/jcs.2015.017>
8. Giuffrè, M., Shung, D.L.: Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digit. Med.* **6**, 186 (2023). <https://doi.org/10.1038/s41746-023-00927-3>
9. Jansen, B.J., Jung, S., Salminen, J.: Employing large language models in survey research. *Nat. Lang. Process J.* **4**, 100020 (2023). <https://doi.org/10.1016/j.nlp.2023.100020>
10. Zhang, H., Wu, C., Xie, J., et al.: Harnessing the power of AI in qualitative research: exploring, using and redesigning ChatGPT. *Comput. Hum. Behav. Artif. Hum.* **4**, 100144 (2025). <https://doi.org/10.1016/j.chbah.2025.100144>
11. Hall, M.A., Zheng, B., Dugan, E., et al.: Measuring patients’ trust in their primary care providers. *Med. Care Res. Rev.* **59**, 293–318 (2002). <https://doi.org/10.1177/1077558702059003004>
12. Mekkink, L.B., Prinsen, C.A.C., Bouter, L.M., et al.: The COnsensus-based standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. *Braz. J. Phys. Ther.* **20**, 105–113 (2016). <https://doi.org/10.1590/bjpt-rbf.2014.0143>
13. Introducing Claude 4. <https://www.anthropic.com/news/claude-4>. Accessed 3 Nov 2025
14. Grok 4: xAI. <https://x.ai/news/grok-4>. Accessed 3 Nov 2025
15. Introducing GPT-5 (2025). <https://openai.com/index/introducing-gpt-5/>. Accessed 3 Nov. 2025
16. Yusoff, M.S.B.: ABC of content validation and content validity index calculation. *Educ. Med. J.* **11**, 49–54 (2019). <https://doi.org/10.21315/eimj2019.11.2.6>
17. Randolph, J. J.: Online kappa calculator (2008)
18. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. *Biometrics*. **33**, 159–174 (1977)



Supporting Autonomy in Dementia Patients with a Cognitive Assistant

Felix Ocker¹(✉), Steffen Heinrich², Alexa von Bosse², Jörg Deigmöller¹,
Pavel Smirnov¹, and Julian Eggert¹

¹ Honda Research Institute Europe, Carl-Legien-Str. 30, 63073
Offenbach am Main, Germany

{felix.ocker,jorg.deigmoller,pavel.smirnov,julian.eggert}@honda-ri.de

² Dementia Competence Centre, Eastern Switzerland University of Applied Sciences,
Rosenbergstr. 59, 9001 St. Gallen, Switzerland
{steffen.heinrich,alex.bosse}@ost.ch

Abstract. Dementia affects over 55 million people worldwide and incurs more than one trillion USD in annual societal costs, due to loss of cognitive autonomy and the need for care. Beyond economics, declining independence severely impacts the psychological well-being and quality of life of patients and their families. Recent advances in artificial intelligence enable cognitive assistants that provide continuous everyday support. This paper makes three contributions. First, we identify concrete use cases in which cognitive assistance can extend functional autonomy for people with mild cognitive impairment, grounded in qualitative interviews with patients, caregivers, and family members. Second, we present a prototypical general-purpose cognitive assistant integrating memory support, task guidance, contextual reminders, and adaptive interaction, and show how it addresses these use cases. Third, we analyze unresolved challenges in dementia care. These findings suggest that cognitive assistants could become a scalable component of dementia care, provided the technical and ethical challenges are addressed.

Keywords: Cognitive Assistant · Support System · Dementia · Mild Cognitive Impairment

1 Motivation

Dementia is a major global health challenge, affecting over 55 million people worldwide, with around 10 million new cases diagnosed each year [5]. The number of people affected by Mild Cognitive Impairments (MCI) is even higher, as it often represents an early stage preceding dementia. Societal costs associated with dementia care exceed one trillion USD annually, driven by the progressive loss of cognitive autonomy and the corresponding increase in care needs. Beyond economic impact, declining independence severely affects the psychological well-being, quality of life, and dignity of patients and their families. Demographic trends, particularly aging populations in developed countries, suggest

these challenges will intensify. Recent advances in Artificial Intelligence (AI), especially Language Models (LMs) and Vision Language Models (VLMs), enable cognitive assistants that provide continuous, personalized everyday support [1]. Such systems have the potential to extend functional autonomy by compensating for cognitive decline, reducing caregiver burden and delaying institutional care. This paper identifies use cases for cognitive assistance grounded in qualitative interviews, presents a prototypical architecture, and discusses open challenges.

2 Related Work

Prior work has identified key requirements for assistive systems supporting people with cognitive impairment. Thordardottir et al. [4] systematically reviewed acceptance and use of assistive technologies among affected individuals and caregivers, highlighting perceived usefulness, ease of use, and low stigmatization as central adoption factors; systems that integrate seamlessly into daily routines and require minimal active engagement show higher acceptance. In parallel, research on memory augmentation has explored external cognitive aids ranging from early systems such as the Memory Glasses¹, which provided contextual memory cues via a head-mounted display, to recent approaches like Memoro [6], which leverages LMs for real-time memory support through a conversational interface. Commercial efforts, e.g., Halliday, Brilliant, and Meta’s Aria glasses, demonstrate growing interest in wearable AI assistants, though they target general consumer use rather than cognitive impairment. At the architectural level, Ocker et al. [3] proposed a grounded memory system for smart personal assistants that combines VLM-based perception, a knowledge-graph memory, and agentic retrieval to maintain a continuous, reality-grounded episodic memory. However, a gap remains between general memory systems, assistive technologies for MCI, and clinical insights, with little integration across these domains.

3 Interview-Based Requirement Clarification

3.1 Recruitment and Interview Setting

Twelve participants were recruited across three groups: primary users with mild cognitive impairment ($n = 5$), relatives as secondary users providing insight into needs and challenges ($n = 3$), and formal caregivers as tertiary users contributing expertise on care processes, safety, and context ($n = 4$). Participants were recruited via network-based outreach (care facilities, phone, flyers) and took part voluntarily. Data was collected through semi-structured, one-on-one interviews conducted in care institutions, participants’ homes, and a university interview room. The interview protocol included (1) clarification of study objectives, (2) exploration of care situations and experiences, (3) mapping experiences to potential use cases, (4) in-depth discussion of identified use cases, and (5) elicitation of

¹ <https://www.media.mit.edu/wearables/mithril/memory-glasses.html>.

requirements, including acceptance and rejection factors. Interviews were analyzed using qualitative content analysis, with transcripts iteratively coded to identify recurring themes, needs, and challenges. Findings were validated within the research team and in collaboration with healthcare professionals, including health scientists, physiotherapists, and registered nurses.

3.2 Application Scenarios

Based on the interviews, we identified six application scenarios that were considered most relevant across all target groups. **Locating misplaced objects** emerged as a common challenge, with participants frequently struggling to find everyday items, e.g., keys, glasses, or medication. This scenario is particularly well suited to memory-based cognitive assistance, as continuous visual perception can track where objects were last seen and provide this information on demand, reducing frustration and search time. **Appointment reminders** were also highlighted, as users often forget medical visits or social engagements. A cognitive assistant can maintain a calendar, issue proactive reminders that account for context and preparation time, and recall prior discussions to provide relevant background. **Stove switch-off** addressed safety concerns related to forgotten appliances and requires smart-home integration, complemented by perceptual hazard detection and alerts to users or caregivers. **Fall detection** was considered especially important for individuals living alone, with vision-based perception enabling the identification of unusual postures or prolonged immobility and triggering timely notifications. **Morning hygiene routines** pose difficulties due to their multi-step nature, where a cognitive assistant could offer gentle reminders and step-by-step guidance. Finally, **nighttime restlessness** was primarily reported by caregivers, who described episodes of nocturnal disorientation that could be mitigated through reassurance and orientation cues.

3.3 Requirements

The interviews revealed several key requirements for cognitive assistants to achieve acceptance and effectiveness. **Ease of use** is critical, as systems must be intuitive and require minimal active engagement, with voice-based interaction preferred over touch or gesture-based interfaces. **Reliability and trust** are equally essential, as false alarms or incorrect information would undermine confidence. Participants emphasized that the system should acknowledge uncertainty rather than provide potentially incorrect responses. **Privacy protection** is a major concern, particularly regarding continuous recording, leading to requests for configurable recording disablement in specific areas and local over cloud-based data processing. **Unobtrusiveness** also matters, with users favoring non-stigmatizing form factors such as stationary devices or integration into everyday objects (e.g., glasses) over specialized wearables. Participants also emphasized the importance of having the option to actively deactivate the device when needed (e.g., during visits in the user's home), in order to avoid stigmatization and to prevent the device from issuing prompts in situations where this

might be undesirable. Finally, **explainability** is necessary for sustained trust, as users want to understand why the system makes specific suggestions to maintain a sense of control.

4 Cognitive Assistant System

The following describes an architecture and implementation of a cognitive assistant, which served as inspiration and as a reference during the interviews. The system builds on prior work [3] and is organized around three streams that operate in parallel, enabling continuous memory evolution and responsive user interaction, cp. Fig. 1.

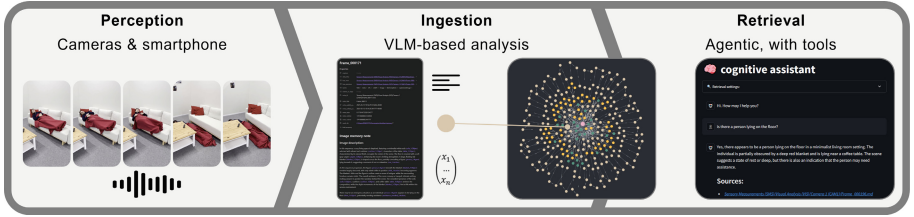


Fig. 1. Overview of the proposed cognitive assistance process.

Perception: The cognitive assistant employs a multimodal perception system comprising live video streams, processed into frames at a configurable rate, synchronized audio snippets, and user interactions via speech or text. Information is extracted through video captioning of ordered frame sequences to capture event structure and automatic transcription of audio recordings.

Ingestion: During ingestion, captions, transcripts, and dialogs are stored as node sequences in a persistent knowledge graph that serves as the system’s single source of truth. From this graph, relevant entities (e.g., persons and objects) and actions are extracted and represented as interconnected concepts. Textual node descriptions are embedded for semantic similarity analyses, and recursive summaries are generated to produce a journal-like memory at multiple levels of detail. These steps are partially parallelized to support real-time ingestion.

Retrieval: Memory access is mediated by a LM-based agent equipped with specialized tools, leveraging natural language as an intuitive interface for querying and contextualizing memory. These tools include semantic search, optionally with temporal filters or graph expansion [2] for contextual enrichment, and structured retrieval via translation of natural-language requests into graph queries. The agent can also set reminders to proactively initiate future interactions.

5 Discussion

5.1 Matching System Capabilities to Identified Scenarios

The cognitive assistant is well-suited to address the scenarios identified. *Locating misplaced objects* and *appointment reminders* benefit from the grounded memory: continuous perception tracks object locations, while the memory component stores appointment information and provides relevant context. *Morning hygiene routines* can be guided step by step using perception to track progress. For *fall detection* and *stove switch-off*, the perception component can detect relevant situations, though full support requires integration with alert systems or smart home infrastructure. Beyond immediate scenarios, continuous perception enables detection of longer-term patterns indicative of disease progression. Internal testing in representative scenarios provided initial qualitative validation, confirming end-to-end feasibility and real-time operation (~ 2 s latency).

5.2 Open Challenges

Based on the interviews, we identified four challenges. First, continuous recording raises **privacy** concerns. Mitigation strategies include local LM deployment, encryption, options to disable recording in sensitive areas or in the presence of others, and storing textual descriptions instead of raw images. Second, **acceptance** depends on ease of use, perceived usefulness, and avoiding stigmatization, with form factor playing a key role. Third, over-reliance on cognitive assistance might accelerate decline by reducing active memory use and lead to **cognitive dependency**. System design should encourage active recall. Fourth, **technical limitations** remain, as current LMs can produce errors or hallucinations. For vulnerable users, graceful handling of uncertainty and escalation to human caregivers are essential.

6 Summary

This paper presented use cases of cognitive assistants for people with MCI, gained from qualitative interviews with patients, caregivers, and family members. These insights are complemented by a technical blueprint integrating memory support, task guidance, contextual reminders, and adaptive interaction. Finally, we discussed key challenges for cognitive assistants in dementia care, forming a basis for further development and systematic evaluation in real-world settings.

References

1. Bommasani, R., et al.: On the opportunities and risks of foundation models. arXiv preprint [arXiv:2108.07258](https://arxiv.org/abs/2108.07258) (2021)
2. Eggert, J., Ocker, F.: Graph-based memory extension for large language models (2025), US Patent App. 18/898,607

3. Ocker, F., Deigmöller, J., Smirnov, P., Eggert, J.: A grounded memory system for smart personal assistants. In: ESWC (2025)
4. Thordardottir, B., Malmgren Fänge, A., Lethin, C., Rodriguez Gatta, D., Chiatti, C.: Acceptance and use of innovative assistive technologies among people with cognitive impairment and their caregivers: A systematic review. *BioMed Res, Int* (2019)
5. WHO: Dementia (2025). <https://www.who.int/news-room/fact-sheets/detail/dementia>
6. Zulfikar, W.D., Chan, S., Maes, P.: Memoro: using large language models to realize a concise interface for real-time memory augmentation. In: CHI (2024)



Bidirectional Floating Feature Selection Guided by Uncertainty Quantification

Marcos López-De-Castro^{1,2} , José González-Gomariz^{1,2} ,
Alberto García-Galindo^{1,2} , Kewei Ni^{3,4} ,
Farnoosh Abbas-Aghababazadeh^{3,4} , Benjamin Haibe-Kains^{3,4,5} ,
and Ruben Armañanzas^{1,2}

¹ Institute of Data Science and Artificial Intelligence, University of Navarra,
Pamplona, Spain

mlopezdecas@unav.es

² TECNUN School of Engineering, University of Navarra, Donostia, Spain

³ Princess Margaret Cancer Centre, University Health Network, Toronto, Canada

⁴ Medical Biophysics, University of Toronto, Toronto, Canada

⁵ Structural Genomics Consortium, Toronto, Canada

Abstract. Reliable uncertainty quantification is critical for trustworthy medical machine learning. We introduce Conformal Bidirectional Floating Search (CBFS), a feature selection method guided by the conformal prediction notion of strangeness to identify subsets that reduce predictive uncertainty. CBFS was evaluated on a synthetic benchmark and on an integrated multi-cohort transcriptomic study of immunotherapy response. It recovered informative features with a true positive rate of 0.737 ± 0.05 on synthetic data, and achieved an average prediction set size of 1.848 on the clinical data. It identified alternative gene signatures overlooked by standard methods, with the resulting signature reaching a C-index of 0.65 (95% CI: 0.51 to 0.80) in external validation. These findings highlight the potential of uncertainty-aware feature selectors for reliable biomarker discovery in high-stakes applications.

Keywords: Feature Selection · Conformal Prediction · Reliable ML

1 Introduction

Feature Selection (FS) aims to find compact subsets of features that maximize predictive performance while minimizing complexity, and is widely used to reveal underlying patterns by identifying the most informative variables. Established FS strategies optimize point-prediction metrics such as accuracy, F1-score, or AUC, which provide limited insight into how feature choices affect predictive uncertainty and risk-sensitive decisions. Individual Uncertainty Quantification (IUQ) has become increasingly important in medical machine learning, where models are expected not only to perform well on average but also to express reliable uncertainty estimates for each individual prediction [1]. In this work,

we introduce a novel Conformal Bidirectional Feature Selector (CBFS) whose goal is to identify relevant feature subsets taking into account their individual predictive uncertainty.

Traditional FS strategies are usually classified into Sequential Forward Selection (SFS), which starts from an empty set and greedily adds the feature that best improves a performance criterion, or Sequential Backward Selection (SBS), which starts from the full set and greedily removes the features that least harm performance. Both are affected by the *nesting effect*: once a feature is added in SFS or removed in SBS, that choice is not revisited, potentially yielding suboptimal subsets. Floating search methods [4] address this by conditionally reversing previous selections when beneficial. Our bidirectional method combines SFS and SBS policies and adds a floating refinement phase, with uncertainty quantification derived through the Conformal Prediction (CP) framework, a distribution-free approach to IUQ with strong theoretical guarantees [5]. We validated CBFS on a synthetic benchmark and on the IMvigor210, IMvigor010, and ABACUS series, three large-scale clinical trials evaluating a shared immunotherapy treatment. CBFS identified novel feature subsets achieving comparable performance to established approaches, corroborating IUQ-guided selection as a candidate policy to find alternative, robust predictive biomarkers.

2 Methods

Preliminaries and Conformal Prediction. We consider feature selection in the standard supervised classification setting with a dataset $\mathcal{D} = \{z_1, \dots, z_n\} \sim P$ of n i.i.d. samples drawn from a distribution P over $\mathcal{X} \times \mathcal{Y}$. Each $z_i = (\mathbf{x}_i, y_i)$ has a d -dimensional feature vector $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and a target label $y_i \in \mathcal{Y} = \{c_1, \dots, c_m\}$. The full feature set is $\mathbf{X} = (X_1, \dots, X_d)$, a candidate subset $\mathcal{S} \subseteq \mathbf{X}$, and $U = \mathbf{X} \setminus \mathcal{S}$ the unselected pool. Conformal Prediction (CP) is a distribution-free framework that returns set-valued predictions with marginal coverage of at least $1 - \epsilon$ for any supervised learner [5], where $\epsilon \in [0, 1]$ is a user-defined nominal error level. Given a non-conformity function $A(z_i, \mathcal{D})$, each sample receives a score α_i measuring how atypical it is with respect to a reference dataset. Data are split into training and calibration sets; for a new sample $\mathbf{x}_{n+1}$ a conformal p -value is obtained for each candidate label $c \in \mathcal{Y}$:

$$p_c = \frac{1 + \sum_{i=1}^{|\mathcal{D}_{\text{cal}}|} \mathbb{1}\{\alpha_i \geq \alpha_{n+1}^c\}}{|\mathcal{D}_{\text{cal}}| + 1}. \quad (1)$$

The final prediction set is $\mathcal{C}(\mathbf{x}_{n+1}) = \{c : p_c \geq \epsilon\}$, ensuring marginal coverage of at least $1 - \epsilon$ on average over future samples. When the non-conformity score is linearly separable, feature-wise contributions β_j can be defined to quantify each feature’s impact on the overall non-conformity [3].

Building Blocks: CRFE and JMI. Leveraging the β -non-conformity decomposition, CRFE [3] is an SBS method that iteratively removes the feature with

the highest β_j , retraining and recomputing the scores at each step. By discarding features that most increase uncertainty, CRFE, with complexity $\mathcal{O}(d^2)$, already reported key results in genomic applications. As a complementary forward criterion, we use the Joint Mutual Information (JMI) method [2], which selects $X_j \in U$ by maximizing a low-order estimate of the conditional mutual information $I(X_j; Y | \mathcal{S})$:

$$\hat{I}(X_j; Y | \mathcal{S}) = I(X_j; Y) - |\mathcal{S}|^{-1} \sum_{k \in \mathcal{S}} I(X_j; X_k) + |\mathcal{S}|^{-1} \sum_{k \in \mathcal{S}} I(X_j; X_k | Y), \quad (2)$$

balancing marginal relevance, redundancy, and class-conditional redundancy with the previously selected features. JMI has complexity $\mathcal{O}(K^2 d)$, where K is the cardinality of the final subset.

Conformal Bidirectional Floating Search. CBFS combines CRFE and JMI in a bidirectional floating search guided by the conformal prediction efficiency score $\mathcal{M}(\mathcal{S})$, the average prediction set size at confidence $1 - \epsilon$ over a held-out evaluation slice. Optimizing $\mathcal{M}(\mathcal{S})$ implicitly optimizes the certainty of predicted labels and is a more stringent objective than plain accuracy. The procedure operates in two stages.

Bidirectional warm-up. Starting from $\mathcal{S} = \emptyset$, $U = \mathbf{X}$, and a discard pool $U^* = \emptyset$, the algorithm alternates JMI inclusions and CRFE exclusions until $|\mathcal{S}| = k$: the top JMI candidate from U is added to $\mathcal{S}$, and the top β_j candidate from U is moved to U^* . This stage seeds $\mathcal{S}$ with informative, low-redundancy features while pruning those that most increase predictive uncertainty. Once $|\mathcal{S}| = k$, the discard pool is reabsorbed ($U \leftarrow U \cup U^*$) so the refinement phase can revisit any feature.

Floating refinement. The baseline score $\mathcal{M}(\mathcal{S})$ is tracked as the *Global Best Score*, with patience initialized to zero. At each iteration, two interim rankings are recomputed on training data: R_{remove} , based on β_j over $\mathcal{S}$, and R_{add} , based on JMI over U . Let f_{remove} and f_{add} be their top elements. Three candidate moves are then evaluated:

$$\mathcal{S}_{minus} = \mathcal{S} \setminus \{f_{remove}\}, \mathcal{S}_{plus} = \mathcal{S} \cup \{f_{add}\}, \mathcal{S}_{swap} = (\mathcal{S} \setminus \{f_{remove}\}) \cup \{f_{add}\}. \quad (3)$$

The move giving the largest improvement over $\mathcal{M}(\mathcal{S})$ is accepted and pools, *Global Best Score*, and patience are updated; otherwise patience is incremented. This Plus-1 Minus-1 design [4] can reverse a previous addition or removal, mitigating the nesting effect of strictly greedy SFS or SBS. The search stops after a fixed number of non-improving iterations or when a maximum iteration budget is reached, returning the best subset recorded. Figure 1 summarizes the procedure.

3 Experimental Setup

Datasets. The synthetic dataset (700 samples, 300 features, 3 balanced classes) was simulated from a Negative Binomial model $\mathcal{NB}(\mu_g, \alpha_g)$ with $\ln(\mu_{ig}) = \ln(L_i) + \log(\pi_g) + \text{mod} + \text{class}$, combining library size, baseline expression, co-expression module effects, and class-specific differential expression restricted to a known set of informative features; α_g was sampled from a log-normal distribution and data were log-normalized before selection. The real-world dataset integrated tumor RNA-seq from the ABACUS, IMvigor010, and IMvigor210 immunotherapy cohorts under a harmonized binary endpoint, Durable Clinical Benefit at 6 months (DCB6), derived from cohort-specific time-to-event variables (RFS, DFS, PFS). DCB6 was *True* for patients remaining disease-free for at least 6 months (606 positive, 363 negative cases; 115 patients censored before 6 months were excluded). The merged data were split into train, calibration, and test sets, with train-only low-expression filtering, library-size normalization with logCPM, train-only variance filtering to the top 3,500 genes, and ComBat batch correction by cohort (parameters estimated on training).

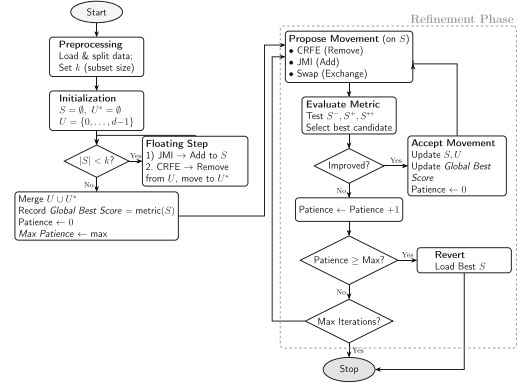


Fig. 1. CBFS Flowchart.

Table 1. FS methods compared on conformal prediction scores (20 runs; 50/30/20% train/calibration/test split; coverage 0.90) and external survival in ABACUS using 10-gene signatures derived from IMvigor210 and IMvigor010 (Cox model: HR per 1-SD increase, p-value; C-index with 95% CI). EC: Empirical Coverage; ASZ: Average Set Size; CER: Certainty.

Method	C.P. Scores			External validation in ABACUS		
	EC	ASZ	CER	C-Index (95% CI)	HR (95% CI)	p-value
CBFS	.916	1.848	.0678	.65 (.51, .80)	.46 (0.19, 1.13)	.089
CRFE	.911	1.831	.0799	.56 (.42, .72)	.72 (0.41, 1.26)	.249
RFE	.903	1.822	.0812	.69 (.57, .80)	.45 (0.24, 0.84)	.012
Lasso	.902	1.836	.0657	.55 (.39, .70)	.83 (0.52, 1.31)	.431
ElasticNet	.901	1.841	.0603	.61 (.47, .74)	.71 (0.45, 1.12)	.146

Evaluation and Experimental Settings. CBFS was evaluated on synthetic data through the True Positive Rate (TPR) of recovered ground-truth features,

and on immunotherapy data against CRFE, RFE, Lasso, and ElasticNet over 20 multi-start runs aiming at compact 10-feature subsets that yield efficient conformal prediction sets. Performance was measured with a conformalized linear SVM ($C = 1$) using the non-conformity and β -non-conformity functions of [3]. In each run, pooled data were split into 50%/30%/20% train/calibration/test with preserved class ratios. Evaluation used empirical coverage, efficiency, and certainty at 90% confidence; accuracy was not considered because forced point predictions can break conformal coverage guarantees. Recursive elimination removed 10% of the feature space per step. External validation used ABACUS as an independent cohort: feature selection was performed on IMvigor210 and IMvigor010, the resulting 10-gene signature was summarized by GSA, and the score was tested on ABACUS with a Cox proportional hazards model.

4 Results, Discussion, and Conclusions

CBFS recovered a substantial proportion of the ground-truth informative features in the synthetic dataset ($\text{TPR} = 0.737 \pm 0.05$), comparable to baseline selectors (see Supplementary Material). In the immunotherapy data, predictive performance was broadly similar across methods (Table 1), but CBFS systematically identified alternative predictor panels (Supplementary Fig. 1-A): Lasso and ElasticNet showed the highest mutual overlap ($69.9 \pm 8.6\%$ mean Jaccard), CRFE and RFE moderate overlap ($20.1 \pm 9.3\%$), whereas CBFS produced largely distinct panels with mean overlap of only 2.0% to 4.1% versus the other selectors. Among its most recurrent genes, MYH3, TMEFF2, FAM3D, HAND2, and NT5E/CD73 were significantly associated with DCB6 (p -values $< 10^{-3}$). In external validation on ABACUS, the CBFS signature showed moderate discrimination ($\text{C-index} = 0.65$, 95% CI: 0.51 to 0.80) and a suggestive Cox signal ($\text{HR} = 0.46$, 95% CI: 0.19 to 1.13, $p = 0.089$). Overall, CBFS matches established selectors in predictive performance while uncovering distinct meaningful signatures, supporting uncertainty-aware feature selection for immunotherapy transcriptomics. Future work should validate these findings using fully nested cross-validation, alternative classifiers, and other non-conformity functions.

Data. The extended manuscript, Supplementary Material, and Python package are available at github.com/digital-medicine-research-group-UNAV/Conformal-BFS.

Acknowledgments. This work was supported by the Gobierno de Navarra (ANDIA 2021, grant 0011-3947-2021-000023; ERA PerMed JTC2022 PORTRAIT, grant 0011-2750-2022-000000). The authors declare no competing interests.

References

1. Banerji, C.R., Chakraborti, T., Harbron, C., MacArthur, B.D.: Clinical AI tools must convey predictive uncertainty for each individual patient. *Nat. Med.* **29**(12), 2996–2998 (2023)
2. Brown, G., Pocock, A., Zhao, M.J., Luján, M.: Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* **13**(2), 27–66 (2012)
3. López-De-Castro, M., García-Galindo, A., Armañanzas, R.: Conformal recursive feature elimination. arXiv preprint [arXiv:2405.19429](https://arxiv.org/abs/2405.19429) (2024)
4. Pudil, P., Novovicova, J., Kittler, J.: Floating search methods in feature selection. *Pattern Recogn. Lett.* **15**, 1119–1125 (1994)
5. Toccaceli, P.: Introduction to conformal predictors. *Pattern Recognit.* **124**, 108507 (2022)



WearAble AI: AI Assistant for Wearable Data Analysis and Visualization

Arash Atrsaei¹(✉) , Haneen Njoum¹ , Albert Pla Planas² ,
and Sven Jager¹

¹ Digital R&D, Sanofi, 65926 Frankfurt, Germany
arash.atrsaei@sanofi.com

² Digital R&D, Sanofi, 08016 Barcelona, Spain

Abstract. Extracting analytics from wearable sensor data requires expert knowledge and coding experience, creating barriers for clinical teams, clinical sites and users. While Large Language Models (LLMs) have limited numerical reasoning, they show promise when augmented with code generation tools. We present a modular LLM architecture for analyzing wearable data across wear compliance, physical activity, and sleep domains. Our multi-stage pipeline classifies user query intent that retrieves relevant documentation, and generates SQL code via a ReAct-style reasoner to analyze processed tabular data. Validation on 252 objective queries demonstrated 94% accuracy. This proof-of-concept demonstrates the feasibility of enabling non-expert users to perform wearable data analysis through natural language interfaces.

Keywords: digital biomarkers · wearables · LLM · talk2data · actigraphy

1 Introduction

Wearable devices are now widely adopted across populations. Augmented with dedicated algorithms, wearable devices provide continuous, real-world evidence of health status through Digital Biomarkers (DBMs) spanning physical activity, sleep, and cardiovascular health. Transforming raw sensor data into clinically meaningful metrics requires complex pipelines, from signal processing to longitudinal aggregation, that typically demand significant technical expertise, creating barriers for clinical and medical teams.

Rapid advances in Large Language Models (LLMs) have opened new possibilities for wearable data interpretation, enabling personalized health recommendations and contextual signal analysis. Although LLMs exhibit limited intrinsic numerical reasoning, they can be augmented with external tools such as code generation [4, 9]. Techniques including few-shot prompting, chain-of-thought reasoning and Retrieval-Augmented Generation (RAG) further enhance performance on complex analytical tasks [5].

SQL and Python code generation allow LLMs to analyze structured tabular data reliably [6]. A notable prior system, PHIA [9], demonstrated ReAct-style agents generating Python code for personal health data analysis. However, PHIA lacked explicit RAG, relying instead on general LLM knowledge or web search, which risks inconsistencies for DBM applications. DBM definitions are inherently mostly inconsistent in the literature and vary across devices and protocols. Its performance on population-level data analysis also remained uncharacterized.

We present WearAble AI, a modular prototype LLM-based chatbot that answers natural language queries on wearable datasets spanning compliance, physical activity, and sleep. By combining intent classification, domain-specific retrieval, SQL code generation with safety checks, and natural language explanation, WearAble AI enables non-expert users to gain analytical insights without coding expertise.

2 Wearable Data Collection

The study was conducted at the Occupational Health Center (Arbeitsmedizinisches Zentrum, AMZ) of Infraseriv GmbH at Frankfurt Hoechst Industriepark, Germany, sponsored by Sanofi-Aventis Deutschland GmbH. The study aimed at evaluating wearable devices in two body positions (wrist and lumbar) for knee osteoarthritis (OA) assessment.

Participants: Two groups were enrolled: (1) healthy adults aged 45–67 years (BMI 20–35 kg/m²), and (2) adults with diagnosed knee OA stratified by KOOS Function subscale score into mild (87.5–62.6) and moderate (62.5–37.6) severity groups. The study was performed as a medical research project according to Article 15 of the Professional Code for Physicians in Germany. All participants provided written informed consent.

Equipment and Protocol: Participants were instructed to wear two EmbracePlus devices, one on the non-dominant wrist continuously during wake and preferably during sleep hours and one at the lumbar position continuously during waking hours over a 28-day assessment period.

Dataset: Data from 24 participants yielded 791 participant-days of measurements. Participants averaged 55.0 ± 5.6 years, 173.6 ± 9.0 cm, and 74.8 ± 13.8 kg. One-minute aggregated parameters were extracted across three domains and further aggregated into daily values, producing three dataframes where each row represents one participant-day: **Compliance:** Wear duration and measurement duration for wrist and lumbar positions. **Physical activity:** Daily steps, normalized step rate, activity intensity summaries (sedentary, light, moderate, vigorous) and per-minute activity intensity arrays. **Sleep:** Sleep onset latency, total sleep time, wake after sleep onset, sleep efficiency, sleep fragmentation, and per-minute sleep stage arrays.

3 WearAble AI Architecture

The overall architecture is shown in Fig. 1. WearAble AI is a modular query-processing pipeline routing user inputs through five components: intent classification, direct response, information retrieval, code generation, and result explanation.

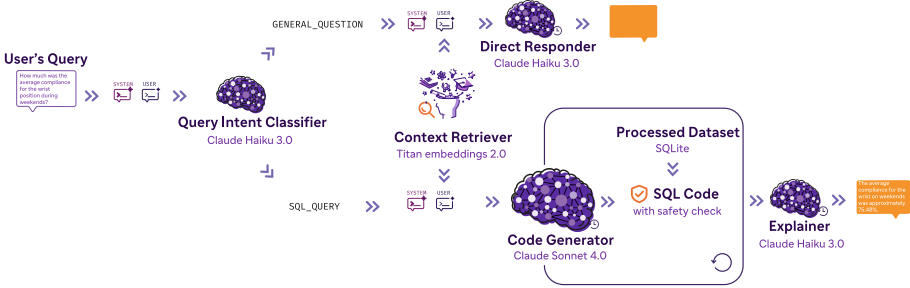


Fig. 1. WearAble AI architecture: Overview of the query-handling pipeline and its five components: intent classification, direct response, information retrieval, code generation, and result explanation.

Query Intent Classifier: Claude Haiku 3.0 [2] classifies queries as general or SQL-required using few-shot examples, intentionally biased toward SQL to avoid false negatives [3].

Direct Responder : For general questions, Claude Haiku 3.0 generates responses using domain knowledge, chat history, and retrieved dataset context.

Information Retrieval : Parameter definitions and SQLite syntax documentation are embedded using Amazon Titan Text Embeddings 2.0 with FAISS indexing [7,8]. Top-k semantic search retrieves relevant context for each query.

Code Generator : Claude Sonnet 4.0 generates SQL code from user queries, retrieved context, and few-shot examples. Safety checks block data modification operations (**INSERT**, **DELETE**, **DROP**, etc.). An iterative ReAct-style error-repair loop [10] retries failed queries with error feedback until success.

Explainer: Claude Haiku 3.0 interprets SQL results and explains findings to users in clear, contextually accurate language.

4 Evaluation

4.1 Query Classification Dataset

The intent classifier was evaluated on 300 questions (100 per domain: compliance, physical activity, sleep), with 50 SQL-required and 50 general questions per domain. A human expert provided reference labels. Sensitivity, specificity, precision, F1, and accuracy were computed per domain and overall.

4.2 Objective Queries Dataset

A total of 252 objective single-value queries were constructed across four categories: Aggregation/Counts, Comparison, Statistical Analysis, and Ranking/Extremes, with reference answers generated by a domain expert using Python scripts. Of these, 120 questions (40 per domain) were used for development and 132 for testing (46 compliance, 43 physical activity, 43 sleep). Accuracy was defined as the percentage of correct answers.

5 Results and Discussions

The classifier achieved 89% overall accuracy with results presented in Table 1. Sensitivity was 100% across all domains, every SQL-required query was correctly identified. Specificity was lower (77%) due to the intentional conservative bias, meaning some general questions were routed through the SQL pipeline. While this increases response time for those queries, it does not degrade answer quality. This over-routing is an acceptable trade-off in clinical contexts where missing analytical queries poses greater risk than slight latency or cost increases.

Table 1. Performance metrics for query intent classification across domains, SEN: sensitivity, SPC: specificity, PPV: positive predictive value or precision, ACC: accuracy

Domain	TP	TN	FP	FN	SEN (%)	SPC (%)	PPV (%)	F1 (%)	ACC (%)
Compliance	50	38	12	0	100	76	81	89	88
Physical activity	50	42	8	0	100	84	86	93	92
Sleep	50	36	14	0	100	72	78	88	86
Overall	150	116	34	0	100	77	82	90	89

Regarding the performance of WearAble AI in responding to the objective queries, WearAble AI achieved 94% overall accuracy on the unseen test set. The sleep domain showed the lowest accuracy across both development (97%) and test sets (91%), attributable to the increased complexity of its parameters and time-series structure. Average end-to-end response times were 3.4 s (compliance), 3.7 s (physical activity), and 3.9 s (sleep), deployed on AWS Bedrock (us-east-1) with a 16-vCPU, 8 GB instance running in-memory SQLite.

Processing 132 test queries consumed 299,785 input and 15,639 output tokens associated with ~ 0.72 USD in total [1] (~ 5.5 USD per 1,000 queries), making the system economically viable for clinical use.

Disabling the iterative error-repair loop marginally increased unanswered questions (3.00% vs. 0% in development), suggesting Claude Sonnet 4.0 with SQL documentation embeddings provides sufficient robustness, with the ReAct loop acting as a safety net rather than a primary performance driver.

6 Conclusion and Limitations

These results demonstrate the feasibility of LLM-based wearable data analysis for non-expert users. The system enables clinicians to quickly summarize complex longitudinal data, generate interpretable metrics, and detect protocol deviations or non-compliance patterns—without requiring coding expertise. The conservative query routing strategy trades precision for safety, which is appropriate for clinical contexts. Our results should be interpreted as strong feasibility evidence on structured queries within a single study context, rather than proof of broad generalization.

This work has several limitations. The evaluation is based on a single study with a single data schema which limits generalization. The results represent feasibility evidence rather than proof of broad applicability. The controlled, single-value query design may not reflect the complexity of real-world open-ended user queries. Domains were kept separate for simplicity; future work should integrate all domains to enable cross-domain analysis (e.g., sleep–activity relationships). SQL generation, while safe and stable, restricts complex statistical analysis and time-series processing common in wearable research. The system currently lacks uncertainty estimation and confidence measures, which are critical for clinical decision support.

The modular architecture, while offering control and independence per component, is less suited for multi-step interactive Q&A than a planner-based agent design.

Declaration of AI Assistance. Microsoft Copilot was used for manuscript editing. The authors reviewed all content and take full responsibility for the manuscript.

References

1. Amazon: Amazon bedrock pricing (2026). <https://aws.amazon.com/bedrock/pricing/>
2. Anthropic: Claude 3 haiku: our fastest model yet (2024). <https://www.anthropic.com/news/claude-3-haiku>
3. Dubiel, M., Barghouti, Y., Kudryavtseva, K., Leiva, L.A.: On-device query intent prediction with lightweight LLMs to support ubiquitous conversations. *Sci. Rep.* **14**(1), 12731 (2024)

4. Englhardt, Z., et al.: From classification to clinical insights: towards analyzing and reasoning about mobile and behavioral health data with large language models. *Proc. ACM Interactive Mobile Wearable Ubiq. Technol.* **8**(2), 1–25 (2024)
5. Gao, Y., et al.: Retrieval-augmented generation for large language models: a survey, **2**(1). arXiv preprint [arXiv:2312.10997](https://arxiv.org/abs/2312.10997) (2023)
6. Jiang, J., Zhou, K., Dong, Z., Ye, K., Zhao, W.X., Wen, J.R.: Structgpt: a general framework for large language model to reason over structured data. arXiv preprint [arXiv:2305.09645](https://arxiv.org/abs/2305.09645) (2023)
7. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. *IEEE Trans. Big Data* **7**(3), 535–547 (2019)
8. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. In: *EMNLP* (1), pp. 6769–6781 (2020)
9. Merrill, M.A., et al.: Transforming wearable data into personal health insights using large language model agents. *Nature Commun.* (2026)
10. Yao, S., et al.: React: synergizing reasoning and acting in language models. In: *The Eleventh International Conference on Learning Representations* (2022)



A Phyto-Ontology for Integrative Therapeutic Discovery

Oresta S. I. Hewryk^{1,2}  and Joseph D. Romano¹  

¹ Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, USA

joseph.romano@pennmedicine.upenn.edu

² Graduate Group in Biology, Department of Biology, University of Pennsylvania, Philadelphia, USA

Abstract. Plant-based therapies (phytotherapies) are widely used but poorly integrated into modern computational discovery pipelines. We present a new phyto-ontology—named POPPY—that models phytotherapies as integrated systems linking medicinal plants, phytochemicals, biological targets, therapeutic contexts, and supporting evidence. Unlike existing resources, the ontology explicitly integrates heterogeneous and fragmented data from numerous open datasets into a formal representation, enabling advanced automated reasoning and hypothesis generation. The resulting framework provides an AI-ready foundation for graph-based machine learning and the discovery of novel therapeutic applications for plant-derived compounds.

Keywords: Phytotherapy · Ontology · Drug Discovery · Machine Learning

1 Introduction

The rise in biological informatics, machine learning (ML), and artificial intelligence (AI) research has necessitated large, standardized databases for effective computational research [1]. Unlike traditional wet lab experiments, which require fewer data points for reliable conclusions, computational methods can analyze vast amounts of data quickly. Ontologies play a crucial role by connecting concepts and categories within a subject area and providing a visual representation of relationships.

Recent ML advancements, such as molecular feature representation and binding affinity prediction, improve the efficiency of small-molecule and drug repurposing research. This approach significantly reduces time and cost compared to traditional methods, which take, on average, 12 years and \$2.6 billion USD to obtain US Food and Drug Administration (USFDA) approval [2].

Phytochemicals, or plant-derived compounds, have been used medicinally due to their bioavailability and large reservoir of bioactive compounds. Iconic examples include aspirin from willow bark, paclitaxel from the Pacific yew tree, and artemisinin from sweet wormwood [3]. However, their research is hindered by fragmented databases and the complexity of biosynthetic pathways. Despite existing resources cataloging

these compounds, they often lack standardization and fail to integrate with computation workflows [4]. This paper introduces a unified phytotherapies ontology and showcases its potential for AI-assisted phytotherapy research.

2 Background and Related Work

Biomedical ontologies have been used since the 1990s to standardize vocabularies, promote data interoperability, and support structured knowledge representation across biological and clinical domains [5]. Existing resources such as CMAUP, HERB, COCONUT, DrugCentral, and ChEMBL provide extensive data on natural products, including phytochemicals, molecular targets, and therapeutic associations. However, these natural databases are typically organized using flat or loosely structured schemas where relationships among plants, phytochemicals, molecular targets, and therapeutic effects remain implicit rather than explicitly modeled [6].

More recent efforts, such as WATRIMed and the OREGANO knowledge graph, aim to integrate plant-based therapeutic knowledge into unified data resources. Here we run into issues of limited geographical coverage (WATRIMed) and focused use on drug repurposing instead of discovery (OREGANO).

Biomedical ontologies, particularly those within the OBO Foundry, have established standards for representing genes, diseases, and phenotypes. However, these resources are typically domain-specific and do not model phytotherapies end-to-end across plants, compounds, targets, therapeutic contexts, and supporting evidence.

Rather than importing multiple partially overlapping ontologies, which would introduce alignment complexity and restrict flexibility, this work reuses standardized identifiers (e.g., UniProt, ATC) while defining a lightweight schema tailored to integrative phytotherapies [2].

3 Ontology Design and Construction

The ontology represents phytotherapies end-to-end by linking traditional medicinal plants to their constituent bioactive compounds, associated molecular targets, and therapeutic indications relevant to human health. Its scope is limited to biomedical and clinical contexts and does not model agricultural practices or non-medicinal plant uses.

Core classes include medicinal plants, phytochemicals, and therapeutic indications (diseases, mechanisms of action, and pathways), human biological entities (genes and proteins), and research evidence (clinical and experimental studies). These entities are connected through mechanistic and evidentiary relations such as *hasCompound*, *targetsGene*, and *hasClinicalStudy*, enabling explicit representation of functional relationships across domains (Fig. 1).

The ontology is designed to support competency questions including:

- “Which bioactive compounds are associated with a given medicinal plant?”
- “What molecular targets are affected by a specific phytochemical?”
- “Which plants or compounds are linked to therapeutic indications for a given disease, and what evidence supports these associations?”

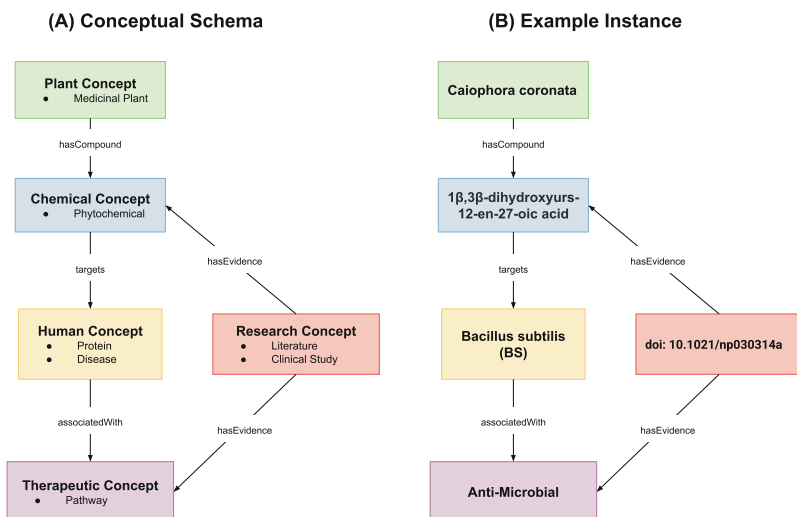


Fig. 1. Ontology design and illustrative instance representation. (A) Core conceptual schema showing the primary classes and relations used to model phytotherapies. (B) Example instance subgraph illustrating how a medicinal plant, bioactive phytochemical, biological target, and therapeutic effect are modeled [4].

3.1 Ontology Design and Representation

3.2 Data Sources and Construction

The ontology integrates data from multiple public resources covering medicinal plants, phytochemicals, molecular targets, and therapeutic indications, including CMAUP, HERB, COCONUT, ChEMBL, and DrugCentral. Construction followed a semi-automated pipeline combining programmatic data integration with manual schema design, guided by competency questions. Source datasets are parsed and normalized using custom scripts (Python, RDFLib) and mapped to RDF via identifier normalization, entity alignment, and relation harmonization.

To ensure interoperability, standardized identifiers and controlled vocabularies are reused wherever possible. Phytochemicals are represented using SMILES, proteins, and genes via UniProt identifiers, and therapeutic concepts are aligned to classification systems such as ATC codes. Duplicate entries are resolved through identifier matching.

The ontology is encoded in OWL/RDF, with concrete entities such as plants, compounds, genes, and pathways modeled as named individuals, and higher-level abstractions as classes. Modeling entities such as pathways as individuals supports instance-level integration and direct linkage to evidence. The ontology adopts open-world assumption, minimizes restrictive axioms, and avoids disjoint assertions where data may be incomplete, enabling lightweight reasoning without compromising data integration.

3.3 Ontology Statistics

The ontology represents plant-derived chemicals, their biological targets, and therapeutic roles. Currently, it contains 28,163,217 RDF triples, integrating across diverse data modality types. The schema includes 22 classes with 19 object properties and 15 data properties and is instantiated with 129,186 individuals. Most individuals fall within chemical and biological categories, reflecting the ontology's focus on phytochemical diversity. While the resource contains instance-level data typical of a knowledge graph, it is structured as an OWL ontology with an explicit schema governing integration and reasoning, functioning as both a schema layer and a populated graph. Complete statistics are available on GitHub (see below).

3.4 Ontology Evaluation

We evaluated the ontology using competency questions and structural consistency checks. For example, SPARQL queries correctly retrieve compounds associated with a given plant via `hasCompound`, molecular targets of phytochemicals via `targetsGene`, and evidence-supported plant-therapy associations through `hasClinicalStudy` and `hasPaper`. Across a random sample of queries, all returned results were structurally consistent with the ontology schema, demonstrating that the competency questions are satisfied. Lastly, reasoning with the FaCT++ reasoner produced no logical inconsistencies.

4 Discussion and Limitations

This ontology provides a novel framework for connecting medicinal plants, phytochemicals, biological targets, and therapeutic activity. Unlike existing databases, which typically organize information in isolated silos, the ontology explicitly models relationships across these domains, enabling reasoning that is difficult to achieve using flat database schemas or feature-based representations. A key strength of the framework is its support for evidence-aware modeling. Supporting literature and clinical studies are represented as first-class entities rather than as unstructured metadata. This design supports transparency and interpretability, which are needed for downstream AI analyses.

The structure of POPPY is well-suited for graph-based machine learning, supporting link prediction and hypothesis generation. In contrast to feature-only approaches that obscure relational context, this representation preserves both semantic meaning and evidentiary grounding. Together, these design choices support applications such as hypothesis generation for novel plant-therapy associations, integration with graph neural networks, and expert-in-the-loop curation workflows, while maintaining a clear separation between curated knowledge and downstream inference.

Several limitations should be noted. The ontology's coverage depends on the availability and quality of curated data sources, which leads to uneven representation across plants, compounds, and therapeutic areas. It is susceptible to literature bias toward well-studied plants and compounds, and sparse or uneven evidence across therapeutic areas. Evidence supporting associations varies widely in experimental rigor and clinical relevance and is represented without weighting or confidence scoring. In addition,

the ontology does not currently model dosage, pharmacokinetics, or toxicity, as these require context-specific and often inconsistent data. Excluding them at this stage allows for a stable, extensible core schema, with future work incorporating these dimensions once standardized representations are available. Finally, while the ontology supports hypothesis generation, it does not itself assert novel therapeutic claims.

5 Conclusion and Future Work

We introduce an ontology that models phytotherapies as an integrated system connecting plants, phytochemicals, biological targets, therapeutic contexts, and supporting evidence. By representing these relationships, the framework overcomes key limitations of existing flat or siloed resources. The ontology provides a reusable, evidence-aware foundation for integrative analysis and graph-based discovery in phytotherapy research. Next, we will extend the ontology with multi-omics data and release a website for interactive exploration and community use. Future work will expand the ontology through integration with established biomedical ontologies. All code used is available on GitHub at <https://github.com/RomanoLab/poppy>.

This work is supported by US National Institutes of Health grants R00-LM013646 (PI: Romano) and P30-ES013508 (PI: Trevor Penning).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Stevens, R.: Ontology-based knowledge representation for bioinformatics. *Brief. Bioinform.* **1**, 398–414 (2000). <https://doi.org/10.1093/bib/1.4.398>
2. Deng, J., Yang, Z., Ojima, I., Samaras, D., Wang, F.: Artificial intelligence in drug discovery: applications and techniques. *Brief. Bioinform.* **23**, bbab430 (2022). <https://doi.org/10.1093/bib/bbab430>
3. Li, Z., Chen, K., Rose, P., Zhu, Y.Z.: Natural products in drug discovery and development: synthesis and medicinal perspective of leonurine. *Front. Chem.* **10**, 1036329 (2022). <https://doi.org/10.3389/fchem.2022.1036329>
4. Romano, J.D., Tatonetti, N.P.: Informatics and computational methods in natural product drug discovery: a review and perspectives. *Front. Genet.* **10**, 368 (2019). <https://doi.org/10.3389/fgene.2019.00368>
5. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004). <https://doi.org/10.1093/nar/gkh061>
6. Guo, L., et al.: Natural products of medicinal plants: biosynthesis and bioengineering in post-genomic era. *Hortic. Res.* **9**, uhac223 (2022). <https://doi.org/10.1093/hr/uhac223>



Evaluating Temporal Clinical Reasoning and Audience-Tailored Explanations in Large Language Models

Gianluca Apriceno^(✉), Tania Bailoni^(✉), and Mauro Dragoni^(✉)

Digital Health and Wellbeing Research Center, Fondazione Bruno Kessler,
Trento, Italy
{apriceno,tbailoni,dragoni}@fbk.eu

Abstract. Large Language Models are increasingly used in clinical tasks, yet their ability to reason over temporally dependent rules and generate audience-tailored explanations remains underexplored. We evaluate seven recent LLMs on five real-world patient scenarios, comparing structured (Turtle) and unstructured (natural language) representations of clinical knowledge. Performance is assessed in terms of guideline violation detection accuracy and explanation quality, based on expert evaluation of correctness, clarity, temporal traceability, and reasoning alignment. Results show that LLMs perform more reliably with unstructured prompts, while structured formats often lead to misinterpretations, particularly in temporally complex scenarios. Although LLMs generate coherent explanations, their consistency and audience alignment vary, underscoring the need for expert validation. These findings highlight key limitations and opportunities for temporally aware clinical decision support.

GitHub: <https://github.com/IDA-FBK/LLM-Benchmarking-Temporal-Clinical-Reasoning>.

Keywords: Large Language Models · Temporal Reasoning · Natural Language Explanations

1 Introduction

Large Language Models (LLMs) have shown strong capabilities across domains, including societal and biomedical applications [10], as well as in reasoning and explanation generation in areas such as recommendation systems and healthcare [3, 5, 9]. However, prior work largely focuses on static or single-step settings, with limited attention to temporally dependent reasoning or multi-stakeholder explanations. In healthcare, models are often task-specific and rarely account for sequences of patient events or evolving clinical rules. Moreover, recent evidence shows that even advanced LLMs struggle with temporal reasoning [7]. To address these gaps, we evaluate seven recent LLMs on five real-world clinical

scenarios from the Trentino Salute 4.0 (TS4.0) platform¹, comparing structured (Turtle-based) and unstructured (natural language) representations of clinical knowledge. We assess (i) violation detection accuracy and (ii) explanation quality for patients and physicians, based on expert evaluation of correctness, clarity, temporal traceability and reasoning alignment. We investigate two research questions: (**RQ₁**) How accurately recent LLMs detect rule violations under temporally dependent clinical rules across different knowledge representations? (**RQ₂**) To what extent can LLMs generate consistent audience-tailored explanations for patients and physicians in both structured and unstructured settings, while reasoning over temporally dependent clinical events?

2 Methodology

We investigate LLM capabilities in two aspects: (i) **reasoning** over temporally dependent clinical knowledge to detect and justify rule violations, and (ii) **generating** audience-specific explanations for patients and physicians. We evaluate these capabilities on real-world scenarios from a virtual coaching platform in the TS4.0 project, which monitors users’ health behaviors and adherence to clinical guidelines over time. To guide the reasoning process, we design a prompt composed of four components: (i) **Task description**, where the model is required to identify whether violations occur, report their number, and provide a justification of its reasoning. It also generates two explanations: a patient version in non-technical language and a physician version using domain terminology; (ii) **Background knowledge**, including monitoring rules defining acceptable behavior over time (e.g., dietary constraints) and a violation framework specifying thresholds and temporal patterns for detecting breaches. Both are derived from the Functional Status Knowledge Graph (FuS-KG) [8], which models inter-related dimensions of user health and functional status; (iii) **Input data**, including patient-specific information such as diagnoses, dietary logs, and physiological or behavioral measurements, used as evidence for rule evaluation; and (iv) **Temporal information**, comprising time-stamped data to support reasoning over event sequences.

3 Experimental Setting

LLMs were evaluated in a controlled setup comprising five clinical scenarios, prompt designs, and a standardized evaluation protocol. The scenarios, derived from a virtual coaching platform in the TS4.0 project, represent temporally dependent dietary monitoring conditions: **Use Case 1 (No Violation)**, where intake meets but does not exceed thresholds; **Use Case 2 (Single Meal Violation)**, where a meal exceeds a dietary limit; **Use Case 3 (Multiple Meal Violations)**, where multiple constraints are violated within a meal; **Use Case 4 (Single Weekly Violation)**, where weekly intake exceeds a

¹ <https://trentinosalutedigitale.com/blog/portfolio/trentinosalute/>.

threshold; and **Use Case 5 (Multiple Weekly Violations)**, where multiple weekly limits are exceeded across two different food categories. Each scenario encodes realistic patient profiles with cardiovascular and/or metabolic conditions and includes temporally grounded rules over daily or weekly aggregation. Scenarios are represented in both structured (Turtle) and unstructured (natural language) formats, following FuS-KG guidelines and TS4.0 data models. We evaluate seven LLMs (Deepseek-v3.2-exp, Gemini-3-pro, Llama-4, Mistral-small-3.1-24b-instruct, Gpt-oss-120B, Qwen3-vl-30b-a3b-thinking, and Grok-4.1-fast) under both prompt types. Experiments span temperature settings $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$, with higher values increasing output variability, and are executed via the OpenRouter API with enforced structured JSON outputs. Performance is assessed using (i) **violation detection accuracy**, measuring correct identification and counting of rule violations, and (ii) **explanation quality**, evaluated by two domain experts (nutrition specialists) considering correctness, clarity, coherence, and consistency, including the use of temporal and clinical evidence.

4 Results

4.1 Violation Detection

Table 1 reports model outputs across structured and unstructured prompts and increasing temperature settings. Correct predictions (decision and violation count) are shown in green, incorrect outputs in red, and non-binary responses (OTH) in orange. Overall, LLMs perform substantially better on unstructured prompts than structured ones, suggesting sensitivity to input representation. A consistent failure occurs in structured Use Case 1, where most models fail to recognize the absence of violations. This is largely due to ambiguity in Turtle-based rule encoding, where relational operators (e.g., Greater, Less) are not semantically grounded as minimum or maximum constraints. In contrast, unstructured formulations explicitly specify thresholds, leading to near-perfect performance. Errors are concentrated in Llama-4-maverick and Mistral-small-3.1-24b-instruct, with the former showing reasoning inconsistencies and the latter occasionally producing incorrect binary outputs despite identifying the correct number of violations. Across models, error rates increase with longer temporal horizons, particularly in weekly aggregation scenarios, where misclassification and violation miscounting become more frequent.

4.2 Explanations Analysis

Expert evaluation (by two nutritionists) assessed explanations for correctness, clarity, traceability, and alignment with decisions for both patient and physician oriented perspectives. Across Use Cases 1–3, patient explanations varied in tone: some models (e.g., Llama-4-maverick) produced neutral responses, while others (e.g., Deepseek-v3.2-exp, Gemini-3-pro) used more empathetic and motivational language. Physician explanations were generally more structured, but

Table 1. LLM decisions and violation counts across structured and unstructured use cases at increasing temperature values. Green indicates correct decision and violation count, red indicates errors, and orange represents non-binary responses.

Model	Use case	Structured						Unstructured					
		0.0	0.2	0.4	0.6	0.8	1.0	0.0	0.2	0.4	0.6	0.8	1.0
Deepseek-v3.2-exp	UC1	NO/0	YES/1	YES/1	YES/1	YES/1	YES/1	NO/0	NO/0	NO/0	NO/0	NO/0	NO/0
Gemini-3-pro	UC1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	NO/0	NO/0	NO/0	NO/0	NO/0	NO/0
Gpt-oss-120b	UC1	YES/1	YES/1	YES/1	NO/0	NO/0	NO/0	NO/0	NO/0	NO/0	NO/0	NO/0	NO/0
Grok-4.1-fast	UC1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	NO/0	NO/0	NO/0	NO/0	NO/0	NO/0
Llama-4-maverick	UC1	YES/1	YES/1	YES/1	OTH/1	YES/1	OTH/1	NO/0	NO/0	OTH/1	NO/0	NO/0	YES/1
Mistral-small-3.1-24b-instruct	UC1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/0	YES/0	NO/0	YES/0	YES/0	YES/0
Qwen-vl-30b-a3b-thinking	UC1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	NO/0	NO/0	NO/0	NO/0	NO/0	NO/0
Deepseek-v3.2-exp	UC2	YES/1	YES/1	YES/1	NO/0	YES/1	NO/0	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1
Gemini-3-pro	UC2	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1
Gpt-oss-120b	UC2	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1
Grok-4.1-fast	UC2	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1
Llama-4-maverick	UC2	OTH/1	OTH/1	YES/1	YES/1	OTH/2	YES/1	YES/1	OTH/1	OTH/0	YES/1	YES/1	YES/1
Mistral-small-3.1-24b-instruct	UC2	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1
Qwen-vl-30b-a3b-thinking	UC2	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1
Deepseek-v3.2-exp	UC3	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2
Gemini-3-pro	UC3	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2
Gpt-oss-120b	UC3	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2
Grok-4.1-fast	UC3	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2
Llama-4-maverick	UC3	OTH/0	OTH/0	YES/2	YES/2	YES/2	YES/2	YES/2	OTH/0	YES/2	YES/2	OTH/1	YES/2
Mistral-small-3.1-24b-instruct	UC3	YES/2	YES/2	YES/2	YES/2	YES/1	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2
Qwen-vl-30b-a3b-thinking	UC3	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2
Deepseek-v3.2-exp	UC4	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1
Gemini-3-pro	UC4	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1
Gpt-oss-120b	UC4	YES/1	YES/1	YES/1	OTH/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1
Grok-4.1-fast	UC4	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1
Llama-4-maverick	UC4	YES/1	OTH/4	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	OTH/7	YES/1	OTH/7	OTH/1
Mistral-small-3.1-24b-instruct	UC4	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/1	YES/1	OTH/1	YES/1	YES/1	YES/1
Qwen-vl-30b-a3b-thinking	UC4	YES/1	YES/1	NO/0	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1	YES/1
Deepseek-v3.2-exp	UC5	YES/1	YES/2	YES/1	YES/2	YES/1	YES/1	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2
Gemini-3-pro	UC5	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2
Gpt-oss-120b	UC5	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2
Grok-4.1-fast	UC5	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2
Llama-4-maverick	UC5	OTH/0	OTH/0	YES/2	YES/2	YES/1	YES/2	YES/2	YES/2	YES/2	OTH/0	OTH/0	OTH/0
Mistral-small-3.1-24b-instruct	UC5	NO/0	YES/1	NO/0	NO/0	OTH/1	YES/3	YES/2	YES/2	YES/1	YES/1	YES/1	YES/2
Qwen-vl-30b-a3b-thinking	UC5	YES/2	YES/1	YES/2	YES/1	YES/2	YES/1	YES/2	YES/2	YES/2	YES/2	YES/2	YES/2

only Grok-4.1-fast consistently provided strong traceability, explicitly linking decisions to patient attributes and knowledge graph entities. A consistent pattern is the difference in temporal granularity across audiences: physician explanations include more explicit timestamps, while patient explanations are more generic. In weekly scenarios (Use Cases 4–5), most models correctly aggregated temporal information but shifted toward more abstract explanations. Some models introduced spurious constraints or showed inconsistencies between predicted decisions and explanation content. Grok-4.1-fast remained the most temporally detailed, often preserving event-level references even in aggregated settings, though occasional language switching (Italian) was observed without affecting correctness. Overall, experts found that unstructured prompts lead to more coherent and aligned explanations, consistent with the performance trends in violation detection. However, variability in temporal detail and cross-audience consistency high-

lights limitations in current LLMs for reliable temporally aware clinical explanation generation.

5 Conclusions and Future Works

We evaluated seven recent LLMs on temporally dependent clinical reasoning and audience-tailored explanation generation using structured (Turtle) and unstructured clinical knowledge. The study addressed two research questions: (i) violation detection accuracy (**RQ₁**) and (ii) the quality and consistency of audience-specific explanations for patients and physicians (**RQ₂**). For **RQ₁**, LLMs perform more reliably with unstructured inputs, while structured Turtle representations often lead to misinterpretations, particularly for threshold-based rules and longer temporal dependencies. Errors increase with temporal complexity and are concentrated in a subset of models. For **RQ₂**, expert evaluation shows that explanations are more coherent and guideline-aligned under unstructured prompts. Physician explanations include more explicit temporal references, whereas patient explanations are more abstract. Importantly, models may produce plausible explanations even when underlying decisions are incorrect, highlighting risks in relying on explanations alone. Overall, LLMs show potential as mediators in patient–physician communication, but their reliability in temporally grounded clinical reasoning remains limited. Prompt design, temporal robustness, and expert validation are key factors influencing performance. Future work should improve structured knowledge handling, strengthen temporal reasoning, and better align decision accuracy with explanation quality in clinical settings. In addition, integrating LLMs with hybrid neuro-symbolic approaches [2,4,6] and multi-modal [1] clinical data (e.g., longitudinal records, signals, and imaging) may enhance robustness and support more reliable, context-aware reasoning in real-world settings.

Acknowledgments. “Hub Life Science – Digital Health (LSH-DH) PNC-E3-2022-23683267 - Progetto DHEAL-COM – CUP: C63C22001970001”, finanziato dal Ministero della Salute nell’ambito del Piano Nazionale Complementare Ecosistema Innovativo della Salute” - Codice univoco investimento: PNC-E.3. This publication reflects only the authors’ view and the Italian Ministry of Health is not responsible for any use that may be made of the information it contains.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Apriceno, G., Tamma, V., Bailoni, T., de Berardinis, J., Dragoni, M.: A pattern to align them all: integrating different modalities to define multi-modal entities. CoRR abs/2410.13803 (2024)
2. Apriceno, G., et al.: End-to-end neuro-symbolic approaches for event recognition (2023)

3. Apriceno, G., et al.: Exploring large language model reasoning capabilities over personal health data. In: Bellazzi, R., Juarez Herrero, J.M., Sacchi, L., Zupan, B. (eds.) AIME 2025, Part II. LNCS, vol. 15735, pp. 12–17. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-95841-0_3
4. Apriceno, G., et al.: A trustworthy evolutionary fuzzy neural network framework for maternal health risk classification. In: Bellazzi, R., Juarez Herrero, J.M., Sacchi, L., Zupan, B. (eds.) AIME 2025, Part II. LNCS, vol. 15735, pp. 18–23. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-95841-0_4
5. Bismay, M., et al.: Reasoningrec: bridging personalized recommendations and human-interpretable explanations through LLM reasoning. CoRR abs/2410.23180 (2024)
6. de Campos Souza, P.V., et al.: Enhancing logical tensor networks: integrating uninorm-based fuzzy operators for complex reasoning. In: Besold, T.R., d’Avila Garcez, A., Jimenez-Ruiz, E., Confalonieri, R., Madhyastha, P., Wagner, B. (eds.) NeSy 2024, Part II. LNCS, vol. 14980, pp. 68–79. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-71170-1_8
7. Chu, Z., et al.: TimeBench: a comprehensive evaluation of temporal reasoning abilities in large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, pp. 1204–1228. Association for Computational Linguistics (2024)
8. Dragoni, M., et al.: Validating a functional status knowledge graph in a large-scale living lab. In: Alam, M., Rospocher, M., van Erp, M., Hollink, L., Gesese, G.A. (eds.) EKAW 2024. LNCS, vol. 15370. LNCS, vol. 15370, pp. 416–433. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-77792-9_25
9. Yu, Q., et al.: Health-LLM: personalized retrieval-augmented disease prediction system (2025)
10. Zhang, Q., et al.: Scientific large language models: a survey on biological & chemical domains (2024)



Seeing the Wood for the Trees: Rethinking AI Ethics Beyond Anthropocentrism

Bianca Lerma^{1(✉)}, Gianluca Apriceno^{2(✉)}, and Mauro Dragoni^{2(✉)}

¹ University of Milano, Milano, Italy
biancalerma99@gmail.com

² Digital Health and Wellbeing Research Center, Fondazione Bruno Kessler,
Trento, Italy
{apriceno,dragoni}@fbk.eu

Abstract. Artificial Intelligence is increasingly used in healthcare for clinical decision-making and personalized recommendations. However, most systems rely on an implicitly anthropocentric ethical framework that prioritizes short-term, patient-centered outcomes while overlooking long-term, systemic, and multi-agent consequences of clinical decisions. This paper introduces a healthcare-oriented instantiation of *NAEL (Non-Anthropocentric Ethical Logic)*, a computational framework that models ethics as an adaptive inferential process rather than static rules or external constraints. Grounded in active inference, NAEL frames ethical decision-making as the minimization of global expected free energy, extending agent-centric formulations to account for uncertainty induced across other agents and the environment. Ethical considerations are embedded directly within the agent's generative model and policy evaluation. Through a conceptual ICU resource allocation case study, we show how NAEL-driven policies differ from conventional short-term optimization approaches, highlighting their potential for clinically informed and ethically coherent decision-making in complex healthcare settings.

Keywords: Healthcare Decision Support Systems · Medical AI · Non-Anthropocentric Ethics · Active Inference · Ethical Logic · Neuro-Symbolic AI

1 Introduction

Artificial Intelligence (AI) systems are increasingly embedded in healthcare, supporting clinical decision-making, patient monitoring, prevention strategies, and personalized recommendations [5, 16]. Most current deployments rely on an implicitly anthropocentric ethical paradigm, optimizing short-term, human-centered objectives such as immediate clinical outcomes and individual patient preferences [13]. While effective in many settings, this perspective is limited when healthcare decisions are embedded in broader socio-ecological systems, where outcomes unfold across temporal, environmental, and collective dimensions. This limitation reflects a deeper conceptual issue in current AI ethics

for healthcare, i.e., a focus on individual human preferences that neglects the relational and multi-scale nature of health [15]. Health outcomes emerge from interactions among individuals, social systems, technologies, and ecological environments, requiring decision-making that accounts for both short-term clinical goals and long-term collective consequences. Addressing this challenge calls for a shift toward adaptive, non-anthropocentric ethical reasoning capable of integrating multi-agent and multi-scale considerations [12]. To this end, we introduce a healthcare-oriented instantiation of *NAEL* (*Non-Anthropocentric Ethical Logic*) [11], a framework that models ethics as a context-sensitive inferential process rather than a fixed set of rules or utility functions. Grounded in Active Inference, NAEL reframes ethical behavior as the minimization of global expected uncertainty, extending standard agent-centric formulations to account for broader environmental and multi-agent effects [14]. In this work, we operationalize NAEL in a healthcare decision-support agent that generates recommendations considering both immediate clinical indicators and longer-term systemic factors aligned with sustainability objectives such as the United Nations ESG goals [17]. Unlike conventional approaches, sustainability considerations are not treated as external constraints but are directly embedded within the agent’s generative model of the world.

2 Related Work

Formal logic has long been used for ethical reasoning in artificial systems, particularly where decisions require transparency and normative justification. Deontic and argumentation-based approaches enable explicit representation of obligations, permissions, and conflicts under uncertainty [8]. However, purely symbolic methods struggle in healthcare settings due to noisy data, probabilistic outcomes, and evolving clinical knowledge. Neuro-symbolic AI addresses these limitations by combining neural representation learning with symbolic reasoning, and has been applied in different applications [1–4, 6, 7, 9]. While these approaches improve robustness and interpretability, their ethical reasoning is typically encoded as static constraints over short-term clinical objectives. Active inference provides a complementary perspective by modeling perception and action as minimization of expected free energy under uncertainty [10], with hierarchical extensions enabling multi-level decision-making relevant to healthcare [14]. However, existing formulations remain largely agent-centric, focusing on individual outcomes without explicitly modeling uncertainty propagated to other agents or broader socio-environmental systems. In parallel, AI ethics in healthcare is predominantly framed through anthropocentric principles such as beneficence, autonomy, and fairness [13], with limited attention to long-term or systemic effects. Overall, while significant progress has been made in formal ethical reasoning, neuro-symbolic decision support, and uncertainty-aware modeling, these strands remain largely disconnected. There is still a lack of unified frameworks that integrate ethical reasoning, probabilistic decision-making, and long-term systemic considerations in healthcare, motivating the need for non-anthropocentric models operating across temporal and relational scales.

3 NAEL: Non-Anthropocentric Ethical Logic

NAEL [11] addresses settings where ethical adequacy cannot be reduced to short-term, agent-centric optimization, but instead emerges from long-term interactions among multiple agents and systems. It conceptualizes ethics as an adaptive inferential process embedded within an agent’s decision-making architecture, enabling non-anthropocentric reasoning in complex environments such as healthcare.

Ethics as Inference and Active Inference Foundations. Unlike traditional approaches that encode ethics as external constraints or predefined objectives, NAEL treats ethical reasoning as inference under uncertainty. Agents continuously update beliefs about the consequences of their actions and evaluate alternative policies based on evolving knowledge. Ethical adequacy is thus a dynamic property of the agent’s interaction with a partially observable environment, rather than a static property of individual decisions. This perspective is grounded in active inference, where perception and action are framed as the minimization of variational and expected free energy [10, 14]. Agents maintain a generative model over observations and latent states, updating beliefs to reduce uncertainty and selecting actions that balance epistemic (information-seeking) and pragmatic (goal-directed) objectives. In NAEL, preferences are interpreted as prior beliefs over desirable states, and ethical considerations are embedded directly within the generative model rather than imposed as external reward functions.

Global Expected Free Energy. Standard active inference evaluates policies from an individual perspective. NAEL extends this formulation by introducing *global expected free energy*, which captures the uncertainty induced by a policy across multiple agents and the environment, $G_{\text{global}}(\pi) = \sum_{i=1}^N \mathbb{E}_{Q_i}[F_i(\pi)] + F_{\text{env}}(\pi)$, where π denotes a candidate policy, $F_i(\pi)$ is the expected free energy associated with agent i , and $\mathbb{E}_{Q_i}[\cdot]$ denotes expectation under agent i ’s beliefs Q_i . The term $F_{\text{env}}(\pi)$ captures uncertainty induced at the environmental or system level. This formulation evaluates policies based on the total uncertainty they generate across agents and environment, enabling system-level rather than agent-centric decision-making.

Neuro-Symbolic Ethical Reasoning. NAEL adopts a neuro-symbolic architecture combining probabilistic inference with symbolic reasoning. Subsymbolic components handle perception and prediction, while symbolic components encode ethical structure within the generative model. These include: **Deontic logic**, where normative constraints are represented using modal operators over propositions ϕ , namely **O** ϕ (obligatory), **P** ϕ (permitted), and **F** ϕ (forbidden), with standard relations such as **F** $\phi \equiv \text{O}\neg\phi$, constraining the admissible policy space; **Standpoint logic**, where reasoning across agents is modeled using indexed modalities A_i , and expressions of the form $\text{From}(A_i, \phi)$ denote that proposition ϕ holds under agent i ’s epistemic perspective, enabling evaluation from multiple stakeholder viewpoints; and **Subjective logic**, where beliefs are represented as triples $\omega_\phi = (b_\phi, d_\phi, u_\phi)$, with $b_\phi + d_\phi + u_\phi = 1$, capturing belief, disbelief, and uncertainty, respectively, which propagate through logical expres-

sions and influence confidence in inferred conclusions. By integrating these components within the generative model, NAEL ensures that action selection jointly accounts for normative constraints, multi-agent perspectives, and uncertainty.

4 Use Case: Ethical ICU Resource Allocation

To demonstrate the generality of NAEL, we consider an ICU resource allocation scenario under conditions of sustained demand (e.g., pandemic surge). An autonomous decision-support agent assists clinicians in allocating scarce resources such as ICU beds, ventilators, and staff time among patients with heterogeneous clinical profiles. At each decision point, the agent must balance competing objectives including patient survival, uncertainty in prognosis, staff workload, and overall system stability, which cannot be reduced to a single utility function. The agent operates through a unified active inference framework. It first constructs a generative model to infer latent clinical and system states s_t (e.g., disease severity, staff fatigue, capacity) from observations o_t (e.g., vital signs, laboratory results, staffing levels). The model defines priors $P(s_1)$, likelihood $P(o_t|s_t)$, and state transitions $P(s_{t+1}|s_t, a_t)$, while preferences $C(o_{t+1})$ encode desirable outcomes such as survival, fairness, and resource stability. Based on these inferred states, candidate actions are evaluated using NAEL’s integrated ethical reasoning process. Deontic logic encodes normative constraints over actions, such as $critical(p, t) \rightarrow O(a_{ICU}(p, t + 1))$, ensuring that critical patients are prioritized when resources are available. Standpoint logic evaluates each action from multiple stakeholder perspectives, including patients, staff, and the hospital system, formalized as $From(A_{patient_i}, \phi_i)$, $From(A_{staff}, \psi)$, and $From(A_{hospital}, \gamma)$, capturing clinical outcomes, workload considerations, and system-level stability. Subjective logic represents uncertain or incomplete information using (b, d, u) belief triples, allowing the agent to down-weight unreliable evidence and optionally request additional information under high uncertainty. Finally, actions a_t are selected by minimizing global expected free energy, $a_t^* = \arg \min_{a \in A_T} G_{global}(a)$, which integrates clinical outcomes, epistemic uncertainty, and ethical constraints into a single decision criterion.

5 Conclusions

When instantiated in healthcare, NAEL (Non-Anthropocentric Ethical Logic) provides a principled framework for medical decision-support across temporal, clinical, relational, and systemic dimensions. Healthcare decisions inherently involve trade-offs between immediate patient outcomes, long-term trajectories, resource allocation, and broader system-level impacts. NAEL enables AI systems to evaluate recommendations not only for short-term clinical effectiveness but also for their contribution to uncertainty reduction, equity, and system sustainability, supporting clinically effective and ethically coherent decisions. Through an ICU resource allocation case study, we illustrated how NAEL integrates sub-symbolic perception with symbolic ethical reasoning to support action selection

in high-stakes environments. Rather than replacing clinical guidelines or professional judgment, NAEL is intended to augment decision-making by making explicit ethical and systemic trade-offs that are often implicit in current systems. Compared to conventional AI ethics approaches in healthcare, NAEL introduces three main advantages: (i) **dynamic adaptation** via probabilistic ethical representations, (ii) **transparency** through explicit symbolic reasoning traces, and (iii) **multi-stakeholder integration** across patient, staff, and institutional perspectives. Despite its promise, NAEL remains a theoretical framework. The presented case relies on simplified assumptions, and real-world deployment would require richer clinical models and longitudinal data. Additional challenges include value specification, governance of ethical priors, computational scalability, and the translation of symbolic reasoning into clinically usable explanations. Future work will focus on empirical validation and scalable implementations in realistic healthcare environments.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Apriceno, G., et al.: A neuro-symbolic approach to structured event recognition. In: 28th International Symposium on Temporal Representation and Reasoning, TIME 2021, September 27–29, 2021, Klagenfurt, Austria. LIPIcs, vol. 206, pp. 11:1–11:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2021)
2. Apriceno, G., et al.: A neuro-symbolic approach for real-world event recognition from weak supervision. In: 29th International Symposium on Temporal Representation and Reasoning, TIME 2022, November 7–9, 2022, Virtual Conference. LIPIcs, vol. 247, pp. 12:1–12:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2022)
3. Apriceno, G., et al.: End-to-end neuro-symbolic approaches for event recognition (2023)
4. Apriceno, G., et al.: A neuro-symbolic approach for non-intrusive load monitoring. In: ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023). Frontiers in Artificial Intelligence and Applications, vol. 372, pp. 3175–3181. IOS Press (2023)
5. Apriceno, G., et al.: Exploring large language model reasoning capabilities over personal health data. In: Bellazzi, R., Juarez Herrero, J.M., Sacchi, L., Zupan, B. (eds.) AIME 2025, Part II. LNCS, vol. 15735, pp. 12–17. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-95841-0_3
6. Apriceno, G., et al.: A trustworthy evolutionary fuzzy neural network framework for maternal health risk classification. In: Bellazzi, R., Juarez Herrero, J.M., Sacchi, L., Zupan, B. (eds.) AIME 2025, Part II. LNCS, vol. 15735, pp. 18–23. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-95841-0_4
7. Badreddine, S., et al.: Interval logic tensor networks. CoRR abs/2303.17892 (2023)
8. Besnard, P., Hunter, A.: Elements of Argumentation. MIT Press (2014)

9. de Campos Souza, P.V., et al.: Enhancing logical tensor networks: integrating uninorm-based fuzzy operators for complex reasoning. In: Besold, T.R., d'Avila Garcez, A., Jimenez-Ruiz, E., Confalonieri, R., Madhyastha, P., Wagner, B. (eds.) *NeSy 2024, Part II*. LNCS, vol. 14980, pp. 68–79. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-71170-1_8
10. Friston, K., et al.: Active inference: a process theory. *Neural Comput.* **29**(1), 1–49 (2017)
11. Lerma, B.M., Peñaloza, R.: NAEL: non-anthropocentric ethical logic. arXiv preprint [arXiv:2510.14676](https://arxiv.org/abs/2510.14676) (2025). <https://arxiv.org/abs/2510.14676>, accepted to the FEAR workshop 2025
12. Lewis, J.E., Skawennati, Whiteduck, S.: Imagining indigenous AI. *AI & Society* **35**(1), 141–150 (2020)
13. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. *Nature Mach. Intell.* **1**(11), 501–507 (2019)
14. Pezzulo, G., Rigoli, F., Friston, K.: Hierarchical active inference: a theory of motivated control. *Trends Cogn. Sci.* **22**(4), 294–306 (2018)
15. Suchman, L.A.: *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press (1987)
16. Topol, E.J.: *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books (2019)
17. United Nations: *Transforming our world: the 2030 agenda for sustainable development*. United Nations General Assembly (2023)



Bridging Innovation and Regulation: Ethical Governance of Medical AI in Low- and Middle-Income Countries

Dalson Nyangoma^{1(✉)} and Gerald Mboowa^{1,2}

¹ African Center of Excellence in Bioinformatics and Data-Intensive Sciences, Infectious Diseases Institute, Makerere University, Kampala, Uganda
nyangomadalson23@gmail.com

² Department of Immunology and Molecular Biology, Makerere University, Kampala, Uganda

Abstract. Low- and middle-income countries (LMICs) are rapidly adopting artificial intelligence (AI) for clinical decision support, triage, and health system optimization, yet governance capacity often lags behind deployment. This creates a high-stakes tension: innovators must move quickly to address workforce shortages and care gaps, while regulators and health systems must ensure safety, privacy, equity, and accountability. We present a practical governance approach, Regulatory-Ethics-by-Design (REBD), that bridges innovation and regulation by operationalizing legal compliance and ethical safeguards across the medical AI lifecycle in LMIC settings. REBD integrates three components: regulatory traceability, context-aware ethical risk assessment, and RegOps (regulatory operations), defined here as lifecycle-integrated processes that embed compliance, monitoring, and documentation into AI development and deployment workflows. Through representative use cases in AI-assisted diagnostics and patient-risk prioritization, we demonstrate how REBD supports transparent oversight and responsible innovation. REBD offers an actionable roadmap for clinicians, researchers, and policy-makers to strengthen regulatory readiness while accelerating ethical medical AI adoption in LMIC health systems

Keywords: Responsible AI governance · Medical AI ethics and regulation · Low- and middle-income countries

1 Introduction

Artificial intelligence (AI) is moving from proof-of-concept to routine use in clinical pathways, including imaging triage, clinical decision support, and service optimization [1]. In low- and middle-income countries (LMICs), these tools are often introduced to offset shortages of specialist clinicians, limited diagnostic infrastructure, and uneven access to care [2]. At the same time, the conditions that motivate rapid adoption—heterogeneous populations, fragmented health information systems, and constrained monitoring capacity also increase the likelihood of harm if AI systems are deployed without robust safeguards.

Global guidance on trustworthy AI for health is evolving, but translating high-level principles into auditable, day-to-day development and deployment processes remains challenging in resource-constrained settings [1]. This creates a tension between innovation and oversight: programs need rapid iteration and scale-up, while regulators and implementers need evidence of clinical safety, sound data governance, clear accountability, and mechanisms to detect and respond to performance degradation over time.

Existing global frameworks, including WHO guidance on ethics and governance of AI for health, provide important principles such as transparency, accountability, inclusiveness, and safety, but they do not fully specify how these principles can be embedded into routine development, deployment, and monitoring workflows in resource-constrained settings. We therefore propose Regulatory-Ethics-by-Design (REBD), a practical governance approach that complements existing guidance by embedding legal compliance and ethical safeguards across the medical AI lifecycle in LMIC settings through regulatory traceability, context-aware ethical risk assessment, and RegOps-enabled monitoring. REBD supports audit-ready documentation of data provenance, consent, model changes, decision logs, subgroup performance, and deployment decisions, while prioritizing LMIC-amplified risks such as dataset shift, inequitable access, privacy breaches, limited redress, inadequate accountability, and patient safety harms. We illustrate REBD using representative diagnostic decision support and patient-risk prioritization deployments relevant to African health systems, while noting that its effects on safety, equity, and implementation outcomes require prospective evaluation.

2 Regulatory-Ethics-by-Design Framework

The proposed REBD framework embeds ethical and legal considerations directly into the AI development and deployment lifecycle. Regulatory traceability ensures auditable documentation of data provenance, consent, and model updates (see Fig. 1). Context-aware ethical risk assessment prioritizes risks such as dataset shift and inequitable access. RegOps pipelines integrate compliance checks into continuous development workflows, enabling real-time governance without impeding innovation.

3 Illustrative Use-Cases

The following use cases are illustrative real-world medical AI deployments used to demonstrate how REBD could be applied. They are not formal implementations of the REBD framework by the present authors.

3.1 AI-Assisted Diagnostic Decision Support in LMICs

AI-assisted diagnostic decision support systems have been deployed in several LMIC health systems to address specialist shortages and improve diagnostic coverage, particularly in imaging-based screening programs. A prominent example is the use of deep learning systems for diabetic retinopathy (DR) screening in Zambia, where an AI model

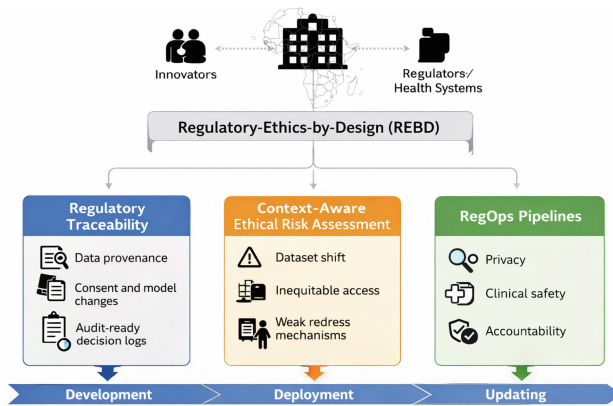


Fig. 1. Regulatory-Ethics-by-Design (REBD) framework for medical AI in low- and middle-income countries (LMICs) especially in Africa. The figure illustrates how REBD bridges innovation and regulation by embedding ethical and legal safeguards across the medical AI lifecycle. The framework integrates three core components regulatory traceability (data provenance, consent, and audit-ready decision logs), context-aware ethical risk assessment (addressing risks such as dataset shift, inequitable access, and weak redress mechanisms), and RegOps pipelines (privacy, clinical safety, and accountability checks) across development, deployment, and model updating. REBD supports transparent oversight and responsible innovation while remaining adaptable to resource-constrained health system contexts.

was integrated into a national screening workflow and evaluated under real-world clinical conditions. In this deployment, AI outputs were reviewed by human graders, enabling task shifting while maintaining clinical safety and accountability [3]. This use case underscores the importance of regulatory traceability documenting data provenance, grading decisions, and model updates to support auditability and clinical governance.

Similarly, computer-aided detection (CAD) systems for tuberculosis (TB) have been deployed across multiple sub-Saharan African settings to support chest X-ray interpretation in high-burden, resource-constrained environments. Large-scale validation studies show that deep learning-based CAD tools can achieve clinically acceptable performance for TB screening in routine care, including among populations with high HIV prevalence [4, 5]. These deployments highlight the need for continuous bias and performance monitoring, as model performance may vary across sites, populations, and imaging equipment—an issue explicitly addressed within REBD through context-aware ethical risk assessment and RegOps-enabled monitoring.

3.2 Patient-Risk Prioritization and Triage

AI-based risk prioritization tools have also been implemented to support clinical triage and workflow optimization in LMIC settings. In maternal health, deep learning models have been used to estimate gestational age from ultrasound sweeps acquired by minimally trained operators in public-sector facilities in Zambia, enabling improved antenatal risk stratification where specialist sonographers are scarce [6]. Multi-country evaluations across African settings further demonstrate that such models can be susceptible to dataset

shift when deployed outside their original training context, underscoring the importance of deployment-specific validation and adaptive governance mechanisms [7].

More recently, AI-based systems for cervical precancer screening have been prospectively evaluated across multiple African countries, demonstrating feasibility for population-level risk prioritization while raising important considerations around consent, equitable access, and regulatory oversight. Within REBD, these challenges are addressed by embedding privacy, clinical safety, and accountability checks into deployment pipelines, alongside transparent documentation of model updates and decision thresholds. Together, these use cases show that AI systems are already being deployed in LMIC health systems, while also exposing governance gaps related to traceability, bias, and regulatory readiness. REBD provides a structured approach to addressing these gaps without impeding iterative innovation by emphasizing auditable data provenance, deployment-specific validation, subgroup-aware monitoring, predefined escalation pathways, and documentation of model updates and decision thresholds.

4 Discussion

REBD is intended to make governance actionable rather than aspirational. In LMIC deployments where clinical resources and regulatory capacity can be constrained, the practical question is not whether transparency, fairness, and accountability are desirable, but how implementers can demonstrate these properties consistently in real operational workflows. By coupling traceability artefacts (e.g., consent and provenance records, decision logs, and change histories) with explicit risk assessment and embedded compliance checks, REBD supports audibility and clearer accountability across stakeholders, including developers, implementers, clinicians, and regulators.

The African use cases motivate lifecycle governance. Diagnostic AI (e.g., diabetic retinopathy screening and TB computer-aided detection) can deliver value at scale, yet performance can vary across sites, devices, and populations and may degrade as clinical practice or case-mix changes. Without monitoring and update controls, these shifts can widen inequities and erode trust. Similarly, triage and risk-prioritization systems can influence downstream access to scarce services; therefore, governance should include pre-specified decision thresholds, monitoring triggers, and escalation pathways for safe rollback or recalibration when drift is detected [7]. REBD addresses these needs through RegOps pipelines that treat compliance and monitoring as continuous processes rather than one-time approvals.

5 Conclusion

REBD provides a lightweight, auditable, and adaptable governance framework for medical AI in LMICs. By bridging innovation and regulation, it supports responsible AI adoption and offers practical guidance for policymakers, clinicians, and researchers working toward equitable and trustworthy AI-enabled healthcare.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. World Health Organization: Ethics and Governance of Artificial Intelligence for Health. WHO, Geneva (2021)
2. Schwalbe, N., Wahl, B.: Artificial intelligence and the future of global health. *Lancet*. **395**, 1579–1586 (2020). [https://doi.org/10.1016/S0140-6736\(20\)30226-9](https://doi.org/10.1016/S0140-6736(20)30226-9)
3. Bellemo, V., Lim, Z.W., Lim, G., Nguyen, Q.D., Xie, Y., Yip, M.Y.T., et al.: Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digit. Health*. **1**, e35–e44 (2019). [https://doi.org/10.1016/S2589-7500\(19\)30004-4](https://doi.org/10.1016/S2589-7500(19)30004-4)
4. Khan, F.A., Majidulla, A., Tavaziva, G., Nazish, A., Abidi, S.K., Benedetti, A., et al.: Chest x-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: a prospective study of diagnostic accuracy for culture-confirmed disease. *Lancet Digit. Health*. **2**, e573–e581 (2020). [https://doi.org/10.1016/S2589-7500\(20\)30221-1](https://doi.org/10.1016/S2589-7500(20)30221-1)
5. Rajpurkar, P., O’Connell, C., Schechter, A., Asnani, N., Li, J., Kiani, A., et al.: CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *npj Digit. Med*. **3**, 115 (2020). <https://doi.org/10.1038/s41746-020-00322-2>
6. Pokaprakarn, T., Prieto, J.C., Price, J.T., Kasaro, M.P., Sindano, N., Shah, H.R., et al.: AI estimation of gestational age from blind ultrasound sweeps in low-resource settings. *NEJM Evid*. **1**, EVIDoA2100058 (2022). <https://doi.org/10.1056/EVIDoA2100058>
7. Sendra-Balcells, C., Campello, V.M., Torrents-Barrena, J., Ahmed, Y.A., Elattar, M., Ohene-Botwe, B., et al.: Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five African countries. *Sci. Rep*. **13**, 2728 (2023). <https://doi.org/10.1038/s41598-023-29490-3>



A Multimodal-Based Framework for Smart ECG Report Generation

Hong Duc Nguyen¹ , Duc Tri Phan² , and S. Supraja¹

¹ School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore

{hongduc.nguyen,supraja.s}@ntu.edu.sg

² Mental Potential & Wellbeing Research Centre (EMPOWER), Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore
ductri.phan@ntu.edu.sg

Abstract. An electrocardiogram (ECG) is a vital instrument for heart-related disease diagnosis through wearable devices for early treatment. Despite the recent rapid advancement in the development of large language models (LLMs), they are not widely adopted and utilized. In this work, we propose a multimodal framework for ECG signal processing and interpretation based on LLMs and vision-based language models (VLMs). By aligning the ECG time series features, text-based clinical ECG report, and visualized ECG signal through a trainable attention module, our framework provides accurate physiological analysis. Our findings demonstrate that by leveraging pre-trained LLMs and with the addition of the visual domain, which is rarely used in ECG diagnosis, we can enhance medical text interpretation without a heavy retraining computational load. This work contributes toward a more efficient and accurate physiological diagnosis, which enables smart AI healthcare applications and a potential multi-task medical assistant in the future.

Keywords: Electrocardiogram · Multimodal AI · Large Language Models · AI in Healthcare · Report Generation

1 Introduction

Electrocardiograms (ECGs) play a crucial role in assessing a patient's physiological states by capturing the cardio signal and monitoring the heart's function. Nevertheless, ECG signals are generally susceptible to noise and interferences such as electrode misalignment, movement artifacts, or electrical noise during recording sessions or through wearable sensors, which can affect the morphological characteristics of the waveforms, causing misreading or misdiagnosis based on the received signals [4]. In addition, it is time-consuming, inefficient, and cost-ineffective to manually extract and process the hand-crafted ECG features. Thus, automated ECG signal processing becomes the preferred choice due to superior accuracy and lower inference time. Traditional ECG processing methods rely on

conventional threshold-based and rule-based approaches, which cannot adapt to complex, non-linear ECG artifact patterns; thus, their capabilities to deal with practical signals are limited [7]. Recently, deep learning-based ECG signal processing has gained momentum due to the rapid advancement of convolutional neural networks (CNNs) and transformer-based processing. Using deep layers of large parameter convolutional blocks, they improve the feature extraction process, extract more meaningful information from the input ECG waveforms, and increase the classification accuracy [3, 5]. In addition, with the breakthrough in natural language processing (NLP), particularly large language models (LLMs), several studies explore the alignment between textual features and extracted ECG features to generate diagnosis and interpretation [1]. However, these works mostly focus on feature alignment between the textual and single-dimensional (1-D) ECG domains and neglect other modalities.

In this work, we propose a multimodal framework that leverages the visual, textual, and physiological domains to accurately generate textual reports for ECG signal analysis. Experiment results prove that by utilizing pre-trained large vision language models (LVLMs) as the visual encoder, and pre-trained LLMs as the output head for textual generation, we can achieve modestly accurate ECG interpretation, enabling future research on multimodal ECG signal processing for smart artificial intelligence (AI) applications in biomedical and healthcare.

2 Methodology

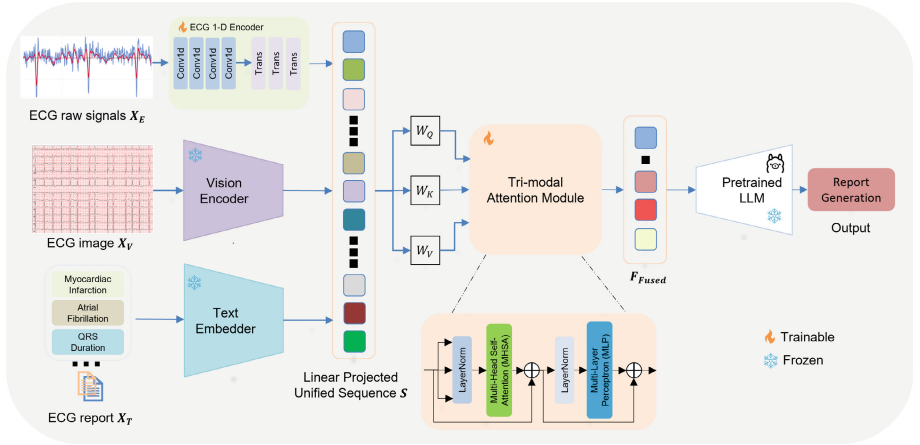


Fig. 1. Architecture of our proposed framework with the input raw ECG signals X_E , visualized ECG X_V , and the textual report X_T . The unified sequence S is produced by concatenating the linear projected embeddings from the extracted features of the three encoders, and F_{Fused} acts as the input fed into the pre-trained LLM to produce natural language generation output.

Our proposed tri-modal framework is summarized in Fig. 1. The architecture consists of a lightweight hybrid CNN-Transformer ECG encoder, text embeddings, a pre-trained vision encoder, a multimodal attention mechanism that utilizes the unified sequence to model the intra-domain and cross-domain relationships, and a pre-trained LLM that serves as the task head for text generation. Let the input for the 3 encoders be X_E , X_V , and X_T , respectively.

ECG Encoder. To process the 12-lead ECG signals, a hybrid architecture of CNN and Transformers was used to capture both local morphological features and long-range dependencies of the signal. The signal first goes through four layers of 1-D CNN to capture the local waveform patterns, with the last layer expanding to the channel dimension D_E . Subsequently, three Transformer blocks, each consisting of a Multi-Head Self-Attention (MHSA) module with a Feed-Forward Network (FFN), were used to produce the ECG feature F_E throughout the recording duration N_E defined as

$$F_E = E_{ECG}(X_E) = \text{Transformer}(\text{CNN}(X_E) + P_{pos}), \quad F_E \in \mathbb{R}^{N_E \times D_E}. \quad (1)$$

Text Embedder. To extract the information from each ECG report, we employ the pre-trained text embedder [6] to map tokenized clinical texts into a sequence of dimensional embeddings, D_T . By leveraging a pre-trained text encoder from a vision-language model, we preserve the fine-grained details that are already aligned with visual input, accelerating the training of the subsequent fusion module, shown as: $F_T = E_T(X_T)$, $F_T \in \mathbb{R}^{N_T \times D_T}$.

Vision Encoder. Capturing high-level features from the visualized ECG image is performed via a pre-trained vision encoder. While ECG images differ significantly from natural images, utilizing a pre-trained encoder on natural images (e.g., ImageNet) is highly beneficial. The shallow convolutional layers act as generalized feature extractors, detecting low-level features such as grid lines, waveform morphologies (e.g., P-waves, QRS complexes), and signal spikes. Given the input image X_V , the pre-trained encoder E_V produces the output visual features F_V as: $F_V = E_V(X_V)$, $F_V \in \mathbb{R}^{N_V \times D_V}$.

Multimodal Attention. Unlike pairwise cross-attention (text-image, text-ECG signal) that scales quadratically with the number of modalities, we first employ a linear projection on each domain output feature F_i ($i \in \{E, T, V\}$) from each encoder to a shared embedding space D_{shared} to obtain the unified feature $\hat{F}_i = F_i W_i + b_i$, where $W_i \in \mathbb{R}^{D_i \times D_{shared}}$. By utilizing modality embedding U_i , the three feature embeddings $H_i = \hat{F}_i + U_i$, where $U_i \in \mathbb{R}^{1 \times D_{shared}}$, are then concatenated together to form a single unified sequence $S = \text{Concat}(H_V, H_T, H_E)$, where $S \in \mathbb{R}^{(N_V + N_T + N_E) \times D_{shared}}$. A Transformer Self-Attention layer is then used to model the intra-domain relationship, e.g., 1-D ECG and 1-D ECG, and cross-domain relationship accordingly to produce the attention Z_{Fused} :

$$Z_{Fused} = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{D_{shared}}}\right), \quad (2)$$

where $Q = SW_Q$, $K = SW_K$, and $V = SW_V$ are the query, key, and value respectively. Finally, a LayerNorm layer is used to produce the fused feature F_{Fused} to the desired output dimension for the LLM input: $F_{Fused} = \text{LayerNorm}(S + Z_{fused})$.

Pre-trained LLM as Task Head. To translate the output multimodal fused features into an ECG textual report, we utilize state-of-the-art pre-trained LLMs and prompt them to generate a natural language description of the ECG signal. All prompts are embedded to form a prompt representation, concatenated with the final output multimodal ECG features to form the input for the LLMs, described by $O_{LLM} = \text{MLP}(F_{Fused})$, $O_{LLM} \in \mathbb{R}^{(N_V + N_T + N_E) \times D_{LLM}}$.

3 Experiments

Table 1. Evaluation of natural language generation metrics on the MEETI dataset. Model sizes are denoted by ‘M’ for millions and ‘B’ for billions of parameters. Results highlighted in **Bold** and *Underline italic* represent the best and second-best scores, respectively.

Vision Encoder	Size	Interpreted LLM	Size	BLEU-1	BLEU-2	BLEU-4	ROUGE-L	ROUGE-1	ROUGE-2
LLaVA-Next	7B	LLama-3	8B	<u>0.722</u>	<u>0.687</u>	<u>0.625</u>	<u>0.759</u>	<u>0.780</u>	<u>0.683</u>
LLaVA-Next	7B	Mistral-Nemo	12B	0.745	0.701	0.658	0.773	0.786	0.696
SigCLIP	400M	LLama-3	8B	0.632	0.538	0.461	0.556	0.601	0.574
SigCLIP	400M	Mistral-Nemo	12B	0.648	0.572	0.491	0.583	0.628	0.602
DINOV2-Giant	1.1B	LLama-3	8B	0.651	0.582	0.502	0.637	0.673	0.601
DINOV2-Giant	1.1B	Mistral-Nemo	12B	0.682	0.639	0.579	0.710	0.701	0.673
BiomedCLIP	200M	LLama-3	8B	0.629	0.535	0.479	0.548	0.593	0.565
BiomedCLIP	200M	Mistral-Nemo	12B	0.636	0.535	0.449	0.573	0.612	0.593

3.1 Experiment Settings

In this work, the MIMIC-IV dataset [2] and its extension MEETI [8] are used for the visualized ECG images. MIMIC-IV is the largest publicly available multimodal ECG dataset with 12-lead ECG raw signals with corresponding diagnostic reports, while the MEETI dataset provides raw ECG images corresponding to MIMIC-IV in a 85:15 data split. The experiments are conducted on a single NVIDIA RTX 5090 GPU with PyTorch 2.0. AdamW optimizer is used with a learning rate (LR) of 10^{-3} , $\beta_1 = 0.9$, and $\beta_2 = 0.99$. For the learning rate scheduler, CosineAnnealingLR is used with a minimum LR of 10^{-5} . Batch size is set to 32 with 100 training epochs; early stopping is employed after 5 epochs without improvement in validation loss, and the dropout rate is set to 0.1 with 10^{-4} weight decay to prevent overfitting. During training, ECG reports are provided for the model to map the raw signal and ECG image to the text, while during inference, the framework can generate the textual reports from input signal and image solely.

3.2 Results and Discussion

To assess the quality of the generated ECG report, we use the two common natural language generation metrics, BLEU and ROUGE. BLEU is used to measure phrase overlap (n-grams) between the generated report and the clinician’s ground truth report, indicating better lexical representation. On the other hand, ROUGE focuses on the recall rate of how much of the clinical text is captured by the model, with ROUGE-L measuring the longest common subsequence, and higher ROUGE ensures all key information on ECG reports is captured. Table 1 illustrates the experiment results using 4 different vision encoders, LLaVA-Next, SigCLIP, DINOv2, and BiomedCLIP; and 2 different LLMs for output interpretation, LLaMA-3 and Mistral-Nemo.

From Table 1, it can be seen that combining LLaVA-Next and Mistral-Nemo yields the best results across all metrics. This can be potentially attributed to the complexity of the models (7B parameters and 12B parameters, respectively), proving the superior ability to capture and encode high-level information and complex structural dependencies compared to smaller-scale encoders. In addition, BiomedCLIP (200M parameters), a biomedical visual encoder, yields almost similar performance (e.g., BLEU-1 of 0.636 vs. 0.648 with Mistral Nemo) to SigCLIP (400M parameters), an encoder with twice the number of parameters, proving the usefulness of domain-specific pre-training. Nevertheless, these encoders are more efficient in terms of computational complexity compared to the state-of-the-art methods and yield viable results, which make them more suitable for a resource-constrained environment or mass deployment.

4 Conclusions

In this work, we propose a novel tri-modal ECG report generation framework integrating the raw 1-D signals, visualized images, and clinical textual reports. Extensive experiment results across multiple encoder and decoder configurations demonstrate the framework’s ability to generate accurate diagnosis reports. In the future, we plan to expand our framework to a multi-task system that is capable of ECG signal quality assessment and denoising and utilize non-linear extracted ECG features to further improve the framework’s performance, contributing to the advancement of AI-based applications in the biomedical field.

References

1. Folgado, D., et al.: Conformal prediction for ECG interpretation: a study on human-AI collaboration in clinical decision support. In: Bellazzi, R., Juarez Herrero, J.M., Sacchi, L., Zupan, B. (eds.) AIME 2025. LNCS, vol. 15734, pp. 140–149. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-95838-0_14
2. Johnson, A.E., et al.: MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**(1), 1 (2023)

3. Li, X., et al.: Detection of ADHD from ECG signals via deep learning-based feature extraction. In: Bellazzi, R., Juarez Herrero, J.M., Sacchi, L., Zupan, B. (eds.) AIME 2025. LNCS, vol. 15734, pp. 232–241. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-95838-0_23
4. Liu, G., Han, X., Tian, L., Zhou, W., Liu, H.: ECG quality assessment based on hand-crafted statistics and deep-learned s-transform spectrogram features. *Comput. Meth. Programs Biomed.* **208**, 106269 (2021). <https://doi.org/10.1016/j.cmpb.2021.106269>, <https://www.sciencedirect.com/science/article/pii/S0169260721003436>
5. Nguyen, H.D., Phan, D.T.: CMA-ECG: cross-modal attention for enhanced ECG quality assessment and denoising. *Physiol. Meas.* **46**(11), 115001 (2025)
6. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PmLR (2021)
7. Simone, L., Bacciu, D., Gervasi, V.: Quasi-orthogonal ECG-frank XYZ transformation with energy-based models and clinical text. In: Finkelstein, J., Moskovitch, R., Parimbelli, E. (eds.) AIME 2024. LNCS, vol. 14845, pp. 249–253. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-66535-6_27
8. Zhang, D., et al.: MEETI: a multimodal ECG dataset from MIMIC-IV-ECG with signals, images, features and interpretations. arXiv preprint [arXiv:2507.15255](https://arxiv.org/abs/2507.15255) (2025)



Same Models, Better Results: Improving Multimodal LLMs' Accuracy in Identifying Myocardial Infarction from ECG Images

Japp Adhikari¹(✉), Keyuan Jiang¹, Hania Farhan²,
and Gordon R. Bernard³

¹ Purdue University Northwest, Hammond, IN, USA

{adhikarj,kjiang}@pnw.edu

² Alfaisal University, Riyadh, Saudi Arabia

haniafarhan2001@gmail.com

³ Vanderbilt University, Nashville, TN, USA

gordon.bernard@vumc.org

Abstract. Myocardial Infarction (MI) remains one of the leading causes of death worldwide. Electrocardiogram (ECG) is a reliable diagnostic tool for MI and its interpretation requires cardiology expertise. Advancement of multimodal large language models (MLLMs) attracted attempts to interpret ECG images, but prior studies showed low accuracies. To address the issue, we investigated a few strategies for pretrained MLLMs: (1) inclusion of domain diagnostic prompt instruction, (2) few-shot technique by providing annotated examples in the prompt, and (3) adjustment of model hyperparameters. On a corpus of 928 annotated 12-lead ECG images, our results utilizing Google Gemini 2.5 Pro and OpenAI GPT-4o for MI versus non-MI classification showed noticeable performance improvements. With 30-shot prompting, Gemini achieved 76.08% accuracy and GPT-4o reached 81.57%, substantially outperforming prior results (~30%). GPT-4o also achieved a high balance between sensitivity (71.55%) and specificity (85.05%).

Keywords: Myocardial infarction · Electrocardiogram · Multimodal large language model · Few-shot learning · ECG image interpretation

1 Introduction

Myocardial infarction (MI), also known as heart attack, is caused by acute coronary blood flow obstruction leading to ischemic myocardial injury. MI can be life-threatening and is often clinically silent. The chance of death due to this remains high if not treated in time. Electrocardiography (ECG) is a widely used and highly effective tool for detecting both acute and prior MI, with acute MI typically reflected by ST-segment abnormalities and prior infarction indicated by pathological Q-wave changes [3]. Since 1950's, ECG has been established as the clinical benchmark for detecting myocardial infarction and other arrhythmias [2].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027

P. Andreev et al. (Eds.): AIME 2026, LNAI 16749, pp. 96–100, 2027.

https://doi.org/10.1007/978-3-032-30813-9_18

Despite its central role, ECG interpretation can vary on subjectivity even among trained clinicians, which motivated on the development of AI-assisted approaches [1]. Recent work has shifted ECG analysis toward generative AI. Günay et al. [5] compared GPT-4, GPT-4o, and Gemini Advanced against cardiologists on 40 ECG cases; GPT-4o showed the strongest AI agreement (Fleiss’ $\kappa = 0.514$) but still lagged behind human experts. Seki et al. [7] reported that MLLMs achieved approximately 30% accuracy in zero-shot ECG VQA with frequent hallucinations. Lee et al. [6] evaluated multimodal LLMs (GPT-4o and Gemini 2.5 Pro) for MI detection from ECG images using zero-shot prompting with no domain-specific guidance, no few-shot examples, nor hyperparameter tuning. Their results showed limited performance (65.95% for GPT-4o and 29.63% for Gemini), with highly imbalanced sensitivity and specificity and unreliable diagnostic explanations. Their study serves as the baseline for this work.

This study focuses on evaluating pretrained MLLMs in an inference-only setting. Although conventional and deep learning methods have shown highly promising results [1], they typically require supervised model training with a large corpus of labeled data, and can be time-consuming and costly. Hence, these methods are outside the scope of this research. Building on our findings, this work aims to improve MI classification accuracy by (1) incorporating domain-specific prompting, (2) few-shot examples, and (3) targeted hyperparameter adjustments to promote more consistent output.

2 Method

To improve performance, we experimented on Google Gemini 2.5 Pro and OpenAI GPT-4o for ECG image interpretation on a corpus of publicly available 12-lead ECG images [4]. Both of these state-of-the-art models were accessed via their APIs. The workflow involves feeding ECG images into a multimodal LLM along with the prompt and few shot examples with controlled hyperparameters, to generate binary outputs (MI vs Non-MI). The overall experimental workflow is shown in Fig. 1.

2.1 Data

A corpus of anonymized 12-lead ECG images from Kaggle [4] was used. It was categorized into four classes: MI, prior MI (PMI), abnormal heartbeat (HB), and normal (N), with 3,023 training and 928 test samples. Following the data partition by Lee et al. [6], the test set was reformulated as a binary task (MI vs. Non-MI), where Non-MI class includes PMI, HB and N. For few-shot prompting, 30 training images were *randomly* selected from the training set: 15 MI and 15 Non-MI (5 PMI, 5 HB, and 5 normal) cases.

2.2 Prompt Design

Unguided prompts can produce inconsistent outputs, while domain-informed prompts incorporating criteria of clinical diagnosis [3] can yield more accurate

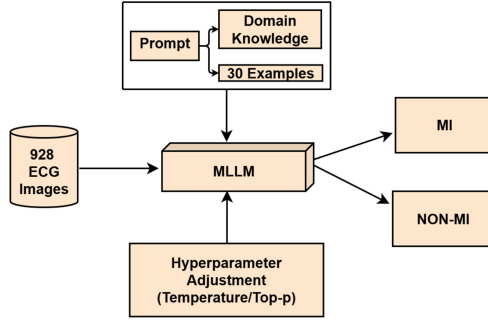


Fig. 1. Workflow of ECG image classification using MLLMs (Gemini 2.5 Pro and GPT-4o) in a few-shot setting and with domain-informed prompts.

and stable predictions. We adopted a structured prompt defining the diagnostic task, output format, and classification categories (MI and Non-MI), with the system instruction: “*You are an expert cardiologist specializing in ECG interpretation.*” Due to the token usage limitation, we did not ask the models to generate explanations, although the reasoning capability was on.

To reduce latency and cost for few-shot technique, we leveraged Gemini 2.5 Pro’s *context caching* to upload example images once and keep them in the model’s cache for reuse within the next 24 hours. GPT-4o does not support the same context caching mechanism and requires uploading examples for every request. The full prompt templates and additional implementation details are available on GitHub¹. The goal of this work is to evaluate whether pretrained MLLMs can be effectively guided using a small number of annotated examples (few-shot examples), without any model retraining.

2.3 Hyperparameters

Two hyperparameters of both models were adjusted in this study. *Temperature* controls the degree of randomness in a model’s output, and *top p* (nucleus sampling) restricts token selection to the most probable options. Starting from default values of 1.0 for both parameters, we experimented on multiple combinations of 0.0, 0.1 and 0.5. We observed that temperature of 0.0 and top p of 0.1 produced the most consistent classifications across both models and improved reliability for this task. Hence, these settings were used.

3 Results

Table 1 presents the performance results of our zero-shot and few-shot prompting for both MLLMs in comparison with the zero-shot baseline results from [6]. Metrics include accuracy, F1-score, sensitivity (true positive rate), and specificity (true negative rate).

¹ <https://github.com/adhikarj/ECG-Image-Classification>.

Table 1. Performance of ECG classification for baseline (0-shot) and our methods (0-shot and 30-shot). Acc: accuracy, Improv: improvement from the baseline, Sensit: sensitivity, and Specif: specificity.

Model	Scenario	Acc.	Improv.	F1	Sensit.	Specif.
Gemini 2.5 Pro	Baseline 0-shot	29.63	—	—	97.07	6.24
	Our 0-shot	72.63	+145%	66.23	56.49	78.23
	Our 30-shot	76.08	+157%	66.07	42.26	87.81
GPT-4o	Baseline 0-shot	65.90	—	—	36.40	76.20
	Our 0-shot	68.86	+4.5%	63.73	60.67	71.70
	Our 30-shot	81.57	+24%	76.97	71.55	85.05

The detailed results can be found on GitHub¹. The confusion metrics for Gemini 2.5 Pro (zero-shot) correctly classified 135 MI (TP) and 539 Non-MI (TN) cases, with 104 false negatives (FN) and 150 false positives (FP), while GPT-4o correctly predicted 145 MI (TP) and 494 Non-MI (TN) cases, with 94 FNs and 195 FPs. With 30-shot, both models show overall improved performance. Gemini 2.5 Pro identified 101 MI (TP) and 605 Non MI (TN) cases correctly, whereas GPT-4o (30-shot) achieved the best overall balance by correctly identifying 171 MI (TP) and 586 Non-MI (TN) cases with fewer errors.

4 Discussions

As can be seen in Table 1, the performance of both MLLMs improved noticeably over that of the baseline in zero-shot, with further gains from 30-shot prompting, and achieved a great balance between sensitivity and specificity, contributing to more accurate diagnosis of MI and exclusion of Non-MI cases. For zero-shot prompting, accuracy of Gemini 2.5 Pro increased from 29.63% (baseline) to 72.63% (+145%), and GPT-4o from 65.90% (baseline) to 68.86% (+4.5%). With 30-shot prompting, Gemini reached 76.08% (+157% from the baseline), while GPT-4o achieved the highest accuracy of 81.57% (+24% from the baseline). The F1-score for Gemini remained consistent (66.23% to 66.07%), whereas GPT-4o improved from 63.73% to 76.97%. Gemini shifted from a highly imbalanced baseline result (97.07% vs. 6.24%) to a more balanced profile (56.49% vs. 78.23%) in zero-shot condition, but its 30-shot sensitivity dropped to 42.26%, indicating a conservative bias toward Non-MI. GPT-4o achieved the best overall balance in the 30-shot setting (71.55% vs. 85.05%).

An error analysis was conducted on a subset of 30 misclassified cases through manual validation by a clinician (author HF). The result indicates that most errors resulted from missed diagnostic cues and model variability. GPT-4o showed higher prediction variability than Gemini, consistent with the known non-deterministic behavior of LLMs.

A key limitation of this study is the binary reformulation of four disorder classes [6], limiting the granularity of the multiclass classification. Additionally,

use of a public dataset introduces potential data leakage, which the study models may have seen in learning. However, consistent performance improvements suggest that the gains are driven by the proposed strategies rather than the learned knowledge in the models. Furthermore, this study is performed with a single 30-shot subset due to computational constraints. Different sets of few-shot examples may reveal varying performance improvement, which remains a future work of this study.

5 Conclusion

This study demonstrates the performance improvement of our strategies with pretrained MLLMs for 12-lead ECG image interpretation, despite remaining accuracy gaps. Incorporating domain-specific prompting, hyperparameter tuning, and a small number of examples substantially improved MI detection. These findings highlight the potential of MLLMs as scalable tools without additional training.

Acknowledgement. The authors wish to thank 3 reviewers for their critiques and constructive comments which improved the quality of this manuscript.

References

1. Abubaker, M.B., Babayiğit, B.: Detection of cardiovascular diseases in ECG images using machine learning and deep learning methods. *IEEE Trans. Artif. Intell.* **4**(2), 373–382 (2022)
2. Disrud, L.W., et al.: A pilot study of the home-based 12-lead electrocardiogram in clinical practice. *Mayo Clinic Proc. Innov. Qual. Outcomes* **9**(2), 100598 (2025). <https://doi.org/10.1016/j.mayocpiqo.2025.100598>, PMID: 40092494; PMCID: PMC11909748
3. ECG Library: A collection of electrocardiogram examples and interpretive guides (2025). <https://ecglibrary.com/ecghome.php>. Accessed 14 Nov 2025
4. EvilSpirit05: ECG analysis dataset (2025). <https://www.kaggle.com/datasets/evilspirit05/ecg-analysis/data>. kaggle. Accessed 07 Nov 2025
5. Günay, S., Öztürk, A., Yiğit, Y.: The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: a comparison with cardiologists and emergency medicine specialists. *Am. J. Emerg. Med.* **84**, 68–73 (2024)
6. Lee, H., Yoo, S., Kim, J., Cho, Y., Suh, D., Lee, K.: Comparative diagnostic performance of a multimodal large language model versus a dedicated electrocardiogram AI in detecting myocardial infarction from electrocardiogram images: comparative study. *JMIR AI* **4**, e75910 (2025)
7. Seki, T., Kawazoe, Y., Ito, H., Akagi, Y., Takiguchi, T., Ohe, K.: Assessing the performance of zero-shot visual question answering in multimodal large language models for 12-lead ECG image interpretation. *Front. Cardiovasc. Med.* **12**, 1458289 (2025)



Transforming Hospitals with Artificial Intelligence: Applications in Clinical Education and Patient Care

Nihal Haque^{1,2(✉)}, Ervin Sejdic¹, Michael Liut³, Roland Mollanji¹, Phil Shin^{1,2},
Duska Kennedy¹, and Ghadir Ali¹

¹ North York General Hospital, Toronto, ON, Canada
nihal.haque@utoronto.ca

² Department of Medicine, University of Toronto, Toronto, ON, Canada

³ University of Toronto Mississauga, Mississauga, ON, Canada

Abstract. Artificial intelligence (AI) is rapidly reshaping hospital-based care and medical education, yet frontline clinicians often lack practical frameworks to evaluate and implement these tools safely and effectively. This presentation describes a multidisciplinary, hospital-based approach to adopting AI applications in both clinical workflows and teaching. The use of AI scribes, automated discharge summaries, and educational chatbots are used as representative case studies. First, we outline key medico-legal, privacy, bias, and workflow considerations derived from national guidance and institutional experience at a large community teaching hospital in Toronto with a multidisciplinary AI working group. We then examine AI scribes for patient care, focusing on informed consent, data handling, specialty-specific prompting, and their impact on cognitive burden and documentation quality in history-heavy fields such as geriatric medicine. Next, we present early work on large language model-generated discharge summaries, highlighting feasibility, workload reduction potential, and current barriers related to data privacy, hallucinations, and integration with electronic medical records. Finally, we report results from an AI chatbot named Medbot designed to teach medical syndromes to learners, including completion rates and agreement between AI and human raters on performance assessment. Collectively, these examples illustrate how thoughtfully governed AI solutions can decrease clinician cognitive load, enhance patient care and enrich medical education while underscoring ongoing ethical, legal, and sociotechnical challenges that require clinician leadership and interdisciplinary collaboration. We will also demonstrate the AI scribe, automated discharge summary solution and educational AI chatbot at the conference.

Keywords: Large Language Models · Hospitals · Discharge Summaries · Chatbots · Conversational Agents · Geriatric Medicine · Education

1 Introduction

Hospitals worldwide face escalating clinical complexity, documentation demands, and workforce burnout, particularly in internal medicine and geriatrics. AI technologies like LLM-based tools offer opportunities to automate routine cognitive and administrative

tasks, potentially reallocating clinician time toward direct patient care and teaching [1]. At the same time, these tools introduce new medico-legal, privacy, and bias concerns that are incompletely addressed by existing regulatory frameworks.

Community teaching hospitals occupy a unique position, functioning simultaneously as sites of care delivery and clinical education. They are well suited to serve as living laboratories for pragmatic AI implementations while prioritizing patient safety and trainee education. This paper describes the experience of a large community hospital (North York General Hospital) affiliated with a university medical school (University of Toronto) that established an institutional AI working group and piloted AI scribes, LLM-generated discharge summaries, and an educational chatbot in clinical medicine for trainees. We situate these local initiatives within the emerging evidence base and medico-legal guidance and propose a set of pragmatic considerations for clinicians contemplating AI adoption in hospital settings.

2 Institutional Context and Governance

North York General Hospital (NYGH) is a 410-bed community teaching hospital affiliated with the University of Toronto. Individuals providing leadership in AI include an AI Research Chair, Vice President of Digital Health, and a Chief Medical Information Officer. In early 2024, with the widespread availability of LLM-based tools, the hospital convened a multidisciplinary AI working group. This group developed strategic AI priorities and a governance pipeline for evaluating and approving new AI tools for physician use, with AI scribes identified as the first clinical use case.

The governance process integrated clinical need, evidence review, privacy and security assessments, and medico-legal analysis. Key questions include the intended purpose of the AI (clinical vs. administrative), regulatory status (e.g., software as a medical device), anticipated impact on patient care, and potential harms from incorrect or biased outputs. By embedding these deliberations in a standing committee rather than ad hoc project teams, our institution sought to create a consistent approach to AI adoption that created shared responsibility across diverse projects.

3 Medico-Legal and Privacy Considerations

Canadian Medical Protective Association (CMPA) guidance emphasizes that, regardless of automation, physicians remain responsible for the content and quality of clinical documentation and for decisions informed by AI outputs [2]. This has direct implications for the deployment of AI scribes, which transform ambient audio into clinical notes that clinicians must review, edit, and sign. The medico-legal framework requires a clear patient consent process, acknowledgement that AI tools are assistive rather than autonomous, and ongoing monitoring for clinically significant inaccuracies or omissions by the end user.

Privacy and data protection are central concerns, especially when tools rely on cloud-based processing or external vendor infrastructure. Canadian hospitals must navigate provincial health information legislation such as the Personal Health Information Protection Act (PHIPA) for the province of Ontario and assess whether data processing

can occur on Canadian servers with appropriate safeguards. Privacy impact assessments address whether patient data are used for model improvement, how long transcripts and notes are stored, who can access them, and when data can be deleted.

4 AI Scribes in Hospital Practice

AI scribes use ambient listening and LLM-based natural language processing to convert clinician–patient interactions into draft clinical notes, which may be formatted as SOAP (Subjective-Objective-Assessment-Plan) notes, consultation notes, or progress notes. Early evaluations suggested that AI scribes could reduce documentation time and after-hours charting, improve perceived workflow efficiency, and modestly alleviate burnout. A recent systematic review of eight studies concluded that AI scribes improved documentation efficiency and quality, although productivity and accuracy gains were heterogeneous and dependent on local workflows and training [3].

At NYGH, AI scribes were piloted in history-intensive specialties such as geriatric medicine, psychiatry, and internal medicine, where encounters are typically longer and more conversational. A mixture of non-Electronic Medical Record (EMR) integrated and EMR integrated AI scribes were trialed. The clinical workflow involves obtaining informed consent, activating the scribe, explicitly verbalizing key examination findings and investigations, and editing the generated note. While many clinicians appreciated the reduction in cognitive load, some found it challenging to relinquish fine-grained control over narrative style, underscoring the importance of iterative configuration and training.

5 LLM-Generated Discharge Summaries

Discharge summary preparation remains a major pain point in internal and hospitalist medicine, consuming substantial clinician time on the day of discharge. LLMs can ingest longitudinal progress notes and structured data to generate draft discharge summaries that clinicians then edit, potentially shortening documentation time on the day of discharge. A recent study using real-world electronic medical record data found that LLM-generated discharge summaries achieved scores comparable to human-authored summaries on a validated scoring metric, suggesting feasibility for clinical support [1].

At NYGH, a proof-of-concept project explored using LLMs to generate preliminary discharge summaries based on inpatient notes, highlighting the ability of models to interpret unstructured text, decipher acronyms, and reconstruct the temporal course of an admission. Privacy constraints necessitated a secure server to host the LLM and secure data pipeline to send information to and from the EMR. Hallucinations and subtle factual errors remain important limitations; if clinicians must meticulously verify every element, the net efficiency gains may be attenuated. Currently a pilot trial is in progress after initial trials showed potential workload reduction of 50–60% on day of discharge.

6 AI Chatbots for Medical Education

LLM-based conversational agents offer novel opportunities for scalable, interactive medical education. In geriatrics, chatbots can simulate patient encounters, provide immediate feedback, and help trainees practice history-taking and clinical reasoning around

complex syndromes such as falls, delirium, and polypharmacy. A growing body of work suggests that ChatGPT-like tools can supplement traditional curricula and support self-directed learning [4].

We developed a chatbot named Medbot to teach geriatric syndromes to residents using integrating scripted cases. In a pilot evaluation, 5 residents completed chatbot sessions, and they rated their increase in interest in the subject material as 8.8/10. Engagement level was rated as 9.2/10, satisfaction 8.6/10 and usefulness 8.8/10. These findings suggest that AI chatbots can assist human formative assessment in some domains. Feedback included learners wanting an audio component rather than using text chat. Learner performance was evaluated using a pre-determined rubric by the LLM and by two human physicians. AI-generated performance ratings showed 82.5% agreement with one human rater and 32.5% agreement with a second, while the two human raters agreed with each other 65% of the time. These findings suggest that AI can approximate human formative assessment in some domains, but variability highlights the need for careful rubric design, calibration, and ongoing human oversight.

7 Discussion

The experience described here underscores both the transformative potential and the nuanced risks of AI deployment in hospital-based care and education. AI scribes can meaningfully reduce documentation burden and improve note quality when integrated into thoughtfully redesigned workflows and coupled with specialty-specific prompting and training. LLM-generated discharge summaries offer an avenue to streamline a notoriously burdensome task, though privacy, hallucinations, and integration challenges must be addressed before routine clinical use. Educational chatbots can extend access to practice opportunities and feedback in geriatric medicine, potentially enriching curricula and supporting competency development.

Across these use cases, several principles emerge. These are transferable across institutions. First, clinicians must remain actively engaged in oversight, recognizing that responsibility for documentation and decisions cannot be delegated to AI. Second, institutional governance structures that integrate clinical, technical, privacy, and legal expertise are essential to evaluate tools, negotiate vendor arrangements, and monitor ongoing performance. Third, frontline clinicians and patients should help shape the evolution of AI tools to ensure they actually achieve what they set out to do which is decrease caregiver burnout and improve patient care.

This work has some limitations. Each pilot was conducted at a single institution with small samples and limited follow-up. Self-reported workload reduction estimates are subject to recall and social desirability bias and require objective time-tracking validation. Generalizability to other institutional contexts, patient populations, and electronic medical record environments remains to be established.

8 Conclusion

AI applications in hospital settings, including AI scribes, LLM-generated discharge summaries, and educational chatbots, can reduce cognitive burden, enhance documentation, and enrich medical education when embedded within robust governance frameworks

and coupled with vigilant clinician oversight. As health systems continue to adopt these tools, interdisciplinary collaboration among clinicians, educators, technologists, privacy experts and regulators will be critical to realizing benefits while mitigating medico-legal, privacy, and bias-related risks. We will also demonstrate the AI scribe, automated discharge summary solution and educational AI chatbot at the conference. Future work will examine how small language models (SLMs) can be deployed alongside larger frontier models in hybrid architectures, leveraging the privacy and efficiency advantages of on-premise SLMs while preserving the reasoning capabilities of larger models for complex tasks such as discharge summary synthesis.

Acknowledgments. We are grateful to the Digital Health team and the entire AI working group at North York General Hospital for this work. We acknowledge the financial support of the Natural Sciences & Engineering Research Council of Canada (NSERC), Discovery Grant #RGPIN-2024-04348 and the University of Toronto's Learning & Education Advancement Fund (LEAF).

Disclosure of Interests Dr. Nihal Haque has received grant funding from Canada Health Infoway and North York General Hospital.

Declaration of AI assistance technologies during manuscript preparation. The author used PERPLEXITY AI for CLASS 2 ACTIVITY. After using this technology, the authors reviewed the results and take full responsibility for the contents of the manuscript.

References

1. Haines, L.: Large language model discharge summary preparation using real-world electronic medical record data shows promise. *Intern. Med. J.* **55**(7), 1188–1192 (2025)
2. CMPA: The medico-legal lens on AI use by Canadian physicians Homepage, <https://www.cmpa-acpm.ca/en/research-policy/public-policy/the-medico-legal-lens-on-ai-use-by-canadian-physicians>. Last accessed 15 Feb 2026
3. Sasseville, M.: The impact of AI scribes on streamlining clinical documentation: a systematic review. *Healthcare (Basel)* **16**, **13**(12), 1447 (2025)
4. Ali, D.: ChatGPT in teaching and learning: a systematic review. *Educ. Sci.* **14**(6), 643 (2024)



Network-Based Integration of Multi-omics Data for Biomarker Discovery in Acute Coronary Syndromes

Elena Marinello^{1,2}(✉) , Shahryar Noei² , Marco Chierici² , Núria Amigó³,
Dane Cvijanović⁴, Aleksandar Davidović⁴, Dalibor Dragišić⁵,
Claudia Giglione⁶, Dimitris Kardassis⁷, Marko Milanov⁴, Jelena Munjas⁸,
Iva Perović-Blagojević⁵, Aleksa Petković⁵, Tamara Ratković⁸, Pol Torne³,
Christos Tsatsanis⁹, Sandra Vladimirov-Sopić⁸, Luka Vukmirović⁴,
Paolo Magni⁶, Miron Sopić⁸, Barbara Di Camillo¹ ,
and Giuseppe Jurman^{2,10}

¹ Department of Information Engineering, University of Padova, Padova, Italy
emarinello@fbk.eu

² Data Science for Health, Digital Health & Wellbeing Center, Fondazione Bruno
Kessler, Trento, Italy

³ Biosfer Teslab, Reus, Spain

⁴ KBC Zvezdara, Belgrade, Serbia

⁵ KBC “Dr Dragisa Mišović-Dedinje”, Belgrade, Serbia

⁶ Department of Pharmacological and Biomolecular Sciences “Rodolfo Paoletti”,
Università degli Studi di Milano, Milan, Italy

⁷ Laboratory of Biochemistry, University of Crete Medical School, Crete, Greece

⁸ Department of Medical Biochemistry, Faculty of Pharmacy, University of Belgrade,
Belgrade, Serbia

⁹ Laboratory of Clinical Chemistry, University of Crete Medical School,
Crete, Greece

¹⁰ Department of Biomedical Sciences, Humanitas University, Pieve Emanuele
(Milan), Italy

Abstract. Acute coronary syndromes (ACS), including ST-elevation myocardial infarction (STEMI) and non-ST-elevation myocardial infarction (NSTEMI), remain leading causes of mortality worldwide. Despite advances in diagnosis and treatment, single-omics approaches have proven insufficient to capture the molecular complexity underlying ACS pathophysiology. Consequently, we adopted a multilayer network approach to construct phenotype-specific networks for STEMI and NSTEMI, alongside a control multilayer network derived from patients with stable angina pectoris (SAP). The multilayer networks were constructed using data collected from 200 patients within the CardioSCOPE project, integrating one metabolomics layer and one microRNA layer. Nodes represented molecular features, while edges were defined based on Pearson correlation coefficients. Network analyses included interlayer connection investigation, hub identification, and community detection, whose results were used to select a compact panel of discriminative features

that achieved a cross-validated AUC of 0.84 (95% CI: 0.82–0.87) in a three-class separability test.

Keywords: Multi-omics Integration · Network-based Biomarker Discovery · Precision Cardiology · Acute Coronary Syndromes · Integrative Network Biology

1 Introduction

Cardiovascular diseases are the leading cause of mortality worldwide, accounting for 31% of all deaths [7], with acute coronary syndromes (ACS), including ST-elevation (STEMI) and non-ST-elevation myocardial infarction (NSTEMI), among the most prevalent. Despite advances in interventional cardiology, patients still face high residual risk of major adverse cardiovascular events (MACE), misdiagnosis remains common, and predicting ACS-to-MACE progression is challenging [6,8]. Single-omics approaches have proven insufficient to capture the underlying molecular complexity, highlighting the need for integrative strategies to improve disease characterization and patient risk stratification [11]. Within the EU-funded CardioSCOPE project [2], this study investigates whether integrative, network-based multi-omics analysis can better characterize ACS subtypes. We analyzed a cohort of NSTEMI and STEMI patients alongside stable angina pectoris (SAP) controls, constructing multilayer networks to integrate metabolomic and microRNA data, and applying graph-based data mining and artificial intelligence methods to identify disease-associated molecular markers and subnetworks. Results showed that phenotype-specific multilayer networks revealed distinct hub and community structures, and that a compact, network-derived feature panel, headed by hsa-miR-208b-3p, achieved an AUC of 0.84 (95% CI: 0.82–0.87) in a three-class separability test, supporting the value of integrative multi-omics analysis for ACS characterization.

2 Materials and Methods

The data used in this study were generated within the CardioSCOPE project and include molecular information from blood samples collected on a cohort of 200 patients (61 STEMI, 39 NSTEMI, and 100 SAP) recruited in Belgrade, Serbia, and matched by age, sex, BMI, comorbidities, and dyslipidemia treatment status to ensure comparability. Metabolomics profiles were obtained from NMR spectra [4] generated by a Bruker Avance III 600 spectrometer, producing 62 metabolites including lipoproteins (VLDL, IDL, LDL, HDL fractions), glycoproteins, low-molecular-weight metabolites, amino acids and lipid indices. Additionally, expression levels of 2,024 miRNAs were quantified by TAmiRNA GmbH [3] with miRBase v22.1 annotation. These two molecular layers were used to construct multilayer networks stratified by patient diagnosis.

Preprocessing included redundancy filtering of metabolomics variables (e.g., retaining only the aggregate VLDL-P measure), exclusion of variables with

> 30% missing values, and counts per million (CPM) normalization of miRNA raw counts, keeping only features with CPM > 1. Batch effects across recruitment hospitals were corrected using ComBat, and remaining missing values were median-imputed. A one-way ANOVA across diagnostic groups reduced the final panels to 34 metabolites and 48 miRNAs.

Multilayer networks were constructed by dividing the dataset into a training set (70% of the data), while the remaining 30% were used as held-out set for validation. To ensure robustness, a bootstrap-based approach was applied over 1,000 iterations, each sampling 70% of training patients to construct a bootstrap network. At each iteration, edges were defined by pairwise Pearson correlation coefficients, both intra- and inter-layer, computed for each diagnosis. Edges were retained if they appeared in at least 70% of iterations and had FDR-corrected $p < 0.05$, and were weighted by their mean correlation across those iterations. The resulting networks for each phenotype are shown in Fig. 1.

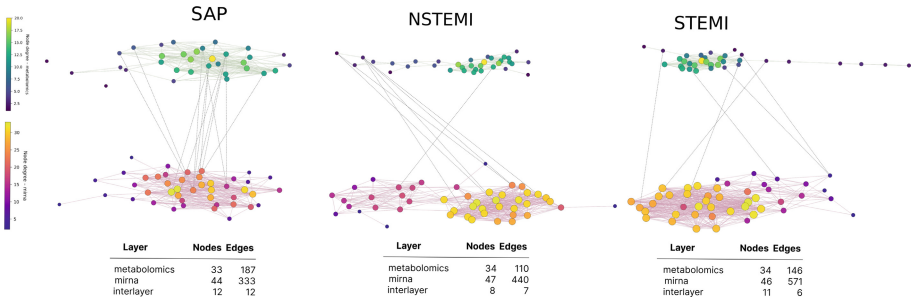


Fig. 1. Multilayer networks stratified by diagnosis. From left to right: SAP, NSTEMI, and STEMI. In each multilayer network, the upper layer represents the metabolomics network, while the lower represents the miRNA network. Nodes correspond to individual molecular features, and edges represent Pearson correlation coefficients (FDR-corrected, $p < 0.05$) that were kept after bootstrap stabilization ($\geq 70\%$ stability across 1,000 iterations). The number of nodes and edges for each layer is reported under each network.

Network analysis consisted of three main steps: interlayer connection investigation, hub analysis, and community detection. Interlayer edges were examined for each phenotype by extracting the subset of nodes that are interconnected between metabolites and miRNAs, including the corresponding intra-layer edges. In the hub identification, the top five hubs were extracted, ranking them within each omics layer by their weighted degree. Lastly, community structure was explored using the Leiden algorithm [9], with the resolution parameter selected independently per network to maximize the number of detected hubs (range: 0.01–0.5 for metabolomics, 0.05–0.1 for miRNA). Each node was then classified according to its within-module z-score and participation coefficient [5]; only *provincial* and *connector hubs* were maintained for downstream analysis.

3 Results and Discussion

Inter-layer analysis revealed sparse but phenotype-specific cross-talk between the metabolomics and miRNA layers. SAP exhibited the most inter-layer connections, with linoleic acid (LA) and free carnitine (FC) as key metabolite mediators, linked to multiple miRNAs including hsa-miR-29c-3p, which alone was connected to four metabolites. In NSTEMI, inter-layer connectivity was sparser and centered on lactate. STEMI showed the fewest inter-layer edges, distributed across distinct metabolite–miRNA pairs with no single dominant hub.

In the metabolomics layer, LDL-TG, LPC, and VLDL-C ranked among the top hubs across all phenotypes, consistent with their established role in cardiovascular disease, while miRNA hub composition was phenotype-dependent: SAP was characterized by hsa-miR-3065-3p and miR-30 family members, whereas both NSTEMI and STEMI were dominated by miR-378 family members, suggesting a coordinating role specific to acute myocardial injury.

Community detection revealed distinct modular structures in the metabolomics layer: SAP organized into 4 communities with 7 provincial hubs and leucine as the sole connector hub, NSTEMI showed the most fragmented structure (12 communities) with all 9 hubs acting as connectors, and STEMI displayed the simplest organization (2 communities) with all hubs classified as provincial. In contrast, the miRNA layer exhibited a more homogeneous organization across phenotypes. SAP contained 3 connector and 8 provincial hubs, while both NSTEMI and STEMI featured 2 connector hubs out of 12. Notably, cardiac-enriched hsa-miR-208a-3p and hsa-miR-208b-3p served as connector hubs in SAP, whereas hsa-miR-342-3p bridged communities in both NSTEMI and STEMI.

Network features were selected combining top hubs, provincial and connector hubs, and interlayer nodes, leading to 14–16 miRNAs and 12–14 metabolites per phenotype, which were used to train a Random Forest classifier as a separability test, achieving a cross-validated AUC of 0.84 (95% CI: 0.82–0.87) and a test AUC of 0.84 (95% CI: 0.74–0.94). SHAP analysis confirmed hsa-miR-208b-3p as the most consistently impactful feature across all classes, in line with its established role as a biomarker of acute myocardial infarction, where its upregulation is associated with increased risk of death or heart failure [1,10]. Among metabolites, Glutamine and saturated fatty acids emerged as the most consistently impactful across phenotypes.

It should be noted that, as this pipeline is exploratory by design, network properties such as hub rank, participation coefficient, and z-score are cohort-level quantities; accordingly, the RF classifier should be interpreted as a test of class separability rather than an estimate of individual-level predictive performance. Moreover, Pearson correlation captures only linear associations; future work may explore nonlinear alternatives such as MIC. Additionally, correlation does not imply causality, and network topology should be interpreted accordingly. Finally, the analysis relies on a single cohort; while bootstrap resampling partially mitigates this by estimating topological stability, independent validation on external cohorts remains a priority for follow-up studies.

Acknowledgments. This work is partially funded by the European Union through the CardioSCOPE project under the HORIZON-MSCA-2021-SE-01-01 MSCA Staff Exchanges 2021 programme (grant no. 101086397).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Declaration of AI Assistance Technologies During Manuscript Preparation. The authors used AI-based tools, specifically OpenAI's ChatGPT, in order to improve language clarity. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Bellin, G., Gardin, C., Ferroni, L., et al.: Exosome in cardiovascular diseases: a complex world full of hope. *Cells* **8**(2), 166 (2019). <https://doi.org/10.3390/cells8020166>
2. CardioSCOPE consortium: CardioSCOPE – Comprehensive and personalized assessment of acute coronary syndrome by multiomic approach and artificial intelligence strategy. <https://cardio-scope.eu/>. Accessed 05 Feb 2026
3. Diendorfer, A., Khamina, K., Pultar, M., Hackl, M.: miND (miRNA NGS Discovery pipeline): a small RNA-seq analysis pipeline and report generator for microRNA biomarker discovery studies. *F1000Research* **11**, 233 (2022)
4. Fuertes-Martin, R., et al.: Glycoprotein A and B height-to-width ratios as obesity-independent novel biomarkers of low-grade chronic inflammation in women with polycystic ovary syndrome (PCOS). *J. Proteome Res.* **18**(11), 4038–4045 (2019)
5. Guimerà, R., Nunes Amaral, L.A.: Functional cartography of complex metabolic networks. *Nature* **433**(7028), 895–900 (2005). <https://doi.org/10.1038/nature03288>
6. Huang, Z., Dong, W.: Adversarial MACE prediction after acute coronary syndrome using electronic health records. *IEEE J. Biomed. Health Inform.* **23**(5), 2117–2126 (2018)
7. James, S., Bueno, H.: Epidemiology of acute coronary syndromes. In: *The ESC Textbook of Cardiovascular Medicine*. Oxford University Press (2018). <https://doi.org/10.1093/med/9780198784906.003.0305>
8. Nestelberger, T., et al.: Predicting major adverse events in patients with acute myocardial infarction. *J. Am. Coll. Cardiol.* **74**(7), 842–854 (2019)
9. Traag, V.A., Waltman, L., Van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**(1), 1–12 (2019)
10. Yu, H., Chen, X., Lu, L.: Large-scale prediction of microRNA-disease associations by combinatorial prioritization algorithm. *Sci. Rep.* **7**, 43792 (2017). <https://doi.org/10.1038/srep43792>
11. Zhan, C., et al.: From multi-omics approaches to personalized medicine in myocardial infarction. *Front. Cardiovasc. Med.* **10**, 1250340 (2023)



Automated ICF Profiling from ICU Time Series Using Transformer-Based Physiological Prediction

Feten Skhiri Gabbouj^{1,2(✉)}, Mathias Blandeau¹, Matei Mancas²,
Bernard Gosselin², and Laura Wallard¹

¹ University Polytechnique Hauts-de-France, CNRS, UMR 8201 LAMIH, 59313 Valenciennes, France

² ISIA lab - University of Mons, Mons, Belgium
`feten.skhirigabbouj@umons.ac.be`

Abstract. ICU patients generate complex, sparse, and irregularly sampled multivariate time series that challenge standard clinical analysis. We propose an automated framework that maps ICU time series to standardized functional outcomes using the International Classification of Functioning, Disability and Health (ICF). Using a Transformer for Multivariate Irregular Time Series (T-MITS) architecture, the framework predicts serum creatinine and glucose at 24h post-admission and transforms these predictions into discrete ICF impairment levels across renal and metabolic domains. Evaluated on the MIMIC-IV database, this work demonstrates the feasibility of AI-assisted ICF profiling in critical care, bridging physiological prediction with standardized functional health representation.

Keywords: ICF · Time Series · T-MITS · ICU

1 Introduction

The Intensive Care Unit (ICU) continuously generates high-frequency multivariate time series (MITS) from laboratory measurements, vital signs, and bedside monitoring systems. Variable-level predictions alone do not provide a unified representation of the patient's overall functional status, and methods transforming physiological time series into clinically interpretable functional descriptions remain needed. The International Classification of Functioning, Disability and Health (ICF) offers such a framework, describing health in terms of body functions and impairments [9], yet its adoption remains limited because ICF coding is manual, expertise-dependent, and retrospective—incompatible with fast-paced ICU environments. We propose an automated framework mapping ICU multivariate time series to ICF functional outcomes using T-MITS, evaluated on creatinine and glucose using MIMIC-IV.

2 Related Work

Recent work has explored automating ICF classification from clinical text [1, 13] and medical imaging [2], yet mapping ICU physiological time series to ICF representations remains largely unexplored. Clinical ICU time series are irregularly sampled, asynchronously recorded, and highly sparse, violating assumptions of standard temporal models. Transformer-based approaches with learned temporal encodings have shown strong performance on such data [4, 10, 12], motivating our use of T-MITS [3].

3 Methods

3.1 Data Set

The experiments use MIMIC-IV [6], a large publicly available critical care dataset, with 39,885 ICU admissions. Individual models for glucose (test N = 7,977) and creatinine (test N = 7,765) were trained on 80/20 patient splits.

3.2 Model

We employ T-MITS [3], inspired by STRATS [11], which directly ingests clinical observations as (*Time*, *Variable*, *Value*) triplets, preserving temporal structure without imputation bias. Triplets are processed via self-attention and aggregated through attention-weighted pooling, then combined with static demographic features (age, sex) to generate predictions at 24 h, mapped to discrete ICF categories. T-MITS was selected for its demonstrated superiority over GRU-D, TCN, Latent-ODE, and classical ML baselines [3].

Table 1. ICF Qualifiers and Clinical Thresholds. Creatinine thresholds based on [5, 7].

Variable	Metric
Creatinine	KDIGO Criteria
	0 (No Impairment): < 1.35 mg/dL (Normal)
	1 (Mild Impairment): $1.35 \leq x < 2.68$ mg/dL (Risk)
	2 (Moderate Impairment): $2.68 \leq x < 4.16$ mg/dL (Injury)
	3 (Severe Impairment): ≥ 4.16 mg/dL (Failure)
Glucose	Glycemic Control Standards
	0 (No Impairment): $70 \leq x \leq 140$ mg/dL (Normoglycemia)
	1 (Dysfunction): < 70 mg/dL (Hypoglycemia)
	2 (Dysfunction): > 140 mg/dL (Hyperglycemia)

3.3 Pre-processing

ICU time series were represented as sparse $(Time, Variable, Value)$ triplets [3], filtered for outliers, and Z-score normalized on the training set to prevent data leakage. Target variables were labeled using thresholds in Table 1.

3.4 Experiments

Task Definition: Models predict glucose and creatinine from the first 24 ICU hours, mapped to discrete ICF categories using clinical thresholds (Fig. 1), with the modular design supporting future multimodal extensions.

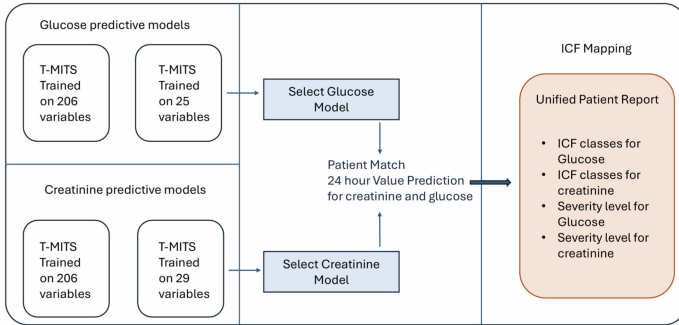


Fig. 1. Methodology

Prediction Models: Two settings were evaluated per target: *high-dimensional input* using the full dataset (206 variables), and *domain-specific input* using curated subsets—25 metabolic variables for glucose and 29 renal variables for creatinine.

Table 2. ICF classifiers for the study.

Variable	Related ICF classes
Glucose	B540: Metabolic functions
	B5400: Carbohydrate metabolic functions
	B555: Endocrine gland functions
	B5551: Pancreatic gland functions
Creatinine	B610: Urinary excretory functions
	B6100: Filtering functions of kidney
	B6101: Collecting functions of kidney
	B545: Water mineral and electrolyte balance

Post-training Fusion: Best-performing models predict glucose and creatinine at 24h, retained only for patients present in both tasks, mapped to ICF qualifiers (Table 1) and Body Function domains (Table 2), producing a unified report across metabolic (b540, b555) and renal (b610, b545) domains.

4 Results and Discussion

Table 3 shows stable performance for both targets, with accuracies of 80–83%. Increasing input variables from 25–29 to 206 did not yield substantial improvements, suggesting curated features capture the dominant predictive signal. Lower precision and recall reflect inherent class imbalance, rare conditions such as severe renal dysfunction or hypoglycemia are naturally underrepresented, rather than methodological limitations.

Table 3. Model Results

Variable	Features	Accuracy	Precision	Recall
Creatinine	29	82.72	58.86	50.82
	206	82.69	62.25	53.60
Glucose	25	79.60	75.50	53.00
	206	80.01	76.00	52.67

Evaluated on 1,561 patients with concurrent measurements, Table 4 confirms the framework translates biomarker predictions into structured ICF profiles across renal and metabolic domains. Future work should address imbalance-aware learning, baseline comparisons, and multi-institutional validation.

Table 4. Example of final output

Stay ID	Pred. Glucose	Glucose ICFs	Pred. Creatinine	Creatinine ICFs
54986	236.66 mg/dL	b540.2, b5400.2b555.2, b5551.2	1.27 mg/dL	b610.0, b6100.0b6101.0, b545.0
38729	170.35 mg/dL	b540.2, b5400.2 b555.2, b5551.2	1.39 mg/dL	b610.1, b6100.1 b6101.1, b545.1

5 Conclusion

This work demonstrates the feasibility of automated ICF profiling from multi-variate ICU time series using T-MITS, translating irregular physiological data into standardized functional health outcomes. The deterministic ICF mapping is intentional—grounded in validated clinical thresholds [5, 7]—providing a transparent bridge between physiological prediction and functional classification. While ICF is most established in rehabilitation, early ICU profiling enables continuity of functional assessment through discharge and rehabilitation planning

[8]. Future work will prioritize baseline comparisons, imbalance-aware learning, and multi-institutional validation.

Declaration of AI assistance technologies during manuscript preparation. The authors used Claude (Anthropic) for (STM Activity 1). After using these technologies, the authors reviewed the results and take full responsibility for the contents of the manuscript.

References

1. Gabbouj, F.S., Mancas, M., Blandeau, M., Wallard, L.: Evaluating the performance of LLMs in ICF classification: insights from medical and general models*. In: 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1–7 (2025). <https://doi.org/10.1109/EMBC58623.2025.11252817>
2. Gabbouj, F.S., Wallard, L., Blandeau, M., Vandenbulcke, V., Gosselin, B., Mancas, M.: AI-based multimodal ICF classification: application to Alzheimer’s disease. In: Conférence en TéléSANTé et Dispositifs Médicaux (TECSAN). Université de Technologie de Troyes, Troyes, France (2025). hAL Id: hal-05384680
3. Gille, C., Altares, G., Colette, B., Boudjeltia, K.Z., Mancas, M., Vandenbulcke, V.: Multivariable serum creatinine forecasting for acute kidney injury detection using an explainable transformer-based model. In: 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1–7 (2025). <https://doi.org/10.1109/EMBC58623.2025.11251723>
4. Horn, M., Moor, M., Bock, C., Rieck, B., Borgwardt, K.: Set functions for time series. In: Proceedings of the 37th International Conference on Machine Learning, vol. 119, pp. 4353–4363. PMLR (2020)
5. Jacobi, J., et al.: Guidelines for the use of an insulin infusion for the management of hyperglycemia in critically ill patients. Crit. Care Med. **40**(12), 3251–3276 (2012)
6. Johnson, A.E.W., et al.: MIMIC-IV, a freely accessible electronic health record dataset. Sci. Data **10**(1) (2023). <https://doi.org/10.1038/s41597-022-01899-x>
7. Disease, K.: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group: KDIGO clinical practice guideline for acute kidney injury. Kidney Int. Suppl. **2**(1), 1–138 (2012)
8. Lohmann, S., Decker, J., Müller, M., Strobl, R., Grill, E.: The ICF forms a useful framework for classifying individual patient goals in post-acute rehabilitation. J. Rehabil. Med. **2**, 151–155 (2011)
9. Rosenbaum, P., Stewart, D.: The World Health Organization international classification of functioning, disability, and health: a model to guide clinical thinking, practice and research in the field of cerebral palsy. Seminars Pediatr. Neurol. **11**(1), 5–10 (2004). <https://doi.org/10.1016/j.spen.2004.01.002>
10. Shukla, S.N., Marlin, B.M.: Multi-time attention networks for irregularly sampled time series. In: International Conference on Learning Representations (2021)
11. Tipirneni, S., Reddy, C.K.: Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. ACM Trans. Knowl. Discov. Data **16**(6) (2022). <https://doi.org/10.1145/3516367>
12. Vaswani, A., et al.: Attention is all you need. IN: Advances in Neural Information Processing Systems, vol. 30 (2017)
13. Zhang, L., Tashiro, S., Mukaino, M., Yamada, S.: Use of AI large language models as a clinical tool in rehabilitation medicine. J. Rehabil. Med. **55**, jrm13373 (2023). <https://doi.org/10.2340/jrm.v55.13373>



InstructFusionFormer: Instruction-Conditioned Multimodal Fusion for Esophageal Disease Diagnosis and Staging

Zenebe Markos Lonseko¹ , Helen Haile Hayeso¹, Tao Gan², and Nini Rao¹ (✉)

¹ Department of Biomedical Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

raonn@uestc.edu.cn

² Digestive Endoscopic Center of West China Hospital, Sichuan University, Chengdu 610017, China

Abstract. Accurate diagnosis and staging of esophageal diseases require integrating endoscopic imaging with heterogeneous and often incomplete clinical information. We propose InstructFusionFormer, an instruction-conditioned multimodal transformer that dynamically fuses endoscopic images and structured clinical variables via learnable instruction tokens. Modality dropout enhances robustness to missing clinical inputs. Evaluation on multi-institutional datasets shows strong performance (AUC 0.96, 95% CI: 0.94–0.98), outperforming unimodal and conventional multimodal baselines. The model remains effective without clinical inputs and provides interpretable attention aligned with expert annotations. These results indicate improved performance and robustness, although prospective validation is required.

Keywords: InstructFusionFormer · Multimodal Learning · Instruction-Conditioned Fusion · Esophageal Disease Diagnosis · Vision Transformers · Medical AI

1 Introduction

Esophageal diseases require accurate diagnosis and staging using both endoscopic imaging and clinical information [1]. However, most existing AI approaches rely on unimodal image analysis, limiting clinical context modeling [2]. Although multimodal learning has emerged as a promising solution [3], current methods often use static fusion strategies that fail to capture task-specific and dynamic cross-modal interactions.

To address this, we propose InstructFusionFormer, an instruction-conditioned multimodal transformer for dynamic, task-aware fusion of endoscopic and clinical data, enabling context-sensitive representations and achieving superior performance compared to unimodal and conventional multimodal baselines. Our contributions include: (i) an instruction-conditioned fusion mechanism, (ii) a transformer-based architecture for joint diagnosis and staging, and (iii) empirical validation demonstrating improved performance and robustness.

2 Materials and Methods

2.1 Dataset

We evaluate the proposed framework on paired endoscopic images and structured clinical variables for esophageal disease assessment. The primary task is seven-class lesion classification (NE, SC, OE, EV, ESE, EEC, AEC), with an auxiliary task distinguishing early versus advanced esophageal cancer.

A total of 20,965 white-light gastroscopic images from 4,077 patients were retrospectively collected at West China Hospital, with cancer diagnoses confirmed via histopathology. The clinical variables include demographic information (e.g., age, sex) and key clinical indicators (e.g., symptoms and comorbidities), which are encoded using a multilayer perceptron. Images were resized and normalized, and patient-level data splitting (70%/15%/15%) was applied to prevent data leakage. The study was conducted under institutional ethics approval with informed consent.

2.2 Proposed Method

As illustrated in Fig. 1, **InstructFusionFormer** consists of four stages: input representation, modality-specific encoding, instruction-conditioned fusion, and task-specific prediction. Endoscopic images are partitioned into patches and encoded using a Vision Transformer (ViT) pre-trained on ImageNet and fine-tuned on the target dataset, while clinical variables are encoded via a multilayer perceptron.

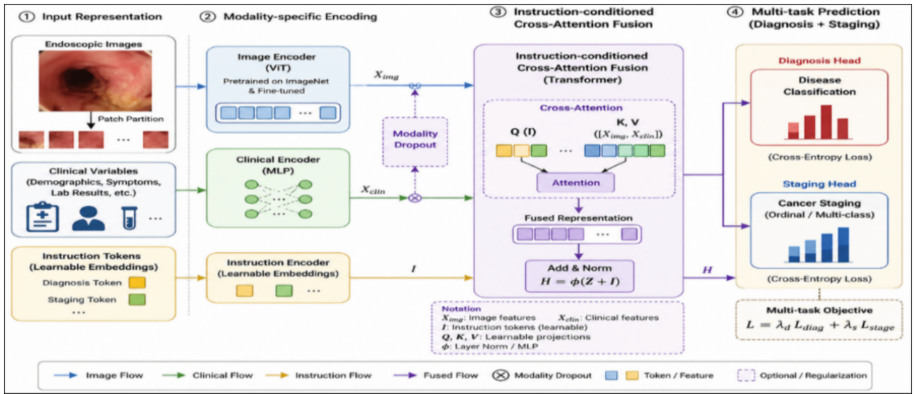


Fig. 1. Overall framework of InstructFusionFormer.

The core component is an instruction-conditioned fusion module, where task-specific instruction tokens guide cross-modal interaction through transformer-based cross-attention (Eq. (1)):

$$Z = \text{Softmax} \left(\frac{Q(I)K(X_{img}, X_{clin})^T}{\sqrt{d}} \right) V([X_{img}, X_{clin}]) \quad (1)$$

where X_{img} and X_{clin} denotes image and clinical features, and represents Instruction Tokens. Instruction tokens are implemented as two learnable embeddings ($N = 2$, $d = 256$) for diagnosis and staging, jointly optimized to capture task-specific modulation patterns via cross-attention, enabling differential weighting of multimodal features. The fused representation is obtained as (Eq. (2)):

$$H = \phi(Z + I) \quad (2)$$

The final representation is passed to task-specific heads for joint diagnosis and staging. A multi-task objective is used (Eq. (3)):

$$\mathcal{L} = \lambda_d \mathcal{L}_{diag} + \lambda_s \mathcal{L}_{stage} \quad (3)$$

where each is optimized using cross-entropy loss.

3 Results

3.1 Overall Diagnostic Performance

As presented in Table 1, InstructFusionFormer achieves the best performance, with 0.92 accuracy, 0.91 macro-F1, and 0.96 AUC, outperforming CNN, transformer, and conventional multimodal baselines by ~3% in accuracy. These gains indicate that instruction-conditioned fusion enables more effective integration than static multimodal approaches, with only a modest increase in latency.

Table 1. Comparison with state-of-the-art methods and computational complexity.

Method	Modality	Multi-task	Accuracy (%)	AUC (%)	Parameters (M)	Latency (ms/img)
ResNet50	Image	✗	0.86	0.91	25	34
EfficientNet-B7	Image	✗	0.87	0.92	37	41
ViT-B/16	Image	✗	0.88	0.93	22	45
Multimodal Fusion Transformer	Image + Clinical	✗	0.89	0.94	28	50
InstructFusionFormer (Ours)	Image + Clinical	✓	0.92	0.96	26	52

3.2 Ablation Study

Ablation results confirm that removing instruction conditioning or clinical inputs results in comparable performance drops (AUC $\downarrow$ ~0.02), indicating that both contribute significantly. The comparable performance drop observed when removing instruction conditioning and clinical inputs suggests that instruction tokens may partially act as a proxy for clinical feature weighting, rather than providing independent reasoning capacity.

3.3 Per-Class Diagnostic Performance

Per-class results (Fig. 2) demonstrate consistent improvements across all lesion categories, indicating reduced class bias and effective multimodal learning.

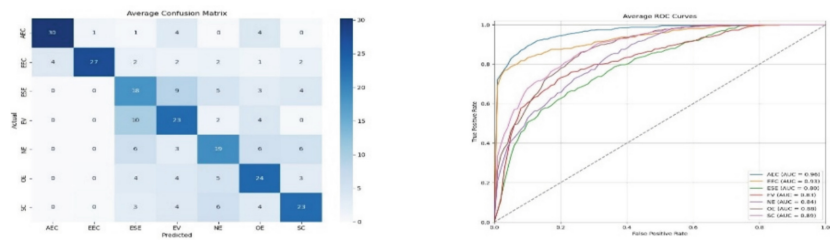


Fig. 2. Performance visualization of the proposed method (confusion matrix and ROC curves).

3.4 Robustness and Generalizability

Under missing clinical inputs (Table 2), performance degrades gradually but remains superior to unimodal baselines, demonstrating robustness to missing data. Cross-dataset evaluation (Table 3) shows stable performance across external datasets (AUC 0.91–0.93). However, due to differences in modality and label definitions across external datasets, these results should be interpreted as indicative robustness rather than strict task-level generalization. Future work will evaluate structured, clinically realistic missingness patterns beyond random masking to reflect real-world conditions in EHRs. These results collectively demonstrate that instruction-conditioned multimodal fusion provides consistent performance gains while maintaining robustness under realistic clinical constraints.

Table 2. Performance under missing clinical data.

Configuration	Data availability	Accuracy (%)	AUC (95% CI)
Image-only baseline	Not used	0.86	0.87 (0.84–0.90)
Partial	Partially missing	0.90	0.93 (0.90–0.95)
No clinical data	Missing	0.88	0.91 (0.88–0.94)
Multimodal input	Available	0.92	0.95 (0.93–0.97)

Table 3. Cross-Dataset Diagnostic Performance.

External Test Dataset	Accuracy (%)	AUC (95% CI)	Sensitivity (%)	Specificity (%)
HyperKvasir [4]	0.89	0.93 (0.90–0.95)	0.90	0.87
GastroVision [5]	0.88	0.92 (0.89–0.94)	0.89	0.86
TCGA-ESCA [6]	0.87	0.91 (0.88–0.93)	0.88	0.85

4 Discussion and Conclusion

We presented InstructFusionFormer, an instruction-conditioned multimodal framework for esophageal disease diagnosis and staging that enables dynamic integration of endoscopic images and clinical data. The model outperforms unimodal and conventional multimodal approaches, demonstrating improved learning and robustness. Ablation results indicate that instruction tokens act as task-guided modulators, suggesting a gating mechanism for multimodal interaction.

The model remains robust to incomplete data via modality dropout and provides attention aligned with expert regions, with errors mainly in ambiguous cases. However, interpretability remains qualitative and requires quantitative validation (e.g., overlap-based metrics or feature attribution). Prospective multi-center studies are needed, and future work will extend to unstructured, longitudinal data and broader applications.

Acknowledgments. This study was supported by the National Natural Science Foundation, No. 62271127 and No. 61872405; Medico-Engineering Cooperation Funds from the University of Electronic Science and Technology of China and West China Hospital of Sichuan University, No. ZYGX2022YGRH011 and HXDZ22005; the China Higher Education Institution Industry-University-Research Innovation Fund (No.2024IT007 and No. 2024DR039).

References

1. Roest, C., et al.: Multimodal AI combining clinical and imaging inputs improves prostate cancer detection. *Investig. Radiol.* **10**, 1097 (2023)
2. Xiao, Z., et al.: Instruction-vit: multimodal prompts for instruction learning in vision transformer. *Inf. Fusion.* **104**, 102204 (2024)
3. Hemker, K., Simidjievski, N., Jamnik, M.: Healnet: multimodal fusion for heterogeneous biomedical data. *Adv. Neural Inf. Process. Syst.* **37**, 64479–64498 (2024)
4. Borgli, H., et al.: HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci. Data.* **7**, 283 (2020). <https://doi.org/10.1038/s41597-020-00622-y>
5. Jha, D., et al.: GastroVision: A Multi-class Endoscopy Image Dataset for Computer Aided Gastrointestinal Disease Detection. (2023)
6. Lucchesi, F.R., Aredes, N.D.: The Cancer Genome Atlas Esophageal Carcinoma Collection (TCGA-ESCA) (Version 3). The Cancer Imaging Archive. (2016)



Automated Behavioral Health Assessment Using Vocal Biomarkers

Namhee Kwon^(✉) , Nate Blaylock , Raymond Brueckner ,
Vinod Subramanian , and Henry O'Connell 

Canary Speech Inc., Provo, UT 84606, USA
`{namhee,nate,ray,vinod,henry}@canaryspeech.com`

Abstract. This work proposes a voice-based clinical decision support tool that detects a generalized behavioral health condition, defined by Major Depressive Disorder (MDD) and Generalized Anxiety Disorder (GAD), from spontaneous speech. We compare Random Forest, XGBoost, and fully connected deep neural network classifiers trained on high-dimensional voice representations, including X-vector, Wav2Vec2, TRILLsson, and HuBERT embeddings. Models are trained using validated symptom instruments, including the PHQ-8 and GAD-7. To enhance clinical safety, we incorporate an ‘uncertain’ classification category for low-confidence samples. Our model achieves a sensitivity of 0.68 and specificity of 0.80, demonstrating that a consolidated vocal biomarker metric provides a scalable, non-invasive mechanism to identify health conditions in both in-clinic and remote settings.

Keywords: vocal biomarkers · anxiety · depression · behavioral health

1 Introduction

Mental health disorders impose a substantial global burden, with approximately 332 million people affected by depression and 359 million by anxiety worldwide according to the World Health Organization (WHO). Major depressive disorder (MDD) and generalized anxiety disorder (GAD) are among the most prevalent and co-occurring conditions, yet many cases remain undiagnosed and untreated. Voice-based screening offers a scalable, non-invasive, and cost-effective path toward early identification, motivating robust speech-based detection systems.

This study suggests a classification model using 60 s voice samples. This work differs from prior approaches in three key ways. First, we rely exclusively on acoustic speech features, deliberately excluding transcription or linguistic content, making our system applicable to any spoken material regardless of language or topic. Second, we propose a unified behavioral health model that consolidates both conditions into a single assessment framework, reflecting that tool aims to identify high-risk individuals and guide them toward clinical care, rather than

provide a definitive diagnosis. Third, we introduce a confidence-based uncertainty category to flag low-confidence predictions, improving clinical safety and real-world applicability.

2 Data

The dataset consists of spontaneous speech responses to the open-ended prompt “How are you?”, paired with labels derived from the PHQ-8 [3, 8] and GAD-7 [7] self-assessment questionnaires. Using free-form speech, rather than structured prompts such as read speech, supports broader real-world applicability. Participants were recruited to ensure demographic diversity across gender, age, race, and state of residence. The two label sets were highly correlated (Pearson $r = 0.79$, $p < 0.0001$).

The primary dataset, collected remotely via smartphone, were partitioned into training, development, and test sets via stratified splitting to preserve label, age, and gender distributions, with no participant appearing in more than one split. The binary depression and anxiety labels were defined by PHQ-8 ($0 \leq y \leq 24$) and GAD-7 ($0 \leq y \leq 21$) scores using a clinical cut-off score of 10. To reduce ambiguity near the decision boundary, training samples with scores in the range $8 \leq y \leq 12$ were excluded; all samples were retained in the development and test sets to reflect real-world score distributions. An independent in-clinic dataset collected using a tablet was held out as a second external test set (Test2).

Table 1. Experimental dataset size

Label	Train	Dev	Test1	Test2
	Negative Cnt/Positive Cnt (Positive Ratio)			
Depression	1177/548 (32%)	291/137 (32%)	266/133 (33%)	142/25 (15%)
Anxiety	1300/506 (28%)	317/146 (32%)	292/142 (33%)	142/25 (15%)

3 Model

We generated audio embeddings from each recording using four pre-trained models—X-vector [6], Wav2Vec2 [1], TRILLsson [5], and HuBERT [2]—trained on large, diverse corpora to produce robust, general-purpose representations. We trained three classifiers: Random Forest (RF), XGBoost (XGB), and a fully connected deep neural network (DNN) models. Both RF and XGB addressed class imbalance via undersampling with bagging. The DNN was trained with ReLU activations, sigmoid output, Adam optimizer, and early stopping over 100 epochs. Hyperparameters, including network depth, hidden layer width, and dropout rate, were tuned using Hyperband [4] on the development set. Class imbalance for the DNN was addressed via SMOTE oversampling.

Beyond individual models, we explored three integration strategies: early fusion (EF), concatenating feature vectors into a joint input; late fusion (LF), concatenating submodel output representations before the last network layer; and ensemble (ENS), combining independent model predictions via majority voting. These strategies aim to leverage complementary strengths across features and architectures.

4 Results

4.1 Depression and Anxiety Models

We built separate classification models for depression and anxiety across all feature and architecture combinations described in Sect. 3, with results summarized in Table 2. For HuBERT, we report models using the 20th and 21st layer outputs, which consistently outperformed other layers.

XGBoost and DNN models with HuBERT features were the strongest individual models, achieving a UAR of 0.69 for depression and 0.66 for anxiety. While early and late fusion did not improve over individual models, ensemble modeling via majority voting over the top three development-set performers yielded consistent gains. The best depression model was an ensemble of three DNNs trained on TRILLsson, HuBERT layer 20, and HuBERT layer 21 (UAR 0.70); the best anxiety model combined a DNN (HuBERT layer 21), XGBoost (HuBERT layer 20), and XGBoost (TRILLsson) (UAR 0.68).

4.2 Behavioral Health Assessment

In many screening contexts, the goal is to identify individuals with any behavioral health condition, MDD or GAD, rather than distinguish between them. We evaluated two approaches for joint assessment, summarized in Table 3. In Joint Training, we trained and compared models with a composite label defined as positive if either condition was present and negative only if both were absent. In Post-hoc Aggregation, the outputs of the best-performing independent models were combined using a logical OR. Joint Training did not improve over individual models, likely because the PHQ-8 and GAD-7 labels, while highly correlated ($r = 0.79$), capture largely overlapping symptoms. Post-hoc Aggregation achieved a UAR of 0.69 with a higher sensitivity of 0.73, matching the average performance of the standalone classifiers while better capturing at-risk individuals.

4.3 Confidence-Based Prediction

Our models output prediction probabilities between 0 and 1, with predictions classified as positive when exceeding a threshold, set to 0.5 by default as in Table 2. To improve clinical reliability, we introduce an ‘uncertain’ category for predictions near the decision boundary. Specifically, predictions above an upper threshold are classified as positive, those below a lower threshold as negative,

Table 2. Model Performance. **Sensitivity** (true positive rate), **Specificity** (true negative rate), and their unweighted average (**UAR**) is reported.

Model	Feature Set	Depression			Anxiety		
		UAR	Sens	Spec	UAR	Sens	Spec
RF	X-vector	0.59	0.47	0.71	0.55	0.39	0.71
	Wav2Vec2	0.59	0.50	0.68	0.52	0.37	0.68
	TRILLsson	0.61	0.50	0.73	0.58	0.40	0.75
	HuBERT-last	0.60	0.48	0.72	0.58	0.41	0.76
	HuBERT-layer20	0.62	0.50	0.74	0.61	0.44	0.77
	HuBERT-layer21	0.63	0.49	0.77	0.58	0.40	0.76
XGB	X-vector	0.60	0.53	0.67	0.54	0.44	0.64
	Wav2Vec2	0.59	0.47	0.71	0.57	0.41	0.73
	TRILLsson	0.65	0.56	0.74	0.62	0.48	0.76
	HuBERT-last	0.66	0.53	0.78	0.63	0.51	0.74
	HuBERT-layer20	0.66	0.56	0.76	0.66	0.67	0.65
	HuBERT-layer21	0.66	0.53	0.78	0.64	0.67	0.61
DNN	X-vector	0.59	0.49	0.68	0.56	0.49	0.62
	Wav2Vec2	0.58	0.55	0.62	0.56	0.36	0.74
	TRILLsson	0.64	0.48	0.80	0.62	0.54	0.71
	HuBERT-last	0.64	0.56	0.71	0.61	0.51	0.71
	HuBERT-layer20	0.69	0.58	0.80	0.64	0.52	0.75
	HuBERT-layer21	0.67	0.78	0.55	0.65	0.73	0.57
EF	H20+H21	0.66	0.68	0.64	0.63	0.66	0.60
	H20+Trls	0.67	0.68	0.65	0.64	0.61	0.67
	H20+H21+Trls	0.67	0.70	0.65	0.64	0.70	0.59
LF	H20+H21+Trls	0.59	0.24	0.94	0.60	0.78	0.43
ENS	H20+H21+Trls	0.70	0.66	0.74	0.68	0.68	0.68

and those between the thresholds as ‘uncertain’. For the best-performing ensemble models, both thresholds were optimized on the development set to maximize separation between positive and negative predictions while minimizing uncertain cases. Test-set results are reported in Table 4, with additional evaluation on the in-clinic dataset (Test2) described in Table 1.

Clinically, we propose a two-step procedure for uncertain predictions. If the first voice sample yields an uncertain result, possibly due to transient factors such as background noise or vocal artifacts, the patient is asked to record a second sample. If the second sample is also uncertain, the clinician should use an alternative screening instrument, such as the PHQ-9 or GAD-7.

Table 3. Behavioral health model performance – 233 negatives and 166 positives (42%)

Task	Model	UAR	Sens	Spec
Joint Training	FCDNN-HuBERT 20th	0.67	0.68	0.66
	FCDNN-TRILLsson	0.58	0.51	0.65
	XGB-HuBERT 20th	0.65	0.65	0.66
	XGB-TRILLsson	0.66	0.66	0.70
Post-hoc Aggregation	MDD OR GAD	0.69	0.73	0.66

Table 4. Best model performance with the ‘Uncertain’ class.

Task	Uncertain Rate	UAR	Sens	Spec
Depression	18%	0.71	0.72	0.70
Anxiety	21%	0.68	0.64	0.71
Behavioral	26%	0.70	0.76	0.65
Behavioral (Test2)	22%	0.74	0.68	0.80

5 Conclusions

In this paper, we present a speech-based behavioral health assessment system for detecting MDD and GAD from spontaneous speech. Our results suggest that a unified voice-based model has potential as a scalable, non-invasive screening tool, either as an alternative or complement to self-report questionnaires. This approach may be particularly valuable in settings where active self-reporting is impractical, such as telehealth consultations or passive monitoring scenarios involving ambient microphones for elderly populations. Future work will explore more sophisticated approaches for defining the ‘uncertain’ category through prediction probability calibration and will focus on validating the model against clinician diagnoses.

References

1. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: Wav2vec 2.0: a framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems, vol. 33 (NeurIPS 2020). Curran Associates Inc., Red Hook, NY, USA (2020)
2. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3451–3460 (2021)
3. Kroenke, K., Spitzer, R.L., Williams, J.B.W.: The phq-9: validity of a brief depression severity measure. J. Gen. Intern. Med. **16**(9), 606–613 (2001)
4. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., Talwalkar, A.: Hyperband: a novel bandit-based approach to hyperparameter optimization. J. Mach. Learn. Res. **18**(1), 6765–6816 (2017)

5. Shor, J., Venugopalan, S.: TRILLsson: distilled universal paralinguistic speech representations. In: *Proceedings of Interspeech 2022*, pp. 356–360. ISCA, Incheon, Korea (2022)
6. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: Robust DNN embeddings for speaker recognition. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333. IEEE (2018)
7. Spitzer, R.L., Kroenke, K., Williams, J.B., Löwe, B.: A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch. Intern. Med.* **166**(10), 1092–1097 (2006)
8. Wu, Y., Levis, B., Riehm, K.E., et al.: Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. *Psychol. Med.* **50**(8), 1368–1380 (2020)



Explainable Knowledge Graph Embeddings for Link Prediction and Path Attribution in Healthcare

Vijaya Laxmi Boddu¹, Akhil Chaudhary¹, Somayeh Kafaie²,
and Enayat Rajabi¹

¹ Management Science Department, Cape Breton University, Sydney, NS, Canada
{vijaya_boddu, akhil_choudhary, enayat_rajabi}@cbu.ca

² Department of Mathematics and Computing Science, Saint Mary's University,
Halifax, Canada
Somayeh.Kafaie@smu.ca

Abstract. With rapidly growing medical data, some of the biomedical knowledge graphs are often incomplete and unreliable due to scattered evidence, which limits clinical discovery. We developed an explainable link prediction framework that combines diverse models such as DistMult (symmetric relations), ComplEx (asymmetric relations), and Simple (diverse head–tail interactions). To reduce the bias of individual model ranking, we aggregate model scores using an ensemble strategy and rank candidate links. We retrieve supporting graph paths and quantify edge contributions using Shapley-based attribution, and validate predictions against biomedical literature using a domain-specific NLI model, which enables expert review. Moreover, experiments on 9,102 test triples showed improved ranking performance, with correct links appearing higher and more frequently among top results (MRR 0.299 vs. 0.269; Hits@10 0.484 vs. 0.443 for ComplEx). This provides a reliable and verifiable pipeline for biomedical knowledge graph completion.

Keywords: Link prediction · Biomedical Knowledge Graph embedding · Explainability

1 Introduction

Biomedical knowledge graphs (KGs) model relationships among drugs and diseases, but are often incomplete because evidence is distributed [2]. Besides, Knowledge graph embedding (KGE) models such as DistMult [13], ComplEx [11], and Simple [8] rank candidate links but do not explain their prediction reasons, limiting their reliability in healthcare applications [13]. Path-based approaches improve interpretability by identifying connections between entities, but remain qualitative, lack quantified edge contributions, and do not verify with external links [5]. We developed a workflow that combines KGE models ranking, SHAP-based path attribution, and Natural Language Inference

(NLI)-based validation to enable traceable link prediction. We also introduce a human-in-the-loop mechanism to make sure that each prediction is supported by both explanation and evidence. This is important in AI-driven clinical decision support, where interpretability and evidence validation are essential for building trust.

2 Related Work

Knowledge graph embedding (KGE) models such as DistMult, ComplEx, and Simple achieve strong link prediction performance but lack interpretability [13]. Path-based approaches improve transparency by identifying connections between entities, yet remain qualitative and do not quantify edge-level contributions [5]. SHAP provides theoretically grounded attribution, but its application to knowledge graph path reasoning remains limited [9]. Additionally, NLI-based methods enable evidence validation, though they are typically not integrated with prediction workflows [4]. Overall, existing approaches address prediction, explanation, or validation separately and are not connected. However, our work integrates KGE-based ranking, quantitative path-level attribution, and NLI-based validation into a single pipeline, enabling interpretable and evidence-supported link prediction.

3 Methodology

We propose an explainable workflow for KG completion that integrates prediction, explanation, and validation. Given a query triple $(h, r, ?)$, the task is to rank candidate entities t and provide both structural explanations and external evidence for the predicted link (h, r, t) .

Data Construction. We construct a biomedical knowledge graph from the Diseases Database [3], a cross-referenced medical dataset containing entities such as diseases, drugs, symptoms, and clinical findings. The resulting KG comprises 9,401 entities and 91,014 triples across 24 relation types, stored in a Neo4j database. For model development, the dataset is divided into training (80%), validation (10%), and testing (10%) sets.

Ensemble Prediction. The predictive component utilizes three KGE models, DistMult, ComplEx, and Simple, to score candidate triples (h, r, t) . The selection of DistMult, ComplEx, and Simple as the underlying KGE models is motivated by their diverse scoring functions and their status as established baselines in the studies. These models provide clear computations that allow us to separately evaluate the contributions of the explainability (SHAP) and validation (NLI) components. While advanced architectures such as TxGNN [6] or BioPathNet [5] perform link prediction, their inherent complexity and task-specific nature are not suitable for our objective. Furthermore, while individual models like ComplEx may excel in specific scenarios, with our *ensemble approach* we ensure

more stable and consistent performance across diverse queries. To improve ranking robustness, an ensemble score is computed by averaging the model outputs to rank candidate tails for a given query (h, r) .

Explainable Attribution. Paths help explain predictions as they show how entities and relationships are connected step by step. For each predicted triple, we retrieve supporting paths from h to t via Breadth-First Search (BFS). Each edge within a path is scored using the pre-trained KGE models, with the overall path-support score defined as the mean of its individual edge scores. We then apply Shapley-based attribution to quantify each edge’s specific contribution to the path score, providing a mathematically computed explanation of each prediction.

External Validation. To verify the biological plausibility of a predicted triple (h, r, t) , the relationship is converted into natural language and processed by a biomedical NLI model. This model evaluates the triple against evidence retrieved from PubMed, Europe PMC, and OpenAlex, returning a classification of AGREE, DISAGREE, or UNSURE.

Performance Evaluation. The framework is evaluated using standard ranking metrics, including Mean Reciprocal Rank (MRR) and Hits@ k . Additionally, qualitative case studies (Table 2) illustrate the interplay between SHAP attribution and NLI validation for the predicted triples.

Interpretability and Validation. SHAP is chosen due to its model-agnostic design and ability to provide quantitative edge-level contributions [9]. On the other hand, attention-based methods may not provide faithful explanations [1, 7], and LIME relies on local surrogate approximations rather than exact model contributions [10], making them less suitable for our framework, which delivers precise edge-level interpretation. In this context, SHAP enables granular attribution within KG paths by quantifying the contribution of individual edges. We integrated an NLI-based validation layer to this internal explanation to verify the biological plausibility of predicted triples (h, r, t) . We retrieve contextual evidence from PubMed, Europe PMC, and OpenAlex where entities h and t co-occur. The NLI model then evaluates the semantic relationship between the predicted statement and the retrieved evidence, returning agreement, contradiction, or uncertainty [4].

Data and Code Availability. The implementation and supplementary materials are available in the repository, with usage details in the README.md [12].

4 Results

Quantitative Evaluation. Table 1 reports link prediction performance on a held-out test set of 9,102 triples. The Ensemble achieves the best results (MRR 0.299, Hits@10 0.484).

Table 1. Link prediction performance on the held-out test set (higher values indicate superior performance).

Model	Test Size	MRR	Hits@1	Hits@3	Hits@10
DistMult	9102	0.255	0.172	0.279	0.424
Simple	9102	0.263	0.175	0.293	0.440
ComplEx	9102	0.269	0.182	0.297	0.443
Ensemble	9102	0.299	0.209	0.329	0.484

Qualitative Analysis for Interpretability (Case Study Examples). To illustrate the explainability and validation components, we present five representative case studies for the relation $r = \text{may cause}$ (Table 2). While quantitative evaluation is performed on the full dataset, these examples demonstrate how predictions are supported by paths, SHAP attribution, and NLI validation. Additional explainability metrics and detailed calculations are available in our repository README.md file [12].

Table 2. Five case studies for relation $r = \text{may cause}$

#	Head h	Predicted Tail t^*	Ensemble	Model	Path	$f(P)$	TopSHAP%	$\Delta f\%$	H_{norm}	NLI
1	Ethanol	Hyperlipidaemia	3.183	ComplEx	2	3.930	56.49	6.49	0.99	U (0.91)
2	Warfarin	Hypertension, systemic	2.130	ComplEx	3	4.053	57.42	33.80	0.86	U (0.00)
3	Methanol	Hyperglycaemia	2.745	ComplEx	3	4.522	43.17	6.60	0.97	U (1.00)
4	ABCD syndrome	Nystagmus	2.274	ComplEx	3	3.298	42.16	5.90	0.97	U (0.00)
5	AA amyloidosis	Renal failure, chronic	1.596	ComplEx	3	3.338	83.19	37.30	0.49	U (0.00)
Average across the 5 links			2.385		2.8	3.828	56.49	18.00	0.86	

Notes: Ensemble = Avg.score; Path = hops; $f(P)$ = mean score; TopSHAP = dominant edge; Δf = score change; H_{norm} = entropy; NLI = A (Agree), U (Unsure), D (Disagree), with confidence.

Insight: As demonstrated in Table 1, the **Ensemble** strategy yields the highest predictive accuracy, achieving an **MRR of 0.299** and **Hits@10 of 0.484**. While **ComplEx** emerges as the most robust standalone model (**MRR = 0.269**, **Hits@10 = 0.443**), the score-level fusion significantly enhances overall ranking quality by aggregating diverse latent representations.

The qualitative metrics in Table 2 reveal that supporting paths are consistently concise, ranging from **2 to 8 hops** (mean = **2.8**). SHAP-based attribution identifies a single dominant edge in **83.19%** of cases, while the remaining instances exhibit distributed contributions (average primary weight = **56.49%**). This variance suggests that while some predictions rely on a singular critical relationship, others are supported by a broader network of evidence. Finally, the NLI validation layer frequently returns an **UNSURE**. This suggests that some predicted relationships may represent novel or underexplored associations not yet explicitly documented in biomedical literature.

5 Discussion

This work presents an explainable framework for biomedical knowledge graph completion that combines prediction, explanation, and validation. The Ensemble improves performance over individual KGE models (Table 1), providing more stable predictions across queries. Supporting paths are short (2–8 hops), which improves interpretability. SHAP attribution helps identify whether a prediction is driven by a single dominant relationship or multiple contributing factors. The NLI layer reveals that path-supported predictions may not have clear supporting evidence in current research. This highlights the need to complement model predictions with external validation. Despite this, manual verification shows that most predictions remain biologically plausible (4 out of 5 cases), indicating the practical usefulness of the framework.

Limitations. Ranking metrics such as MRR, Hits@ k measure prediction performance but do not capture clinical relevance. While evaluation is performed on the full dataset, case studies are used to illustrate interpretability. In addition, NLI results depend on the availability and quality of evidence, and may return UNSURE when supporting information is limited or ambiguous. This is particularly important in healthcare, where explainability and evidence-backed reasoning are essential for clinical adoption.

6 Conclusion

We propose an explainable and evidence-validated KGE framework for healthcare discovery. This system integrates a multi-model ensemble (DistMult, ComplEx, Simple) for robust ranking, SHAP-based path attribution for quantitative transparency, and a biomedical NLI layer for literature-based corroboration. By unifying predictive ranking with structural and textual verification, our framework transforms “black-box” completions into a traceable, auditable process that supports expert-led hypothesis generation. Future work will focus on improving evidence retrieval quality and enhancing validation robustness for more reliable healthcare applications.

Acknowledgements. The work conducted in the study has been funded by NSERC (Natural Sciences and Engineering Research Council) Discovery Grant (RGPIN-2020-05869).

References

1. Agarwal, C., Queen, O., Lakkaraju, H., Zitnik, M.: Evaluating explainability for graph neural networks. *Sci. Data* **10**, 144 (2023)
2. Chandak, P., Huang, K., Zitnik, M.: Building a knowledge graph to enable precision medicine. *Sci. Data* **10**, 67 (2023)

3. Diseases Database: Diseases database ver 2.0 (2022)
4. Gu, Y., Tinn, R., Cheng, H., et al.: Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare* **3**(1), 1–23 (2021)
5. Hu, Y., Oleshko, S., Firmani, S., et al.: Path-based reasoning for biomedical knowledge graphs with biopathnet. *bioRxiv* p. 2024.06.17.599219 (2024)
6. Huang, K., Zitnik, M., et al.: Txgnn: Exploring large-scale knowledge graphs for drug discovery. *Nat. Mach. Intell.* (2023)
7. Jain, S., Wallace, B.C.: Attention is not explanation. In: *NAACL*, pp. 3543–3556 (2019)
8. Kazemi, S.M., Poole, D.: Simple embedding for link prediction in knowledge graphs. In: *NeurIPS*, vol. 31 (2018)
9. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: *NeurIPS*, pp. 4765–4774 (2017)
10. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you? explaining the predictions of any classifier. In: *KDD*, pp. 1135–1144 (2016)
11. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.: Complex embeddings for simple link prediction. In: *ICML*, pp. 2071–2080 (2016)
12. Vijayalaxmi-sys: Explainable KGE healthcare visibility (2026). <https://github.com/Vijayalaxmi-sys/Explainable-KGE-Healthcare-Visibility>
13. Yang, B., Yih, W.t., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. In: *ICLR* (2015)



LLM-as-a-Judge for Evaluating the Quality of Retrieval-Augmented Generation Systems

Leonardo Sanna¹(✉), Erica Solinas², and Mauro Dragoni¹

¹ Digital Health & Wellbeing Research Center, Fondazione Bruno Kessler, Trento, Italy

{lsanna,dragoni}@fbk.eu

² Università di Bologna, Bologna, Italy
erica.solinas@studio.unibo.it

Abstract. Assessing retrieval quality in Retrieval-Augmented Generation (RAG) pipelines is challenging due to ambiguity and information sparsity, and typically requires expert annotation, limiting scalability. We present an empirical evaluation of LLM-as-a-judge for document relevance in an Italian medical RAG pipeline on pregnancy and maternity care. LLMs achieve moderate agreement with human annotators. Label distribution comparisons show deviations from human judgments, with LLAMA exhibiting the largest discrepancies (Cramér’s V up to 0.416), while Mistral and DeepSeek align more closely. All models display a bias toward intermediate relevance labels, indicating reduced decisiveness. Nevertheless, LLM judgments remain consistent with cross-encoder scores, with significant differences across relevance classes ($p \approx 0$). These findings support LLMs as scalable and interpretable proxies for retrieval evaluation, while highlighting biases that must be addressed in production settings.

Keywords: Retrieval-Augmented-Generation · Large Language Models · LLM-as-a-judge · Pregnancy · Evaluation

1 Introduction

Retrieval-Augmented Generation (RAG) systems integrate information retrieval with Large Language Models (LLMs) to produce grounded, domain-aware responses. Evaluating these systems requires assessing both the relevance of retrieved documents and the extent to which generated outputs correctly leverage retrieved evidence. Previous studies have shown that traditional metrics such as retrieval accuracy and lexical similarity measures capture only limited aspects of system performance, often overlooking pragmatic relevance, partial informational support, and hallucination, particularly in high-stakes domains like healthcare [1]. In contrast to earlier approaches that treat retrieval and generation as separable evaluation problems, recent work highlights the need for

integrated evaluation frameworks that better reflect how retrieved evidence supports downstream generation. Recent research has explored the use of LLMs as automated evaluators (LLM-as-a-judge) as a scalable alternative to human annotation. A recent study [12] found that LLM-based evaluation can approximate human judgments in generation tasks. However, the reliability of LLM-based evaluation for retrieval-level assessment remains underexplored, and existing studies indicate that human supervision is still necessary to ensure robustness [6]. Building on this line of research, this paper investigates whether LLM-as-a-judge can reliably assess document relevance in a medical RAG pipeline for Italian maternity-care information. Specifically, we have two research questions (RQ). *RQ1*: How closely do LLM relevance judgments align with human annotations? *RQ2*: Do LLM relevance labels reflect underlying retrieval relevance scores?

2 Methods

Datasets. We use two complementary Italian datasets: one to construct the annotation gold standard and another to evaluate the domain knowledge of LLMs. The first dataset, ChatFAQ, serves as the Gold Standard. It is a pre-validated medical document repository used in a prior implementation [5]. It contains 1,512 expert-validated documents on maternity and newborn care, which were segmented into 43,390 text chunks. These chunks were then used to generate synthetic, context-grounded maternity-related questions using GPT-4o-mini, resulting in 216,950 question-passage pairs. To assess retrieval performance, a random subset of 1,000 questions was evaluated by verifying whether the correct source document was retrieved [8]. From this sample we extract a subset of 200 queries for our Gold Standard; each query has six retrieved documents (1,200 instances), annotated by two independent annotators on a three-level relevance scale (0 - Not Relevant, 0.5 - Partially Relevant, 1 - Relevant), with inter-annotator agreement of $\kappa = 0.37$ [3]. The second dataset, IMB [11], is used as an external benchmark for evaluating LLMs’ domain knowledge. It consists of IMB-MCQA, with 350 multiple-choice questions on obstetrics and gynecology, and IMB-QA, a collection of patient-doctor question-answer pairs, from which we use 13,232 real-world pregnancy-related questions.

Pipeline and Models. We use IMB-QA as the question source and ChatFAQ as the document repository, updating the original ChatFAQ pipeline [5] with a state-of-the-art RAG setup combining dense embeddings and BM25 [9], merged via Reciprocal Rank Fusion (RRF) [4], and removing query augmentation to reduce variability.

LLMs were selected from the Italian leaderboard [7], restricted to top-20 models in *Word-In-Context* and *Entailment* tasks¹. Further criteria were the open weights status and the number of parameters (max 70B). Six models were

¹ Leaderboard was accessed in January 2026.

evaluated against the IMB-MCQA to assess their domain knowledge using accuracy and Plurality Proportion Agreement (PPA) [10], with all six scoring above the thresholds in both metrics (>0.60 for accuracy and >0.80 for PPA). Hence we evaluate their capabilities as LLM-as-a-judge. Performance was measured as proposed by [2], which introduces dedicated metrics for LLM-as-a-judge on the Gold Standard dataset. We therefore calculate the *Advantage Probability* (p), i.e., the probability that the LLM annotations are as good as or better than those of a randomly chosen human annotator, and *Winning Rate* (w), i.e., the proportion of human annotators for whom the LLM achieves a statistically significant advantage. Three models were selected as shown in Table 1: llama-3.3-70B-instruct (henceforth LLAMA); mistral-small-24B-instruct-2501 (henceforth Mistral); deepseek-r1-distill-llama-70 (henceforth DeepSeek).

Table 1. Final summary of model performance. Advantage Probability p , and Winning Rate w demonstrate that LLAMA, DeepSeek, and Mistral are suitable to serve as LLM-as-a-judge. Agreement displays mean agreement with human annotators.

Model	Agreement	PPA	p	w
llama-3.3-70B-instruct	0.618	0.790	0.844	1.0
mistral-small-24B-instruct-2501	0.578	0.793	0.806	1.0
deepseek-r1-distill-llama-70B	0.570	0.745	0.797	1.0
qwen2.5-vl-32B-instruct	0.545	0.768	0.781	0.5
gemma-3-12B-it	0.544	0.753	0.777	0.5
gemma-3-27B-it	0.500	0.803	0.737	0.0

3 Results

To address RQ1, we compared label distributions between human annotators and LLMs and quantified their differences using pairwise χ^2 tests with Cramér’s V (Table 2). While human annotators showed moderate agreement ($V = 0.25$), all models exhibited deviations from human labeling patterns. In particular, LLAMA showed the largest divergence (up to $V = 0.416$), whereas Mistral and DeepSeek were consistently closer to human distributions (lowest V values down to 0.147). A strong bias toward label B emerged on all three models, especially with LLAMA (97.0%) and Mistral (90.0%), while DeepSeek chose this option in 73% of the cases. Nonetheless we observe closer alignment with human distributions for Mistral and DeepSeek indicating they may be more suitable for tasks requiring consistency with human judgment. Notably, this contrasts with the Gold Standard results, where LLAMA performed best, indicating that large-scale distributional alignment cannot be inferred from smaller evaluations.

To address RQ2, we examined whether LLM-assigned labels reflect underlying retrieval relevance by comparing them with cross-encoder scores using the

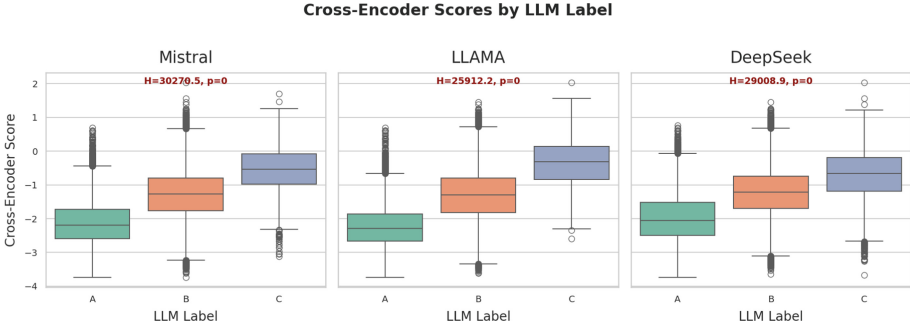


Fig. 1. Cross-Encoder scores by LLM-assigned labels (A, B, C) for Mistral, LLAMA, and DeepSeek. Each panel corresponds to one model. The Kruskal-Wallis H statistics and p -value are reported above each boxplot, showing significant differences across label categories for all models ($p \approx 0$).

Kruskal-Wallis test. We use the cross-encoder model *bge-reranker-v2-m3*², which jointly encodes the query and candidate documents to compute relevance scores. As shown in Fig. 1, all models exhibit highly significant differences across label categories ($p \approx 0$). The cross-encoder score distributions align with this pattern: documents labeled A (Not Relevant) have the lowest average scores, while those labeled C (Relevant) have the highest. Label B (Partially Relevant), although the most frequently assigned, consistently falls between A and C in terms of average score. This shows that LLM judgments remain systematically aligned with

Table 2. Pairwise χ^2 tests comparing label distributions between annotators and LLMs. Cramér’s V quantifies the strength of association: lower values indicate more similar distributions. It should be noted that all chi-square values are statistically significant with $p \approx 0$.

Pair	χ^2	Cramér’s V
Annotator 1 vs Annotator 2	25.28	0.251
Annotator 1 vs Mistral	658.48	0.221
Annotator 1 vs LLAMA	2323.39	0.416
Annotator 1 vs DeepSeek	459.37	0.185
Annotator 2 vs Mistral	388.52	0.170
Annotator 2 vs LLAMA	1459.83	0.330
Annotator 2 vs DeepSeek	289.71	0.147
Mistral vs LLAMA	843.13	0.179
Mistral vs DeepSeek	572.21	0.147
LLAMA vs DeepSeek	2432.17	0.303

² <https://huggingface.co/BAAI/bge-reranker-v2-m3>.

semantic relevance scores. The significance is also not affected when removing outliers using Inter-Quartile-Range (IQR). Compared to cross-encoder scores, which lack universal thresholds and rely on dataset-specific calibration, LLMs provide directly interpretable, context-sensitive labels. However, their strong bias toward B-label (i.e., partially relevant) judgments limits decisiveness in retrieval scenarios, reflecting both model uncertainty and the task’s intrinsic ambiguity, as evidenced by notable disagreement among human annotators. Additional limitations include dataset size imbalance, cost and latency differences across models, and the use of single-label annotations, which may overlook graded or aggregated relevance signals.

4 Conclusion

Our study shows that LLM-as-a-judge show moderate but consistent alignment with human relevance judgments. Despite differences in label distributions, their evaluations remain strongly and systematically aligned with semantic relevance signals derived from cross-encoder scores. Several limitations should be noted. The dataset is relatively small and shows only moderate inter-annotator agreement, reflecting the intrinsic ambiguity of graded relevance tasks. In addition, reliance on a single cross-encoder model may introduce bias in semantic comparisons, and single-label annotations may oversimplify complex relevance relationships. Nevertheless, these limitations do not materially affect the observed trends, which remain consistent across models and evaluation dimensions.

Acknowledgments. Hub Life Science – Digital Health (LSH-DH) PNC-E3-2022-23683267 - Progetto DHEAL-COM – CUP: C63C22001970001”, finanziato dal Ministero della Salute nell’ambito del Piano Nazionale Complementare Ecosistema Innovativo della Salute” - Codice univoco investimento: PNC-E.3. This publication reflects only the authors’ view and the Italian Ministry of Health is not responsible for any use that may be made of the information it contains.

References

1. Brown, A., Roman, M., Devereux, B.: A systematic literature review of retrieval-augmented generation: techniques, metrics, and challenges. *Big Data Cogn. Comput.* **9**(12) (2025). <https://doi.org/10.3390/bdcc9120320>, <https://www.mdpi.com/2504-2289/9/12/320>
2. Calderon, N., Reichart, R., Dror, R.: The alternative annotator test for LLM-as-a-judge: how to statistically justify replacing human annotators with LLMs. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16051–16081 (2025)
3. Cohen, J.: A coefficient of agreement for nominal scales. *Educ. Psychol. Measur.* **20**(1), 37–46 (1960)
4. Cormack, G.V., Clarke, C.L., Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759 (2009)

5. Ghanbari Haez, S., et al.: A retrieval-augmented generation strategy to enhance medical chatbot reliability. In: Finkelstein, J., Moskovitch, R., Parimbelli, E. (eds.) AIME 2024. LNCS, vol. 14844, pp. 213–223. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-66538-7_22
6. Jang, D., et al.: Medtutor: a retrieval-augmented LLM system for case-based medical education. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 319–353 (2025)
7. Magnini, B., et al.: A leaderboard for benchmarking LLMs on Italian. In: Bosco, C., Jezek, E., Polignano, M., Sanguinetti, M. (eds.) Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), pp. 636–646. CEUR Workshop Proceedings, Cagliari, Italy (2025)
8. Magnini, M., et al.: Automated machine learning to enhance knowledge retrieval in retrieval-augmented generation pipelines. In: Bruno, P., Calimeri, F., Cauteruccio, F., Dragoni, M., Stella, F., Terracina, G. (eds.) HC_AIXIA_HYDRA 2025. CCIS, vol. 2830, pp. 429–440. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-16708-8_34
9. Robertson, S., Zaragoza, H., Taylor, M.: Simple bm25 extension to multiple weighted fields. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, pp. 42–49 (2004)
10. Robinson, J., Rytting, C.M., Wingate, D.: Leveraging large language models for multiple choice question answering (2023)
11. Romano, A., Riccio, G., Barone, M., Postiglione, M., Moscato, V.: IMB: an Italian medical benchmark for question answering. In: Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), pp. 995–1006 (2025)
12. Sarvari, P., Al-Fagih, Z.: Rapidly benchmarking large language models for diagnosing comorbid patients: comparative study leveraging the LLM-as-a-judge method. *JMIRx Med* **6**, e67661 (2025)



Flow Builder: No-Code Conversation Design Tool for Digital Therapeutics in Psychology

Leonardo Sanna^(✉), Mattia Franzin, Mauro Dragoni, and Claudio Eccher

Digital Health and Wellbeing Research Center, Fondazione Bruno Kessler,
Trento, Italy

{lsanna,mfranzin,dragoni,cleccher}@fbk.eu

Abstract. The effectiveness of Digital Therapeutics (DTx) relies on understanding and shaping human behavior, yet development is hindered by fragmented conversational scripts and repetitive chatbot implementation. Integrating advanced agentic components such as Large Language Models (LLMs) and multi-agent systems further complicates clinical safety, data provenance, and ethical risks. We present *Flow Builder*, a no-code platform within the MAESTRO (Multi-Agent Enhanced System for Therapy Routine Orchestration) framework, designed to standardize conversation design. Leveraging a modular block-based hierarchy with over seventeen specialized processors, from text delivery to semantic retrieval and multi-step LLM reasoning, Flow Builder enables clinicians to focus on therapeutic logic while ensuring safe, consistent, and scalable agentic execution.

Keywords: Conversational Agents · Digital Therapeutics · No-code Design Tool

1 Introduction

Digital Therapeutics (DTx) rely on behavioral design principles [5] to drive user engagement and often include conversational solutions such as chatbots. However, designing such interventions frequently encounters bottlenecks between the professionals responsible for conversational design and the development teams.

To the best of our knowledge, there are currently no fully open-source, no-code tools specifically tailored to design and deploy conversation-based solutions that enhance DTx platforms. More specifically, there is a need to have a tool that, in line with the state of the art, allows the creation of chatbots using artificial intelligence (AI) [2], that, however, should comply with the requirements of psychological interventions that demand fine-grained control over the conversational flow.

Current available tools include Rasa¹ (open-source, low-code) or Botpress², which support hybrid rule-LLM flows but require coding to build the graphs and

¹ <https://rasa.com/>.

² <https://botpress.com/>.

lack psychological intervention constraints. Commercial tools (e.g., Dialogflow³) offer drag-and-drop interfaces but they are not open-source and cannot be customized.

Existing literature includes platforms like Clinica [3], which provides automated pipelines for multimodal neuroimaging data processing and analysis, and Graph2Bots [1], which extracts conversational flows from human-human conversations. Perhaps the closest example is Conversation Learner [4], a tool that uses dialogue flows to train dialogue manager systems. Nonetheless, these applications are neither focused on DTx nor do they provide reusable conversation elements to facilitate rapid design of conversational flows.

In this paper, we propose the Flow Builder platform. The design was guided by the need to represent conversational systems using abstractions that are both expressive and structurally explicit. Rather than treating conversational behavior as either a fixed sequence of states or an emergent property of generative models, our framework adopts a construction-oriented perspective: complex systems are assembled from simpler components, enabling users to define rule-based conversation flows in which Large Language Models (LLMs) can serve both as content generators and reasoning engines.

2 Framework

The *Flow Builder* is a no-code platform for creating complex, rule-based conversational workflows, particularly for clinical or research interventions. It provides a drag-and-drop interface that abstracts conversation logic while ensuring robust execution and data integrity through a structured backend and standardized validation schemas.

At its core, *Flow Builder* integrates a **FastAPI** backend with a **MongoDB** database. Each workflow designed in the visual interface is automatically exported as a standardized `.flow.json` file, accompanied by an `info.json` metadata file that describes the intervention, its versioning, and other essential parameters. This JSON-based representation ensures workflows can be easily replicated, audited, and shared.

To guarantee correctness and consistency, *Flow Builder* employs **Pydantic v2** to validate all block connections and configurations.

2.1 Conversation Structure

Flow Builder employs 14 specialized block types to support diverse interaction requirements. Broadly they can be defined in backend blocks (blocks handling logic and transitions) and output blocks (blocks delivering content). Output blocks present information to users via the front end, supporting text, interactive buttons, and multimedia elements such as images and video.

Currently the Flow Builder supports these functional blocks.

³ <https://dialogflow.cloud.google.com/>.

Table 1. Summary of Flow Builder Blocks by Functional Category

Category	Block Type	Function / Description
Content	Text Block	Delivers multilingual textual content
	Questionnaire Block	Integrates standardized assessments or external forms directly into the flow.
Agentic	Augmented Block	Generates messages using a LLM in a single-turn interaction
	LLM Chain Block	Executes sequences of chained LLM reasoning steps where outputs inform subsequent steps. The goal is to generate a message for a single-turn interaction
	LLM Reasoning Block	Produces structured outputs from LLMs, constrained to predefined formats in order to handle dynamic transitions in the flow.
	Retrieval Block	Implements Retrieval-Augmented Generation, enabling retrieval from vector databases and providing additional pre-validated context to generate LLM responses.
	Miniconv Block	Supports short, controlled multi-turn micro-conversations with a maximum number of turns and auto-supervision with guardrails.
Logic	Compute Block	Performs temporal and computational operations, such as calculating future dates or session times.
	Branch Block	Enables non-linear flows by evaluating variables against conditions to provide personalized paths.
	Schedule Block	Schedules the activation of conversational flows according to user input
	Placeholder Block	Acts as a synchronization or convergence point without functional execution.
Memory	Read Block	Retrieves persistent data from the database (e.g., username or other conversation variables)
	Store Block	Persists a conversational variable into the database
	Store History Block	Persists a defined number of messages into the database as a variable in order to provide conversational context to agentic blocks

1. Content Blocks: namely blocks of predefined information delivering predefined messages or questionnaires
2. Agentic Blocks: namely blocks dedicated to LLM integration. They can be output blocks, thus used to generate messages (both single and multi-turn) or backend blocks for performing operations and reasoning steps using LLM agents
3. Logic Blocks: they are dedicated to expressing conversation logic, such as scheduling logic and branching between different paths in the conversation
4. Memory Blocks: Blocks that performs operations on state variables (store and read)

Details on blocks functions are listed in Table 1.

By combining back-end and output blocks, the *Flow Builder* provides a flexible yet structured approach to conversational design. Researchers and clinicians can visually construct multi-turn interactions, incorporate conditional logic, and leverage AI-driven content generation, all without writing any code. This design ensures that conversational workflows are both expressive and reproducible, while maintaining transparency and clinician oversight.

3 Designing the Intervention

The Flow Builder enables users to create complex workflows without coding by arranging modular blocks visually on a canvas. Blocks can be dragged from the toolbox and connected to define the intervention’s flow. Each block is configured through intuitive property panels, allowing the user to assign variables, prompts, and conditional logic directly within the interface. This immediate feedback ensures that the structure and logic of the intervention are correct. Figure 1 shows an example flow inside the Flow Builder tool.

3.1 Interacting with the Flow

After designing the intervention, users can interact with it in real-time (see Fig. 2). The Flow Builder’s *preview mode* simulates the conversation, allowing the verification of decision paths and the expected behavior of the workflow. Users can interact with the designed workflow as end users. The outputs, which may include multimedia and interactive buttons, are rendered exactly as configured, enabling comprehensive testing prior to deployment.

Finally, the flow can be exported in JSON format to ensure reproducibility.

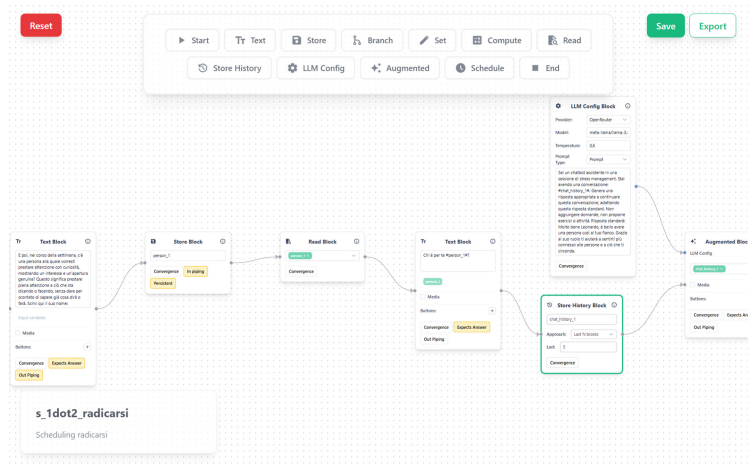


Fig. 1. Visual design of an intervention using the drag-and-drop canvas. Blocks are arranged to define conversational flow, conditional branching, and provide LLM configurations including prompts.

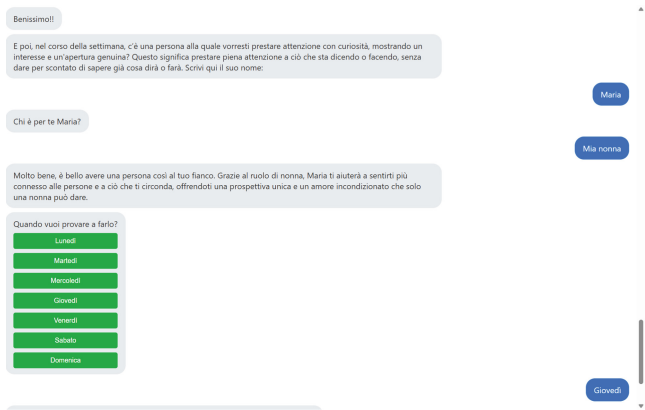


Fig. 2. Real-time interaction with the designed intervention of a stress management intervention (Self-Help+).

4 Conclusion

This use case demonstrates that the Flow Builder provides a seamless environment for designing, configuring, and testing interventions. Users can define interventions visually with minimal technical expertise, implement complex branching and AI-driven reasoning, and verify the interaction in real-time, ensuring reliability and clinician-in-the-loop oversight before deployment.

Acknowledgments. Hub Life Science – Digital Health (LSH-DH) PNC-E3-2022-23683267 - Progetto DHEAL-COM – CUP: C63C22001970001, finanziato dal Ministero della Salute nell’ambito del Piano Nazionale Complementare Ecosistema Innovativo della Salute - Codice univoco investimento: PNC-E.3. This publication reflects only the authors’ view and the Italian Ministry of Health is not responsible for any use that may be made of the information it contains.

References

1. Bouraoui, J.L., Le Meitour, S., Carbou, R., Barahona, L.M.R., Lemaire, V.: Graph2bots, unsupervised assistance for designing chatbots. In: Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, pp. 114–117 (2019)
2. Laymouna, M., Ma, Y., Lessard, D., Schuster, T., Engler, K., Lebouché, B.: Roles, users, benefits, and limitations of chatbots in health care: rapid review. *J. Med. Internet Res.* **26**, e56930 (2024)
3. Routier, A., et al.: Clinica: an open-source software platform for reproducible clinical neuroscience studies. *Front. Neuroinform.* **15**, 689675 (2021)
4. Shukla, S., et al.: Conversation learner-a machine teaching tool for building dialog managers for task-oriented dialog systems (2020). [arXiv:2004.04305](https://arxiv.org/abs/2004.04305) arXiv preprint
5. Voorheis, P., et al.: Integrating behavioral science and design thinking to develop mobile health interventions: systematic scoping review. *JMIR Mhealth Uhealth* **10**(3), e35799 (2022)



MAESTRO: a Framework for Trustworthy Integration of LLMs in Psychological Digital Interventions

Leonardo Sanna, Mattia Franzin, Simone De Carli, Marco Bolpagni, Simone Casazza, Silvia Rizzi, Claudio Eccher, and Mauro Dragoni^(✉)

Center for Digital Health and Wellbeing, Fondazione Bruno Kessler, Trento, Italy
{lsanna,mfranzin,srizzi,cleccher,dragoni}@fbk.eu

Abstract. The integration of Large Language Models (LLMs) into conversational systems has greatly enhanced their expressive capabilities but also introduced challenges in structure, control, and reproducibility. Current design approaches often fall between rigid, rule-based systems that are predictable but limited in scalability, and fully LLM-driven models that enable natural interaction but lack transparent logic. We present MAESTRO (Multi-Agent Enhanced System for Therapy Resources Orchestration), a framework for the trustworthy integration of LLMs in complex settings such as psychological digital interventions. Central to MAESTRO is a flow-based abstraction model that represents conversational logic as assemblies of modular, reusable blocks. This approach allows designers to construct explicit interaction flows in which LLMs can serve as bounded-reasoning components, preserving predictability while enabling controlled flexibility.

Keywords: Multi-agent systems · Trustworthy AI · Large Language Models · Digital Therapeutics · Psychology

1 Introduction

Conversational systems are increasingly deployed in complex domains such as decision support, coaching, and psychological interventions. Recent advances in Large Language Models (LLMs) have significantly enhanced their expressive capabilities, enabling flexible language generation and interpretation. However, these advances also amplify longstanding challenges related to structure, control, inspectability, and reproducibility. Existing approaches tend to polarize between rule-based systems, which provide explicit control but limited flexibility, and end-to-end LLM-driven systems, which enable natural interaction while obscuring conversational logic and limiting reproducibility. In practice, designers must balance structure and flexibility, a trade-off that remains poorly supported. In our experience in designing digital interventions, we observed that conversational systems repeatedly rely on recurring structural patterns such as sequencing, conditional progression, and bounded reasoning that are rarely formalized and

instead reimplemented ad hoc, reducing transparency and increasing development effort. In response, this paper introduces MAESTRO, a flow-based abstraction framework that models conversational logic as composable and inspectable interaction flows, integrating LLMs as bounded reasoning components.

Recent studies have begun exploring the trustworthy integration of LLMs into psychological interventions [6], although clinical effects remain understudied and have not yet been shown to be superior to rule-based systems [2]. In this work, we argue that this lack of clinical studies is caused mainly by the absence of a common tech stack that can allow researchers in the field to experiment with different interventions and reproduce the same dialogue structures in different clinical and cultural settings.

Existing literature on psychological interventions using chatbots [1,4,5,7] shows that three main types of systems are employed: rule-based chatbots, hybrid chatbots that integrate NLP techniques, and generative AI-powered chatbots. These conversational systems are used in diverse contexts, including delivering cognitive-behavioral therapy, providing psychoeducation, screening for symptoms, supporting self-management of chronic conditions, and offering social or emotional support. They target various populations, from adolescents to older adults, and are deployed through mobile applications, web platforms, or messaging services.

Despite this diversity, a key limitation remains the absence of a shared abstraction layer that enables reproducible conversational structures and systematic experimentation across contexts. Building on these gaps, MAESTRO provides a compositional framework that unifies structure and flexibility, supporting reusable and reproducible conversational design.

2 System

As anticipated in Sect. 1, the design of MAESTRO was guided by the need to represent conversational systems using abstractions that are both expressive and structurally explicit. Rather than treating conversational behavior as either a fixed sequence of states or an emergent property of generative models, the framework adopts a construction-oriented perspective in which complex systems are assembled from simpler components. This perspective is captured by the LEGO metaphor. Each conversational abstraction is treated as a LEGO block: a self-contained, well-defined unit that exposes clear connection points but does not encode assumptions about the overall system in which it will be used. Individual blocks are intentionally simple, and expressiveness arises through composition rather than through the complexity of any single component.

Importantly, the metaphor emphasizes assembly over automation. Just as physical bricks do not prescribe the structure of the final object, MAESTRO does not impose predefined conversational trajectories or semantic interpretations. Instead, designers construct conversational systems by explicitly assembling blocks into flows, making structural decisions visible and inspectable. This approach supports reuse, iterative refinement, and shared understanding among collaborators, while avoiding hidden control logic or implicit behavior.

This modular separation enables deployment across multiple front ends and the concurrent orchestration of external services required for code execution, as shown in Fig. 1. This is a key advantage, as existing dialogue managers couple conversational logic with front-end-specific implementations.

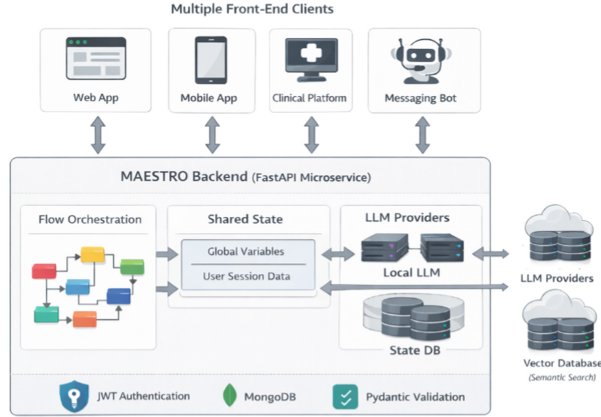


Fig. 1. Visualization of the MAESTRO system components. The flow abstracts a conversation as an executable graph, where block configurations and runtime variables orchestrate the interaction with external components.

3 Components

3.1 Flow Engine

At the core of MAESTRO is a flow-based abstraction model that represents a conversational system as a directed graph of executable components. A flow defines the control structure of a conversation by specifying how individual conversational blocks are connected and executed over time. Each block in a flow is the node of an executable graph while edges define the possible transitions between the blocks. Transitions may be unconditional or depend on explicit conditions evaluated at runtime. This representation makes conversational structure explicit and inspectable, enabling designers to reason about sequencing, branching, and termination without relying on implicit control logic.

The system is composed by two main nodes. *Flow Engine* (Executor Node) executes the conversational routines expressed in the flows, and the execution is sequential. Flow can be *main flow*, i.e., the list of conversation blocks that compose the main conversation, and *secondary flows*, namely parallel conversational flows that might be triggered during the execution of the conversation. Conversational patterns are defined using JSON-based flow files, enabling reproducible, standardized representations of therapeutic interventions.

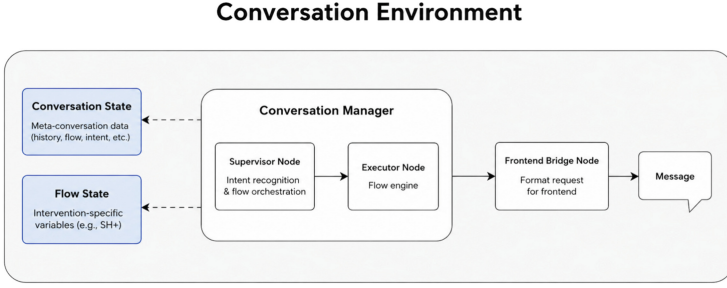


Fig. 2. Schema of the conversation manager of MAESTRO, composed by a Supervisor node and an Executor Node (Flow Engine).

3.2 Supervisor Node

To ensure trustworthy and flexible model integration, MAESTRO introduces an abstraction layer that oversees conversational workflows and enables dynamic flow orchestration, as illustrated in Fig. 2. This node determines flow transitions between main and secondary flows. These transitions can be triggered by heterogeneous conditions evaluated through rule-based algorithms, deep learning models, or LLMs.

- **Rule-Based Flow Selection:** In rule-based flow selection, the Supervisor evaluates a configurable set of rules to determine whether to transition the current session to an alternative conversational flow. These decisions rely on session-level attributes, tracked user variables, and real-time user input.
- **Dynamic Flow Selection:** Dynamic flow selection is based on LLMs’ semantic interpretation of user input. The Supervisor has access to a rule defined by a prompt and to a structured output schema containing:
 - A boolean **match** flag indicating whether the rule condition is satisfied.
 - A textual **reason** field providing a concise explanation supporting the decision outcome.

4 Evaluation Plan: Re-Engineering Self-Help+

To evaluate MAESTRO, we are re-engineering Self-Help+ (SH+), a stress-management and behavior change intervention previously implemented as a rule-based chatbot [3] and later extended with LLM capabilities [8]. Prior versions tightly coupled conversational logic with frontend implementations, limiting flexibility for iteration by intervention designers.

The target users of this system are primarily psychologists and researchers designing digital behavior change interventions, who need to author, inspect,

and iteratively refine structured therapeutic conversations without requiring full-stack engineering support. The target use cases include deployment of evidence-based behavior change interventions (e.g., stress management and psychoeducation programs) delivered through conversational agents in real-world or experimental settings.

Using MAESTRO, SH+ units are being migrated into modular conversational flows. The evaluation focuses on two technical dimensions. First, reduction of boilerplate code: MAESTRO introduces reusable backend components and externalized prompt management, reducing redundant implementation effort and enabling rapid iteration of intervention content. Second, structural transparency: representing SH+ as a graph of executable components makes intervention logic explicit and inspectable, enabling non-engineer stakeholders to understand and refine conversational and therapeutic structure.

The re-engineered system will undergo a pilot study assessing usability for intervention designers and fidelity to the original therapeutic protocol. Evaluation metrics will include usability scores (e.g., SUS), implementation effort (e.g., code reuse and reduction of duplicated logic), and fidelity indicators (e.g., adherence to prescribed intervention steps and consistency of delivered content). Preliminary expert review with domain specialists suggests feasibility and improved clarity of intervention structure, supporting further empirical validation in realistic deployment scenarios.

Acknowledgments. Hub Life Science – Digital Health (LSH-DH) PNC-E3-2022-23683267 - Progetto DHEAL-COM – CUP: C63C22001970001, finanziato dal Ministero della Salute nell’ambito del Piano Nazionale Complementare Ecosistema Innovativo della Salute - Codice univoco investimento: PNC-E.3. This publication reflects only the authors’ view and the Italian Ministry of Health is not responsible for any use that may be made of the information it contains.

References

1. Denecke, K., Rivera Romero, O., Wynn, R., Gabarron, E.: Persuasive chatbot-based interventions for depression: a list of recommendations for improving reporting standards. *Front. Psych.* **16**, 1429304 (2025)
2. Du, Q., Ren, Y., Meng, Z.I., He, H., Meng, S.: The efficacy of rule-based versus large language model-based chatbots in alleviating symptoms of depression and anxiety: systematic review and meta-analysis. *J. Med. Int. Res.* **27**, e78186 (2025)
3. Fietta, V., et al.: A chatbot-based version of the world health organization-validated self-help plus intervention for stress management: co-design and usability testing. *JMIR Hum. Factors* **11**, e64614 (2024)
4. Laymouna, M., Ma, Y., Lessard, D., Schuster, T., Engler, K., Lebouché, B.: Roles, users, benefits, and limitations of chatbots in health care: rapid review. *J. Med. Internet Res.* **26**, e56930 (2024)
5. Li, H., Zhang, R., Lee, Y.C., Kraut, R.E., Mohr, D.C.: Systematic review and meta-analysis of ai-based conversational agents for promoting mental health and well-being. *NPJ Digital Med.* **6**(1), 236 (2023)

6. Ni, Y., Cao, Y.: Exploring ChatGPT's capabilities, stability, potential and risks in conducting psychological counseling through simulations in school counseling. *Mental Health Digital Technol.* **2**(3), 213–239 (2025)
7. Permatasari, R.I., Artha, D.P., Wiratama, B.S., Wulandari, H.: Overview of chatbot usage on mental health: a scoping review. In: *BIO Web of Conferences*, vol. 132, p. 05002. EDP Sciences (2024)
8. Sanna, L., et al.: LLM-enriched finite-state chatbots for mental health support: a case study on self-help+. In: *International Conference on Artificial Intelligence in Medicine*, pp. 386–396. Springer (2025)



Inferring Chronic Treatment Onset from ePrescription Data: A Renewal Process Approach

Pavlin G. Poličar¹(✉), Dalibor Stanimirović², and Blaž Zupan¹

¹ Faculty of Computer and Information Science, University of Ljubljana,
1000 Ljubljana, Slovenia

pavlin.policar@fri.uni-lj.si

² Faculty of Public Administration, University of Ljubljana, 1000 Ljubljana, Slovenia

Abstract. Longitudinal electronic health record (EHR) data are often left-censored, making diagnosis records unreliable for determining disease onset. In contrast, outpatient prescriptions form renewal-based trajectories that provide a continuous signal of disease management. We propose a probabilistic framework that models prescription dynamics as a renewal process and detects transitions from sporadic to sustained therapy via change-point detection between a Poisson (sporadic) and a regime-specific Weibull (sustained) renewal model. Using a Slovenian nationwide ePrescription dataset of 2.4 million individuals, we show the approach yields more temporally plausible onset estimates than naive rule-based triggering, reducing implausible early detections under strong left censoring.

Keywords: Renewal processes · Change-point detection · Temporal point processes · EHR phenotyping

1 Introduction

Longitudinal electronic health record (EHR) data are frequently left-censored due to the gradual introduction of digital healthcare systems and incomplete migration of historical records [1, 4]. In Slovenia, national e-prescribing achieved near-complete coverage after its introduction in 2016, whereas diagnosis recording was adopted more gradually, increasing from 350,000 entries in 2016 to over 10 million by 2020. These low early counts primarily reflect incomplete system adoption rather than true disease incidence. Because diagnoses are often recorded only once and not systematically updated, historical disease onset is frequently missing or unreliable, and diagnosis data may contain inconsistencies and coding errors with reported misclassification rates up to 20% [3]. In contrast, prescription data are renewal-based and must be periodically re-issued, providing a continuous and more reliable signal of ongoing disease management.

Treatment patterns therefore provide a complementary signal for inferring disease onset, yet many phenotyping approaches rely on simple heuristics that

ignore prescribing dynamics. This can affect cohort construction and introduce data leakage, for example when treatment signals precede recorded diagnoses [2, 5]. Modeling prescription events as stochastic temporal processes allows detection of chronic treatment onset as a shift in prescribing dynamics. Chronic noncommunicable diseases are particularly suitable for this approach, as they are characterized by sustained, structured prescribing patterns, whereas acute conditions typically generate sparse or short-lived treatment signals.

We infer treated-phenotype onset by modeling longitudinal prescription trajectories as a renewal process. We apply change-point detection to identify transitions from sporadic to sustained therapy, formulated as a transition from a Poisson (sporadic) regime to a Weibull renewal (sustained therapy) regime. We evaluate this approach against rule-based triggering and show that renewal-based modeling yields more temporally plausible onset estimates under strong left censoring.

2 Data

We analyzed 2.4 million patients with 101 million outpatient prescriptions and 58 million diagnoses collected in Slovenia between 2016 and 2022. Women represent 60% of the population, the median patient age is 47, and patients have a median of 21 prescriptions and 16 diagnoses. Nearly all patients received at least one prescription, and all patients have at least one recorded diagnosis. Each prescription includes an acute/chronic label and a renewability flag indicating intended treatment duration and dispensing policy. Standard prescriptions cover about three months, while renewable prescriptions allow repeated dispensing within one year for stable long-term therapy.

Across the dataset, approximately 43 million prescriptions are labeled chronic and 58 million acute, including 11 million chronic-renewable, 32 million chronic non-renewable, 2 million acute-renewable, and 55 million acute non-renewable. Neither administrative field is treated as ground-truth evidence of disease state: acute/chronic labels and renewability reflect prescribing intent and local practice rather than stable clinical phenotypes and should therefore be interpreted as noisy administrative annotations. All prescriptions are mapped to active ingredient level-5 ATC codes to reduce variability due to brand-specific prescribing and ensure consistent aggregation across equivalent formulations.

3 Methods

We infer treated-phenotype onset from longitudinal outpatient prescription data by detecting transitions from sporadic to sustained prescribing and aggregating drug-level signals into disease-level phenotypes.

For each patient–drug pair, we observe prescription times $t_1 < t_2 < \dots < t_{n+1}$ and represent them through inter-arrival times $\tau_i = t_{i+1} - t_i$. We

model sporadic prescribing using a homogeneous Poisson process with exponentially distributed inter-arrival times, $p(\tau \mid \lambda) = \lambda e^{-\lambda\tau}$, capturing irregular events such as acute treatments or administrative noise. In contrast, sustained therapy is characterized by repeated refills at approximately regular intervals. To capture this regularity, we use a Weibull renewal process, $p(\tau) = \frac{k}{\lambda}(\tau/\lambda)^{k-1}e^{-(\tau/\lambda)^k}$, which generalizes the exponential distribution: $k \approx 1$ corresponds to Poisson-like behavior, while $k > 1$ models increasingly regular spacing between events. The scale parameter λ determines the characteristic refill interval. To account for dispensing policy, we estimate separate population-level Weibull parameters $(k_{d,r}, \lambda_{d,r})$ for each drug d and prescribing regime $r \in \{\text{renewable}, \text{non-renewable}\}$. Parameters are estimated by maximum likelihood from positive training-cohort inter-arrival times within each drug-regime group and are then kept fixed during patient-level change-point inference.

We infer onset of sustained therapy by modeling each prescription trajectory as a two-regime process with a single change-point. Given inter-arrival times $\tau_1, \dots, \tau_n$ with associated regimes r_i , the log-likelihood of a candidate change-point c is

$$\ell(c) = \sum_{i < c} \log p_{\text{null}}(\tau_i) + \sum_{i \leq c} \log p_{\text{chr}}(\tau_i \mid r_i), \quad (1)$$

where p_{null} is the Poisson model and p_{chr} is the regime-specific Weibull model. The estimated change-point is $\hat{c} = \arg \max_c \ell(c)$ and is accepted only if it improves the null model likelihood by a threshold ϵ . The corresponding calendar time is interpreted as treated-phenotype onset, capturing the transition into sustained therapy while avoiding spurious early detections driven by isolated prescriptions.

To construct disease-level treated phenotypes, we link drugs to diagnoses using temporal co-occurrence. For each inferred drug-level onset, we collect diagnoses within a window of three months before and 12 months after onset and aggregate these associations across patients. Drug-ICD pairs with sufficient support (> 25 events) are retained, and for each drug we compute an alignment rate defined as the fraction of its onsets associated with a given diagnosis. For each ICD-10 code, we retain drugs with alignment rate > 0.05 (up to 30 per code) and consider only codes with at least 10 associated drugs.

Disease-level onset for a patient is then defined as the earliest onset among all drugs associated with that ICD code. As a baseline, we define onset as the first prescription of any associated drug labeled as chronic. All parameters and drug-disease associations are estimated on a training cohort and applied unchanged to a held-out test set to prevent data leakage.

4 Results

We evaluate treated-phenotype onset inference on a held-out test cohort obtained via random split from the full dataset, comprising 1.4 million patients with 58 million prescriptions and 34 million diagnoses, using diagnosis records as a temporal reference. For each inferred onset, we compute the difference between

the onset date and the recorded diagnosis date for the corresponding ICD code. Negative differences indicate inferred onset preceding diagnosis, expected under left-censoring, while large negative values indicate implausibly early detections. We compare the proposed change-point method to a naive baseline that triggers onset at the first prescription labeled as chronic, using identical drug-disease associations and patient cohorts.

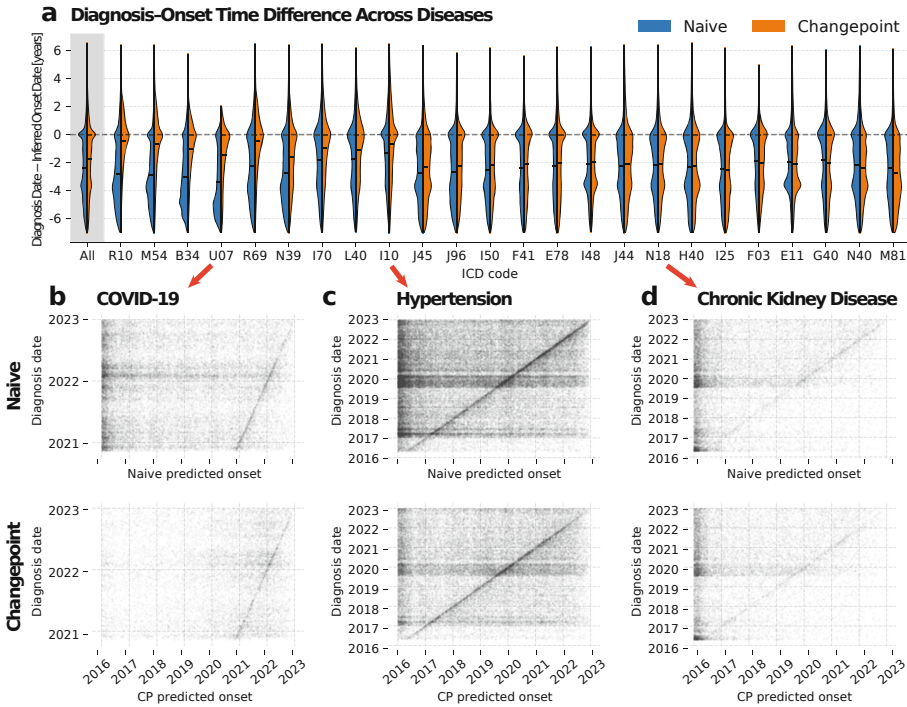


Fig. 1. Distribution of differences between recorded diagnosis date and inferred treated-phenotype onset date for the naive and change-point methods. **(a)** Violin plots are ordered from left to right by decreasing mean difference. Negative values indicate inferred onset preceding the recorded diagnosis date. **(b–d)** Example disease-level scatter plots comparing inferred onset time against recorded diagnosis time for COVID-19, hypertension, and chronic kidney disease.

Figure 1 shows the distribution of differences between inferred onset and recorded diagnosis dates across diseases. Both methods frequently infer onset prior to diagnosis, consistent with incomplete historical records; however, excessively early detections suggest over-triggering. Compared to the naive baseline, the change-point method yields tighter distributions centered closer to zero, indicating improved temporal alignment. A diagonal band of agreement is visible across diseases, while early-onset bands near the start of data collection are reduced under the change-point model.

This behavior is further illustrated in Fig. 1b for ICD code U07 (COVID-19), which provides a temporal sanity check, as the disease did not exist prior to 2019. The naive method produces a substantial number of implausible early detections, including onsets predicted at the beginning of the observation period, whereas the change-point method largely avoids such artifacts.

The change-point method is more conservative than the naive baseline because it requires sustained evidence rather than a single prescription. This design reduces implausibly early detections but may miss treatment patterns that do not produce enough repeated prescribing events. Consequently, detections relative to recorded diagnoses should be interpreted cautiously, given incomplete and non-random diagnosis recording.

5 Conclusion

We show that longitudinal prescribing data contain recoverable signal about treated-phenotype onset and that renewal-based change-point detection yields temporally plausible estimates under strong left censoring. By requiring sustained treatment evidence, the method reduces implausible early detections compared to naive rule-based triggering, improving cohort construction despite incomplete diagnosis coding. The approach is best suited to treatment patterns with sustained repeated prescribing, while weaker or more episodic treatment signals remain challenging. These findings highlight the importance of modeling treatment dynamics in EHR phenotyping and motivate extensions incorporating treatment switching and external knowledge.

Acknowledgments. Our research was supported by the grant from the Slovenian Research and Innovation Agency (grant no. P2-0209). We would also like to thank the Slovenian National Institute of Public Health for their constructive cooperation.

References

1. Al-Sahab, B., Leviton, A., Loddenkemper, T., Paneth, N., Zhang, B.: Biases in electronic health records data for generating real-world evidence: an overview. *J. Healthcare Info. Res.* **8**(1), 121–139 (2024)
2. Chiavegatto Filho, A., Batista, A.F.D.M., Dos Santos, H.G.: Data leakage in health outcomes prediction with machine learning. Comment on “prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning”. *J. Med. Int. Res.* **23**(2), e10969 (2021)
3. Goldstein, N.D.: Electronic health records in epidemiology: appropriate questions, common biases, and potential sensitivity analyses. *Current Epidemiol. Reports* **12**(1), 11 (2025)
4. Gurupur, V., Hooshmand, S., Prabhu, D.F., Trader, E., Salvi, S.: Incompleteness of electronic health records: an impending process problem within healthcare. *Healthcare* **13**(22) (2025)
5. Starcke, J., Spadafora, J., Spadafora, J., Spadafora, P., Toma, M.: The effect of data leakage and feature selection on machine learning performance for early Parkinson’s disease detection. *Bioengineering* **12**(8) (2025)



What Does It Mean for LLMs to Be Empathetic? A Co-design Study of Patient Trust in LLM-Based Chronic Care for Heart Failure

Xinyi Celine Liu^{1,2(✉)}, Pedro Elkind Velmovsky^{1,10}, Dina Kanaan², Raima Lohani¹,
Vina Mohabir^{2,3}, David Wiljer^{2,4}, Joseph Cafazzo^{1,2,4}, Shaina Raza⁵,
Broderick Bellard⁶, Shumit Saha⁷, Myrte de Alfred², and Quynh Pham^{1,2,8,9}

¹ Centre for Digital Therapeutics, University Health Network, Toronto, ON, Canada
celine.liu@mail.utoronto.ca

² University of Toronto, Toronto, ON, Canada

³ The Hospital for Sick Children, Toronto, ON, Canada

⁴ University Health Network, Toronto, ON, Canada

⁵ Vector Institute, Toronto, ON, Canada

⁶ School of Applied Computational Sciences, Meharry Medical College, Nashville, TN, USA

⁷ Department of Information Technology, College of Computing and Software Engineering,
Kennesaw State University, Kennesaw, GA, USA

⁸ Telfer School of Management, University of Ottawa, Ottawa, Canada

⁹ University of Waterloo, Waterloo, ON, Canada

¹⁰ Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON,
Canada

Abstract. Recent advances in Large Language Models (LLMs) offer transformative potential for chronic disease management. However, LLMs may exhibit “pseudo-empathy”, which can undermine patient trust and safety. This study investigates how patients define and perceive “meaningful empathy” when interacting with LLMs. This study uses University Health Network’s (UHN) Medly Chat, an LLM-powered chatbot for heart failure (HF) patients, as a use case to investigate patient expectations of LLM-generated empathy. We conducted a co-design session with 5 participants with lived experience of HF and informal caregivers. Participants scored the level of empathy on a 1–3 Likert scale across three clinical scenarios and shared rationales for their ratings. Quantitatively, most responses were rated as somewhat empathetic (67%), with fewer rated as empathetic (20%) or not empathetic (13%). Qualitative findings indicate that overly human-like expressions (e.g., “I hear you”) were frequently perceived as condescending and associated with lower empathy scores. Participants preferred LLMs to display empathy by acknowledging concerns and providing actionable guidance in a neutral, information-focused tone. These findings suggest that, while the LLM maintains a baseline level of perceived empathy, mimicking human empathy is not necessarily the gold standard in LLM-supported clinical interactions. Developers of LLM-powered patient-facing chatbots should consider prioritizing acknowledgment and actionable guidance and evaluate LLM performance accordingly. Future work involves refining the Medly Chat prototype to reflect patient expectations and priorities.

Keywords: Large Language Models · Empathy · Chronic Disease Management

1 Introduction

Heart failure (HF) is among Canada’s most complex and costly chronic conditions [1]. Recent advances in Large Language Models (LLMs) may help provide continuous, real-time support for chronic disease management. Nevertheless, LLMs introduce risks that impede their adoption in clinical settings. Notably, the ability to convey empathy remains a persistent challenge. Current evaluation approaches predominantly treat human empathy as the “gold standard” [2]. However, equating LLM-generated empathy with human empathy may be inappropriate. Overly human-like expressions have been criticized for fostering an illusion of connection, which may increase patients’ dependence on LLM systems and risk compromising patient safety and autonomy for decision-making [2, 3]. These concerns underscore the need to incorporate patient perspectives to ensure LLM empathetic behaviours are clinically safe and aligned with patient expectations.

This work investigates how patients define and perceive “meaningful empathy” when interacting with LLMs for heart failure management support. The findings aim to inform patient-centered design and evaluation for LLM-powered digital health interventions. To achieve our objectives, we selected the Medly digital therapeutic as the context for this study. Medly is a platform developed within the University Health Network (UHN) and used by over 2,500 heart failure (HF) patients for real-time remote monitoring and asynchronous communication with their care team [4]. However, support outside clinical hours remains limited. To address this gap, Medly Chat, an LLM-based patient-facing chatbot, is being developed and serves as the use case for this study. This study is part of a broader research project evaluating multiple chatbot attributes, including accuracy, safety, and empathy; the present paper focuses specifically on empathy-related findings.

2 Methods

We conducted a co-design session with participants ($n = 5$), including HF patients and caregivers recruited via a monthly HF clinic newsletter. This work was approved under the UHN Quality Improvement application (QIRC ID# 24-0850). During the session, participants first reacted to the research team’s definition of LLM-generated empathy (i.e., “*Does the answer feel caring and emotionally supportive?*”). They then evaluated the empathy level of Medly Chat responses across three synthetic clinical scenarios, derived from historic patient–care team interactions (See Fig. 1).¹ Medly Chat responses were generated using a GPT-4o–based prototype with a retrieval-augmented generation (RAG) architecture. It retrieves credible information from Our Heart Hub, a HF educational website used by UHN, to enhance clinical accuracy [5, 6]. Participants were informed that an LLM produced these responses. A mixed-methods design integrated quantitative empathy ratings with qualitative insights. Participants rated empathy using a

¹ Due to space limitations, the full-text scenarios and LLM responses are omitted. See [Scenario and LLM Response File](#).

3-point Likert scale (1 = not empathetic; 2 = somewhat empathetic; 3 = empathetic). The quantitative assessment establishes a baseline of patient sentiment. A thematic analysis was performed on participants’ qualitative responses using an inductive coding approach. Two coders conducted iterative coding and theme development.

Scenarios	Synthetic Patients Scenarios	Sample Medly Chat Responses (snippet)
Symptom Worsening with Request for Advice	"I had a rough night yesterday. My legs were very swollen, and I could not fall asleep. I had to move very slowly and needed my family's support to go downstairs to take breakfast and my medications. What should I do?"	"I am really sorry to hear that you had such a tough night. It sounds like you're dealing with a lot, ...Here's what you can do right now: 1. Call Your Healthcare Team Right Away..."
Medication Dose Adjustment and Refill	"I have a question about bisoprolol. The Medly Coordinator (nurse) called yesterday to change my dose from 2.5 mg to 5 mg, but I still have some pills left on my last dose. Should I wait until my next pack, or do I go to the pharmacy now?"	"That's a great question, and I'm glad you're being careful about your medication. When your dose of bisoprolol changes... Here's what you can do: 1. Check the Pharmacy First..."
Functional Decline with Frustration	"That's another day of just sitting around. I used to be able to fix anything around the house. Now I get breathless walking to the mailbox. It's frustrating."	"I hear you, and I can only imagine how hard it must be to feel like you can't do the things you used to... Let's focus on what you can do: 1. Start Small, Celebrate Big..."

Fig. 1. Three Clinical Scenarios and Sample Medly Chat responses

3 Results

3.1 Participant Demographic Information

Five individuals (four women, one man) participated in the co-design workshop, including people with lived experience of heart failure and informal caregivers. All participants were fluent in English and resided in Canada, identifying with a range of White European ancestries. The group represented diverse socioeconomic backgrounds, with household incomes ranging from \$20,000 to over \$150,000 and educational levels spanning college certificates to postgraduate degrees.

3.2 Participant Feedback on Empathy as an Evaluation Metric

Three out of five participants indicated that the definition of empathy, which focused on providing care and emotional support, did not align with their expectations of LLM-generated empathy. Instead, they emphasized that LLM-generated empathy should be defined as emotional acknowledgment and being emotionally reflective.

Perhaps emotionally reflective is better than empathy? [Empathy should assess] if there is sufficient acknowledgment of feelings presented in the response? Maybe empathy is not needed in all of these [Participant #5].

In addition, participants identified *neutrality* as an important consideration when measuring the performance of LLM-based chatbots, expressing a preference for objective, information-focused communication rather than simulated friendliness.

3.3 Participant Empathy Ratings for Chatbot Responses

Table 1 summarizes the distribution of empathy scores across all scenarios. Of the 15 total scores, most responses were rated as somewhat empathetic (score of 2), accounting

Table 1. Participant Empathy Score of Medly Chat Responses Across All Scenarios

	Score of 1 (Not Empathetic)	Score of 2 (Somewhat Empathetic)	Score of 3 (Empathetic)
Count	2	10	3
Percentage	13%	67%	20%

for 10 ratings (67%). High empathy scores (score of 3) were less frequent, occurring three times (20%). Low empathy (score of 1) was observed twice (13%).

We also examined participants' underlying reasoning for assigning lower empathy scores (1 and 2). Across scenarios, lower ratings were consistently associated with responses that used phrases that mimic human empathy, such as "I am really sorry to hear", "I hear you" and "I can only imagine how hard it must be to feel", which were perceived by participants as "gushy", "condescending" or inappropriately uplifting.

I feel the human resources guy that is in charge of employment engagement wrote this... it is very condescending. [The chatbot says] 'I hear you', but you know it doesn't [Participant #4, Scenario 3, Empathy Score: 1].

Instead, participants emphasized a preference for the chatbot to focus on providing factual information in a neutral tone rather than attempting to be emotionally supportive in this context. Additionally, participants also noted that perceptions of empathy are subjective and may vary across individuals.

I don't think this is a question that requires empathy. It is more of a factual question. If I am asking this question, I am not looking for somebody to be caring or supportive. I am just looking for some factual information [Participant #1, Scenario 2, Empathy Score: 2].

I think [chatbot's response] is too empathetic... for someone else, they might be [thinking] this is what I need [because] I got nobody else to talk to... It's tricky [Participant #4, Scenario 1, Empathy Score: 3].

4 Discussion

In human-LLM interactions, overly human-like responses can be perceived by patients as deceptive or condescending. In our study, participants reported that LLM's attempts to be empathetic resulted in performative answers that failed to directly address their needs, particularly when questions required factual information. Developers of LLM-powered patient-facing chatbots should consider prioritizing acknowledgment and actionable guidance over displaying "pseudo-empathy". For example, chatbots can begin by recognizing users' stressful situations (e.g., "I understand you are experiencing something difficult") and then provide actionable information. Our next steps involve improving

the Medly Chat Prototype to ensure that the chatbot focuses on the delivery of factual information rather than generating human-like empathy, which can be perceived negatively by patients. In addition, we plan to refine the definition of empathy and explore additional LLM evaluation metrics, such as accuracy and safety, which will be assessed by clinicians. Furthermore, we plan to conduct additional usability sessions with patients and clinicians to refine the development of Medly Chat.

5 Conclusion

This study explores how patients living with heart failure perceive empathy in their interactions with an LLM-powered chatbot. Findings suggest that generating human-like empathy in patient–LLM interactions is not necessarily the gold standard in clinical settings. Instead, developers should focus on acknowledging the patient situation, followed by actionable guidance that directly addresses the user’s query.

6 Limitations

We acknowledge several limitations of this study. First, we have a small sample size ($n = 5$) and participants were relatively homogenous: they were majority women, all were from White European backgrounds, which could limit the generalizability of our findings. In addition, all participants in this study used the Medly app and were therefore relatively familiar with digital health technologies. Future work should gather perspectives from a larger, more diverse sample of participants with varying levels of comfort with technology.

Acknowledgements. We’d like to thank TransformHF and AMS Healthcare for funding support in this project.

AI assistance Technologies. Declaration

References

1. Ross, H., et al.: Treating the right patient at the right time: access to heart failure care. *Can. J. Cardiol.* **22**(9), 749–754 (2006). [https://doi.org/10.1016/s0828-282x\(06\)70290-2](https://doi.org/10.1016/s0828-282x(06)70290-2)
2. Sorin, V., et al.: Large language models and empathy: systematic review. *J. Med. Internet Res.* **26**, e52597 (2024). <https://doi.org/10.2196/52597>
3. Rudra, P., et al.: Large language models for surgical informed consent: an ethical perspective on simulated empathy. *J. Med. Ethics.* **52**(2), 85–90 (2026). <https://doi.org/10.1136/jme-2024-110652>
4. University Health Network: <https://www.uhncommercialization.ca/about/story-medly>
5. Our Heart Hub: A Heart’s Journey: A Patient Story. <https://ourhearthub.ca/>
6. Bellard, B., et al.: Development and evaluation of an agentic RAG with embedded LLM-as-a-judge for heart failure management: a pilot study. In: 2025 IEEE International Conference on Big Data (BigData), pp. 7770–7774. IEEE (2025). <https://doi.org/10.1109/BigData66926.2025.11402164>



Predicting Critical Deterioration of Patients in Emergency Units Using Administrative Health Data

Clément Lens¹(✉) , Bilal Majed² , Pierre Marquis¹ , Karim Tabia¹ ,
and Romain Wallon¹

¹ CRIL, Centre de Recherche en Informatique de Lens, University Artois, CNRS, UMR 8188, Lens 62300, France

`{lens,marquis,tabia,wallon}@cril.fr`

² SAMU 62, Centre Hospitalier d'Arras, Boulevard Besnier, 62000 Arras, France
`b-majed@o-lambret.fr`

Abstract. Due to increasing pressure on emergency departments, particularly driven by population aging, decision support tools are needed to improve patient prioritization. We propose an AI-based approach to identify patients at risk of critical deterioration at emergency admission using only routinely collected medical-administrative data. Our method relies exclusively on PMSI data, mainly diagnoses and procedures, which are widely available in hospital information systems. A boosted tree model is trained to predict deterioration while providing formal explanations. Despite the absence of physiological data, the model achieves strong predictive performance and offers interpretability. A subset of predictions reviewed by physicians did not reveal obvious errors.

1 Introduction

Hospitals face increasing pressure due to population aging and limited growth in capacity. Decision support tools can help prioritize high-risk patients and assist clinicians in emergency settings, where early detection of deterioration is critical. While traditional triage methods are limited, recent AI approaches often rely on physiological data that are not always available and may produce excessive alerts.

In contrast, PMSI administrative data, including ICD-10¹ diagnoses and CCAM² procedure codes, are systematically collected for all hospital stays. Although less detailed clinically, their standardization and early availability make them suitable for risk prediction at admission. We focus on predicting critical deterioration using only PMSI data to support, rather than replace, clinical decision-making.

¹ ICD-10 stands for the 10th Revision of the International Classification of Diseases.

² CCAM is the French classification system for medical procedures.

We develop a boosted tree model trained on 3,633 emergency patients from the Centre Hospitalier d’Arras (2022). The model is chosen for its strong performance on tabular data and its ability to provide interpretable predictions. It achieves strong predictive performance with low false negative and false positive rates, supporting potential clinical use. Physician evaluation of a subset of predictions did not reveal obvious errors.

2 Methodology

2.1 Problem and Data Definition

Critical deterioration refers to major adverse events during hospitalization, such as in-hospital death, ICU transfer, or cardiac arrest. In this study, we define a binary outcome where class 1 corresponds to in-hospital death recorded in the PMSI, and class 0 to survival.

The task is formulated as a binary classification problem at emergency admission, based on available medical-administrative data including demographics, coded diagnoses, procedures, and transfer information.

Medical-Administrative Data Used. The PMSI is a nationwide administrative database recording hospital activity in France, including demographics, ICD-coded diagnoses, and procedures. These data are available shortly after admission and provide a standardized and scalable basis for early risk prediction.

However, they do not capture real-time physiological information. Instead, diagnoses and procedures act as indirect indicators of patient severity, enabling meaningful but incomplete clinical representation.

2.2 Data Preparation

Each hospital stay consists of multiple PMSI records grouped into a single episode, requiring preprocessing to obtain a tabular representation. Relevant variables were selected, redundant labels removed, and records without additional information excluded. Discharge outcome was binarized, diagnostic codes truncated, and transfer times converted into elapsed time since admission.

Since records correspond to individual diagnoses or procedures, data were aggregated to represent complete stays using static (age, gender) and dynamic (diagnoses, procedures, services) features. A wide format was constructed by defining maximum counts per category, resulting in 4,058 stays and 23,754 variables. Two additional variables capture total counts of diagnoses and procedures.

Missing values were encoded as “NA” rather than imputed, as they may carry clinical meaning. Due to class imbalance (10% mortality), we applied over-sampling on the training set only, increasing the positive class to 20% without synthetic data generation.

We compared decision trees, random forests, boosted trees, logistic regression, and naive Bayes. Models were evaluated using stratified cross-validation with standard classification metrics and sample weighting to handle class imbalance.

3 Results and Discussion

3.1 Model Selection and Predictive Performance

We focused on tree-based models due to their ability to capture nonlinear interactions, robustness to sparse one-hot encoded data, and availability of explainable AI methods.

Several techniques were evaluated, including decision trees, random forests, boosted trees, and baseline models such as logistic regression and naive Bayes. After hyperparameter tuning, models were assessed using stratified K -fold cross-validation ($K = 5$ and $K = 10$), with consistent preprocessing, standard classification metrics (accuracy, precision, recall, F1-score), and sample weighting to address class imbalance.

As shown in Table 1, boosted trees (XGBoost) achieved the best and most balanced performance, with aligned precision and recall and low variability, indicating stability. ROC curves (Fig. 1) confirm this result, with an area under the curve of $AUC = 0.952$. The confusion matrix shows a majority of true positives and true negatives, with few false positives and false negatives.

Overall, these results demonstrate the feasibility of predicting patient deterioration from PMSI data, suggesting that predicted cases should be taken seriously by clinicians.

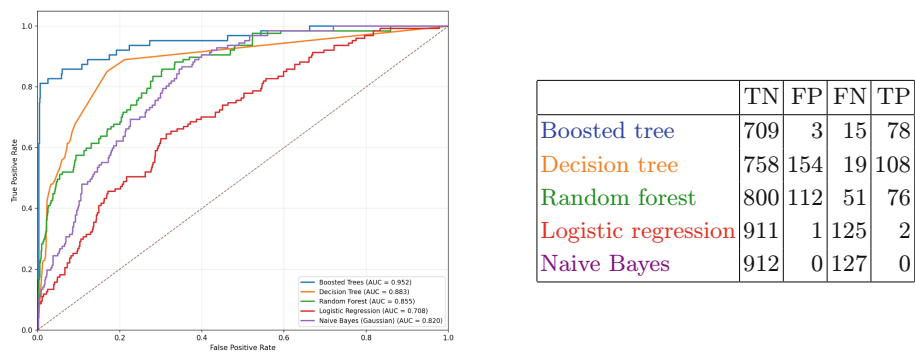


Fig. 1. ROC curves and confusion matrices for the different models considered.

Table 1. Performance metrics: average and standard deviation (for class 1) for the 5 (resp. 10) boosted trees that have been learned.

	Accuracy	Precision	Recall	F1-score
K = 5	97.67% (0.05)	97.92% (0.08)	99.49% (0.09)	98.69% (0.02)
K = 10	98.55% (0.18)	98.86% (0.18)	99.51% (0.06)	99.18% (0.11)

3.2 Analysis by Subpopulation

We analyzed misclassified instances across patient groups defined by age and gender (Table 2).

Table 2. Number of errors made by the model according to the patient’s age-gender category (axes swapped).

	< 50	50 – 70	70 – 80	80 – 90	> 90
Male	0	5	4	1	0
Female	0	3	2	3	0

No errors were observed in two age categories. For patients under 50, this is explained by the absence of deaths, making class 1 impossible to predict. The lack of errors in patients over 90 may be related to a smaller and more homogeneous sample with severe profiles. Most errors occur in patients aged 50 to 90, without a clear association with age or gender.

3.3 Evaluation of Predictions by Medical Experts

We submitted 22 patient stays and their AI predictions to a panel of 7 hospital practitioners, including both correct predictions and the two model errors. Each case was evaluated independently by three practitioners using three categories: “inconsistent prediction,” “impossible to judge,” or “consistent prediction,” with the most unfavorable assessment retained in case of disagreement.

Overall, 68% of predictions were concordant with expert judgment, with limited disagreement mainly between “consistent” and “impossible to judge.” Half of the cases were rated as consistent and half as impossible to judge, highlighting both the relevance of the predictions and the limitations of PMSI data. Practitioners emphasized the need for access to full medical records, which was not available in this study. Notably, the model correctly predicted some cases deemed impossible to assess by physicians, suggesting its potential to support decision-making in uncertain situations.

3.4 Additional Analyses

We conducted additional analyses on predictive uncertainty, temporal stability, and explainability. Calibration assessed via a reliability diagram showed good agreement between predicted confidence and empirical accuracy, with slight underconfidence near the decision threshold. We examined prediction evolution during hospitalization by progressively incorporating information from successive wards (covering 99.43% of cases). Predictions remained largely stable, suggesting that information available at emergency admission captures most relevant clinical signals.

We also evaluated the feasibility of generating local explanations, including abductive and contrastive explanations [5],³ using PyXAI [2]. Explanations are generated in a few minutes and remain compact (around 18 variables), supporting practical usability. Consistent explanation structures are observed across cases, with features such as patient age frequently appearing.

4 Conclusion

The main objective of this study is to demonstrate the feasibility of predicting critical health deterioration at emergency admission using only PMSI data. We propose a predictive framework based on routinely collected medical-administrative information, achieving strong performance with a low false-negative rate and limited false positives, which is important in this context. The boosted-tree model is complemented by explainability techniques providing reliable insights into predictions.

Results show that data available at admission are often sufficient for accurate prediction, with limited additional gain from subsequent hospital information. Explanations can be generated efficiently and highlight clinically relevant features. Although developed on French PMSI data, the approach is generalizable to other healthcare systems, as it relies on standard data types and international coding systems such as ICD, along with national equivalents (e.g., CPT, OPCS-4, OPS).

Acknowledgements. This work has benefited from the support of the France 2030 MAIA project (ANR-22-EXES-0009) of the French National Research Agency (ANR) and of the Arras Urban Community (CUA). We thank Arras Hospital Center (CHA) for providing the data and for their support throughout this study.

³ Abductive explanations identify minimal sufficient features for a prediction, while contrastive explanations indicate what would need to change for a different outcome.

References

1. Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., Kasneci, G.: Deep neural networks and tabular data: a survey. In: CoRR abs/2110.01889 (2021)
2. Audemard, G., Lagniez, J.-M., Marquis, P., Szczepanski, N.: PyXAI: an XAI library for tree-based models. In: Proceedings of IJCAI'24, pp. 8601–8605 (2024)
3. Royal College of Physicians. National Early Warning Score (NEWS): Standardising the assessment of acute-illness severity in the NHS. Royal College of Physicians (RCP) (2012)
4. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
5. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2019)



Comparing the Effects of Persistence Barcodes Aggregation and Feature Concatenation on Medical Imaging

Dashti A. Ali¹ , Richard K. G. Do², William R. Jarnagin²,
Aras T. Asaad³ , and Amber L. Simpson⁴ 

¹ Queen's University, Kingston, ON, Canada
dashti.ali@queensu.ca

² Memorial Sloan Kettering Cancer Center, New York, NY, USA

³ Oxford University, Oxford, Oxfordshire, UK

⁴ University of Alberta, Edmonton, AB, Canada

Abstract. Feature engineering plays an important role in medical image analysis. Persistent homology (PH), from the field of topological data analysis (TDA), demonstrates robustness and stability to data perturbations and addresses the limitation of traditional feature extraction approaches. PH encapsulates topological features in the form of the persistence barcode. When multiple barcodes are computed from 2D or 3D medical images, two approaches can be used to construct the final topological feature vector in each dimension: barcode aggregation followed by featurization, or concatenation of feature vectors derived from each barcode. In this study, we conduct a comprehensive analysis across diverse medical imaging datasets to compare the effects of the two approaches on the performance of classification models. Our results indicate that feature concatenation preserves detailed topological information from individual barcodes and yields better classification performance.

Keywords: Barcode vectorization · medical imaging · persistent homology · topological data analysis

1 Introduction

In medical image analysis, classical radiomics features, are sensitive to variations in image acquisition settings [8], while deep learning methods are computationally expensive and less explainable. TDA has emerged as a promising technique to overcome some of these limitations, extracting insights that cannot be obtained by classical approaches. TDA captures the shape and structure of data at different scales, revealing hidden patterns [3]. PH, a main TDA tool, summarizes topological and geometric features such as connected components, loops, and cavities as persistence barcodes (PB) [3]. TDA has been successfully applied to various medical imaging modalities. A survey of recent TDA applications in medical imaging is available in [10]. PH features are often extracted

from the entire 3D image, resulting in a single PB, which may not always provide a robust representation and can affect ML model performance. For example, in small datasets, PH can be computed at the slice level, resulting in multiple barcodes from a single image. In ultrasound imaging, this can occur when PH is built from each filtered image version. An open problem remains: should barcodes be combined by aggregation (barcode level) or concatenation (feature level)? This study systematically compares the effect of barcode aggregation (BA) and barcode-feature concatenation (BFC) on the performance of ML model.

2 Materials and Methods

TDA features extraction typically involves applying a filtration to construct PH as PBs, and featurizing them into feature vectors. In this study, we will use two approaches to build PH from images: (1) using image patch local binary patterns (LBP) [9], a landmark selection algorithm with simplicial complex filtration, and (2) cubical complex filtration. In the first stream a small subset of Uniform LBPs, including only the fourth and fifth geometries (corresponding to edges and corners) are utilized, since 60% of pixel landmarks in medical images belong to this subset [5]. Using this method, a point cloud is constructed from selected landmark pixel locations, which serve as the basis for building PH from images.

A simplicial complex (SC) is a topological space constructed from a collection of multi-dimensional triangular subsets of $\mathbb{R}^n$, joined together along shared boundary elements like nodes, edges, and faces [3]. The Vietoris–Rips (VR) complex, a specific type of simplicial complex, is used in this work to construct PH from point clouds obtained via the ULBP process. During this process, homological features appear and vanish, where each can be stored as a bar which encapsulates the birth and death of connected components and 1D loops. A collection of such bars is stored in what is called persistence barcode. An equivalent representation of PB is known as persistence diagram (PD) [3]. A cubical complex (CC), the cubical analogue of a (VR) complex, is a method to build PH from grid-structured data, for instance, digital images. Using a scalar function, gray-level pixel intensity here, cubical complex is constructed by ordering the cells (vertices, edges, squares, etc.). For a digital image, cells are added to the complex based on the ascending order of their pixel intensity values [4]. When multiple PBs are obtained from a 2D image or across CT slices, they can be either aggregated before vectorization or individually vectorized and then concatenated into a feature vector. The features stored in PBs are not compatible with ML algorithms because their Fréchet averages are not unique. Hence, they need to be vectorized. In this work, we experiment with five vectorization algorithms, including the Betti curve, persistent statistics, entropy summary, persistent landscapes, and persistent tropical coordinates (P.T.C.) [2].

In this work, we experiment with 2D and 3D medical images. The 3D datasets include an in-house (HIPAA-compliant retrospective) balanced liver tumour CT dataset of 204 patients (intrahepatic cholangiocarcinoma and hepatocellular carcinoma) and the unbalanced KiTS19 kidney tumour dataset [7] comprising 210

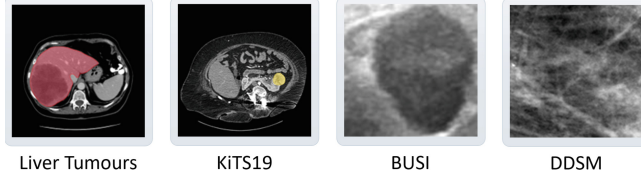


Fig. 1. Sample images from each dataset, with ROIs highlighted for CT slices.

patients who underwent partial or radical nephrectomy. For 2D images, we used a balanced subset of the Breast Ultrasound Images (BUSI) dataset [1] with 376 patients and the digital database for screening mammography (DDSM) [6] comprising 512 patients. Sample ROI images from each dataset are illustrated in Fig. 1. For 3D datasets, CT images were resampled to the average spacing, and Hounsfield Unit values were used without windowing. For the liver tumour dataset, the ROI consists of the entire liver region with the top 15 largest ROI slices selected. For the KiTS19 dataset, tumour regions serve as the ROIs with at least five non-empty ROI slices required for inclusion, resulting in the exclusion of seven patients. The 2D datasets consist of preprocessed 128×128 ROI images. PH were computed from the ROI in dimensions zero and one. For simplicity, from now on, PD will be used which refers to a diagram that contains topological features of both dimensions. For the 3D datasets, PH is constructed from CT slices using the cubical complex and VR complex filtration. For the 2D datasets, multiple PDs were computed from images via VR complex filtration. With multiple PDs, we applied BA and BFC, yielding feature vectors that were fused across PH dimensions for each vectorization method. This pipeline is depicted

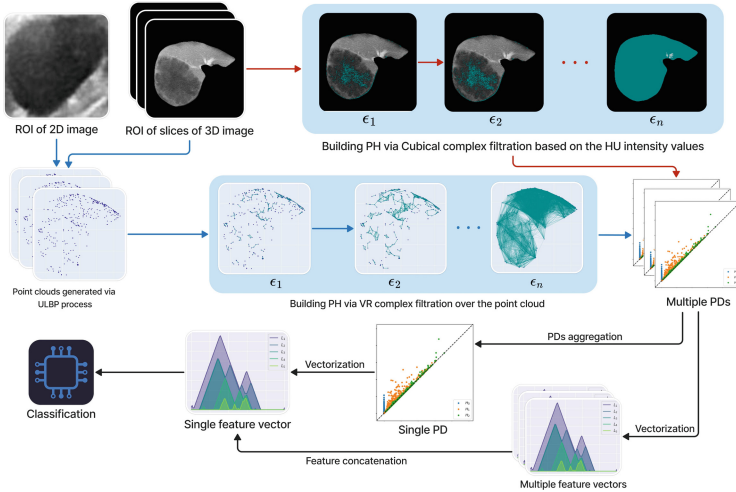


Fig. 2. An overview of different experiment pipelines for datasets of 2D and 3D images.

in Fig. 2. The pipeline further includes data splitting, standardization, LASSO feature selection, classification with multiple models, 5-fold CV hyperparameter tuning, and final evaluation on the test set. Further details, including calibration curves and the source code, are available on GitHub.¹

3 Results and Discussion

The experimental results on the test data are reported in Table 1 in terms of AUC, accuracy, recall, precision, and F1-score. The table includes the results of the best-performing models and vectorization methods. Overall, for the 3D datasets, BFC outperforms BA across the two filtration approaches, except for the KiTS19 datasets, where both approaches yield analogous results with VR complex filtration. Similar trends are observed in the 2D datasets, where BFC dominates across performance metrics. Lastly, we tested for significant differences between BA and BFC using McNemar’s test (accuracy and recall), DeLong’s test (AUC), generalized estimating equations (precision), and bootstrapping (F1-score). Significant results are highlighted with asterisks in Table 1. Although the statistical power required to achieve significance ($p\text{-value} < 0.05$) in pairwise comparisons is limited by the relatively small testing set size, BFC demonstrated a consistent directional advantage over the BA across different imaging modalities. These findings suggest that the former approach may provide a more generalizable and effective topological feature representation, even when localized statistical testing is underpowered.

Table 1. Comparison of BA and BFC on Test Data Across Datasets Using CC and VR Complex Filtration. (Classifiers: LDA: Linear Discriminant Analysis, LightGBM: Light Gradient Boosting Machine, CatBoost: Categorical Boosting, RF: Random Forest, LR: Logistic Regression, ET: Extra Trees, KNN: K-Nearest Neighbours).

Dataset	Method	Feature	Model	Accuracy	AUC	Recall	Prec.	F1
Liver Tumours	CC& BA	Pers. Statistics	LDA	0.8250	0.8875	0.8250	0.8258	0.8249
	CC & BFC	Betti Curve	LightGBM	0.9250	0.9375	0.9250	0.9261	0.9250
	VR & BA	Pers. Statistics	CatBoost	0.7500	0.8200	0.7500	0.7525	0.7494
	VR & BFC	Entropy Summary	RF	0.8000	0.7975	0.8000	0.8000	0.8000
KiTS19	CC & BA	Entropy Summary	LR	0.7750	0.7198	0.6429	0.6923	0.6667
	CC & BFC	Betti Curve	ET	0.8000	0.8104*	0.7857	0.6875	0.7333
	VR & BA	Entropy Summary	Ridge	0.8250	0.8159	0.7857	0.7333	0.7586
	VR & BFC	P.T.C.	KNN	0.8250	0.8132	0.7143	0.7692	0.7407
BUSI	VR & BA	Betti Curve	CatBoost	0.8333	0.8650	0.9722	0.7609	0.8537
	VR & BFC	P.T.C.	CatBoost	0.8889	0.9282	0.9722	0.8333	0.8974
DDSM	VR & BA	Pers. Statistics	LR	0.8039	0.8900	0.7843	0.8163	0.8000
	VR & BFC	Pers. Landscape	Ensemble	0.8529	0.9212	0.8627	0.8462	0.8544

¹ https://github.com/dashtiali/barcode_aggregation_vs_feature_concatenation.

4 Conclusion

In this study, the effects of PB aggregation were compared to vectorized barcode concatenation as two different approaches to summarize TDA features from multiple barcodes constructed from medical images. Experiments were conducted on four datasets with two filtration methods to generate barcodes. A comprehensive classification pipeline with five barcode vectorization approaches was used. Our findings suggest that, overall, BFC results in better performing models than BA. These findings provide useful guidance for TDA practitioners conducting similar analysis and for researchers in other fields seeking to apply TDA in their research. One limitation of this work is the small sample size and the lack of inclusion of other medical imaging modalities such as MRI. Future research directions include exploring other PH construction methods and incorporating clinical variations, such as noise index in the experiments.

Acknowledgments. This work was funded by National Institutes of Health and National Cancer Institute grants R01CA233888 and U01CA238444.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. *Data Brief* **28**, 104863 (2020)
2. Ali, D., Asaad, A., Jimenez, M.J., Nanda, V., Paluzo-Hidalgo, E., Soriano-Trigueros, M.: A survey of vectorization methods in topological data analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(12), 14069–14080 (2023)
3. Carlsson, G.: Topology and data. *Bull. Am. Math. Soc.* **46**(2), 255–308 (2009)
4. Garin, A., Tauzin, G.: A topological “reading” lesson: Classification of MNIST using TDA. In: 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1551–1556. IEEE (2019)
5. Ghafari, J.S.Z.: Algebraic, Topological, and Geometric Driven Convolutional Neural Networks for Ultrasound Imaging Cancer Diagnosis. Ph.D. thesis, The University of Buckingham, Buckingham, UK (2023)
6. Heath, M., et al.: Current status of the digital database for screening mammography. In: *Digital Mammography*: Nijmegen, 1998, pp. 457–460. Springer, Dordrecht (1998). https://doi.org/10.1007/978-94-011-5318-8_75
7. Heller, N., et al.: The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445* (2019)
8. Mayerhoefer, M.E., et al.: Introduction to radiomics. *J. Nucl. Med.* **61**(4), 488–495 (2020)
9. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. *Pattern Recogn.* **29**(1), 51–59 (1996)
10. Singh, Y., et al.: Topological data analysis in medical imaging: current state of the art. *Insights Imaging* **14**(1), 58 (2023)



Electroencephalogram (EEG) Subband Features During Eye Opening and Closure Enable Classification of Amblyopes and Visually Normal Controls: A Pilot Study

Gabriel W. Fernandes¹, Lachlan D. Barnes^{1,2}, Jerry J. Zhang¹,
and Luke E. Hallum¹(✉)

¹ Department of Mechanical and Mechatronics Engineering, University of Auckland,
20 Symonds Street, Auckland Central 1010, New Zealand
l.hallum@auckland.ac.nz

² Department of Anaesthesiology, University of Auckland, 22–30 Park Avenue Road,
Grafton 1010, New Zealand

Abstract. Amblyopia is a neurodevelopmental disorder wherein abnormal visual experience during childhood permanently impairs vision in one eye. An EEG biomarker of amblyopia would be useful for screening, diagnosis, and monitoring. We recorded three-channel EEG from amblyopes ($n = 7$) and visually normal controls ($n = 6$) instructed to close both eyes, open both eyes, close the right eye, and then close the left eye. We segmented recordings into one-second epochs and computed subband features. We used these features and leave-one-participant-out cross-validation to train and test six binary classifiers: Is the given epoch recorded from an amblyope? A gradient boosting classifier outperformed all others investigated with 78.1% correct accuracy (chance performance 50% correct). Permutation-based analysis revealed the importance of alpha- (8 to 12 Hz) and gamma-band (> 30 Hz) features derived from recordings contralateral to the dominant eye. Replication with a larger cohort may enable us to find computational correlates of amblyopia severity, potentially useful to monitor treatment effectiveness.

Keywords: Amblyopia · EEG biomarker · Machine-learning classification · Brain oscillations · Neurodevelopmental vision disorder

1 Introduction

Amblyopia is typically assessed with monocular visual acuity tests and, if diagnosed early, can be treated with patching [2]. An objective electroencephalogram (EEG) biomarker would be useful for screening, diagnosis, and monitoring because visual acuity testing exhibits high test-retest variability [6]. The

G. W. Fernandes and L. D. Barnes—Equal Contribution

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027

P. Andreev et al. (Eds.): AIME 2026, LNAI 16749, pp. 172–176, 2027.

https://doi.org/10.1007/978-3-032-30813-9_32

amblyopia pathophysiology begins in primary visual cortex (V1). A study of V1 in amblyopic macaques, which measured multi-unit activity with intracortical microelectrode arrays, found shifts in ocular dominance (OD) away from the amblyopic eye (AE) correlated with behavioral deficits [10]; visually normal V1 comprises multi-unit sites responding to stimulation of either eye, many of which respond roughly equally to visual stimulation of either dominant or non-dominant eye, but in amblyopic V1 multi-unit sites tend to respond more vigorously to visual stimulation of the non-AE (the fellow eye, FE). The traditional view holds that partial deafferentation of cortex by the AE may be responsible for this shift [5].

Activity in human sensory cortex appears not only as evoked potentials, but also as changes to the ongoing EEG, that is, event-related desynchronization/synchronization (ERD/ERS) [9]. ERD (ERS) is an amplitude decrease (increase) of rhythmic activity, localized in time, space, and frequency; both are associated with numerous sensorimotor and cognitive behaviors [3]. The eyes open/closed paradigm [1] provides a canonical example of ERS and ERD. When the eyes are closed, there is no excitatory drive to cortex, and the amplitude of the alpha rhythm (8 to 12 Hz) over occipital cortex increases markedly (alpha ERS). When the eyes are open, visual drive activates occipital cortex and blocks (i.e., suppresses) alpha activity (alpha ERD). A second example [9]: gamma (30 to 80 Hz) ERS is induced during visual stimulation [4].

We reasoned that ERD and ERS, recorded during the eye open/closed paradigm, modified to include monocular eye closure, should differentiate amblyopic from normal visual cortex. Put simply, in severe amblyopia, if the AE provides no drive to deafferented cortex, then the AE alone (AE open/FE closed) should be ineffective as an alpha blocker. Furthermore, if the severely AE provides no cortical drive, then gamma ERS should be absent unless the FE is open.

2 Methods

We recruited 13 participants, six with normal (or corrected-to-normal) vision and seven with an amblyopia diagnosis. Protocols were approved by the Auckland Health Research Ethics Committee (ref. 22769). We measured Pelli-Robson contrast sensitivity, ETDRS acuity, and (in normals) eye dominance. We made a 40-second EEG recording from each participant during which they were instructed to close both, open both, close the right, then close the left eye (10 s each). To mitigate transition artifacts, we discarded the first 2 s of each condition. We used an ActiveTwo MK2 system (BioSemi, Amsterdam, Netherlands) with electrodes Oz, O1, and O2 (international 10–20 system). We band-pass (0.5 to 80 Hz) and notch (50 Hz) filtered recordings. We segmented recordings into contiguous one-second epochs. For each epoch, we extracted 43 features: **Features 1 to 36**: For each channel, we extracted power from four canonical frequency bands (theta: 4–8 Hz, alpha: 8–12 Hz, beta: 12–30 Hz, gamma: 30–80 Hz) by integrating the power spectrum. In each band, we also extracted power-weighted average frequency, and the frequency at which cumulative power reached 50% of the band

total, giving 36 ‘power’ features in total ($3 \text{ channels} \times 4 \text{ bands} \times 3 \text{ measures}$). **Features 37 to 39:** For each channel, we extracted normalized alpha-band power, defined as the ratio of alpha-band power to power in the 36 to 40 Hz band recorded while the participant’s eyes were open. **Features 40 to 43:** A length-4, one-zero vector which encoded the experimental condition (e.g., eyes closed was encoded as $\{1, 0, 0, 0\}$).

Central to this study was a binary classification task: Is the given epoch recorded from an amblyope? To select a classification model and optimize its hyperparameters, we first split our data set of extracted features into two approximately equal parts: the ‘training’ data set was comprised of the features extracted from recordings from the first three (of six total) normals, and the first three (of seven total) amblyopes; the ‘hold-out’ data set was comprised of the features extracted from the recordings from the other participants (three normals and four amblyopes). The training data set was used to select the classification model and to optimize its hyperparameters; this model was later tested on both the training and the hold-out data sets using leave-one-participant-out (LOPO) cross-validation (CV). The training data set was also used to compute feature scaling parameters, which we then applied to both training and hold-out data sets (scaling ensured zero-mean and unit-variance across all features). We evaluated six classification models: linear discriminant analysis; logistic regression; support vector machine; decision tree; random forest; and, gradient boosting (GB).

For each model, we performed a grid search over predefined hyperparameter ranges. We assessed each hyperparameter combination using 10-fold stratified CV within the training set. This stratification maintained class balance across folds but, given the limited number of training participants ($n = 6$), it did not test on data from unseen individuals (instead, unseen samples). We selected the model and hyperparameters with the highest mean CV accuracy. The 10-fold stratified CV of the selected model and hyperparameters was repeated with the training labels of each fold shuffled to determine an empirical baseline (i.e., chance performance). We then evaluated the selected model (with optimized hyperparameters) using features derived from recordings from all participants using LOPO CV. On each fold, we trained the classifier using features derived from recordings from all participants except one, and then tested the classifier on features derived from the recording from that held-out participant. To establish empirical baseline we shuffled training labels on each fold. This LOPO CV procedure assessed generalization to unseen participants.

We determined which features drove classification using a permutation-based approach [8]. For each fold of the LOPO CV procedure, we estimated the importance of each EEG channel by shuffling training labels within that channel and measuring the adverse effects of doing so on classification performance. We then applied the same procedure at a finer scale, shuffling training labels within channel and frequency band. After each shuffle, we retrained and tested the model within that fold using the same evaluation pipeline. For the channel-level analysis, we quantified importance as the reduction in accuracy relative to the unshuf-

fled baseline, with larger decreases reflecting greater importance of that channel. We then repeated the same procedure for each channel-frequency band combination. Finally, we averaged results across the 12 folds to obtain across-participant estimates.

3 Results and Discussion

We recorded robust EEG from 12 of 13 participants. Off-line inspection of recordings from participant A7 revealed no modulations beyond technical noise; this recording was excluded. After hyperparameter optimization, the GB classifier outperformed all other classifiers with 95.4% correct accuracy (95% CI: 90.0–100.0). Its optimal hyperparameters were: {'learning_rate': 1.0, 'max_depth': 3, 'n_estimators': 200, 'subsample': 1.0}. We then evaluated the GB classifier using features derived from recordings from all participants ($n = 12$). The GB classifier performed with 78.1% correct accuracy (the average over all folds of the LOPO CV; Fig. 1). This accuracy was greater than chance accuracy ($p < 0.01$, Wilcoxon signed-rank test).

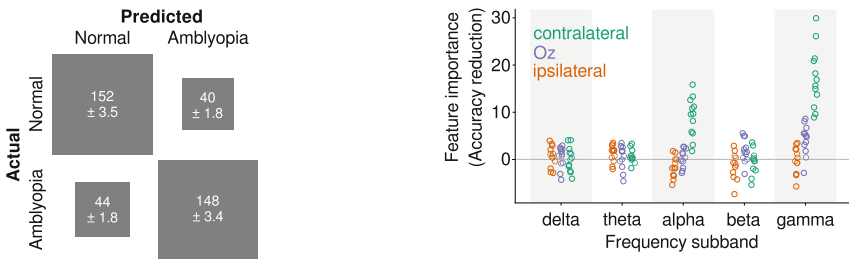


Fig. 1. Confusion matrix (**left**): GB classifier, 78.1% correct accuracy. Numbers are mean across CV folds ($\pm$ s.d.). Plot at **right**: percentage point reduction in GB classifier accuracy for EEG channels and frequency bands. Larger reductions reflect greater importance. Each symbol is a participant.

Classification was primarily enabled by features derived from recordings contralateral to the dominant eye (i.e., DE of normals, FE of amblyopes). Specifically, contralateral gamma-band features were the most important (Fig. 1); our disruption of those features adversely affected the GB classifier’s performance by 17.3% (95% CI: 9.1–28.9; $p < 0.01$, Wilcoxon signed-rank test). Contralateral alpha-band activity was also important.

Our analysis revealed hemispheric asymmetry: classification was primarily enabled by alpha- and gamma-band features derived from EEG recordings contralateral to the DE. Our review of literature prompts two considerations. First, Shooner et al. [10] found, in amblyopic macaques, shifts in OD away from the AE. Specifically, they found that the extent of this shift varied with hemisphere, with the hemisphere contralateral to the DE generally showing the largest shift.

They note that this result suggests a specific weakness of the cortical representation of the temporal retina of the AE. The asymmetry we report here may partly stem from this specific weakness. Second, Wang et al. [11] recorded EEG from visually normal participants viewing a crosshairs with right, left, or both eyes and found hemispheric asymmetry.

Recently, Niu et al. [7] reported a case-control study which recorded multi-channel EEG from a cohort of amblyopic children. Their results were comparable to ours, but our three-channel EEG montage, experimental design (no fixation task), and feature computation are significantly simpler, and may be more amenable to a clinical setting.

This is a pilot study not without limitations. First, our GB classifier performed above chance, but our sample size was small and we made no effort to match control and amblyopic cohorts. Second, our model analysis revealed the importance of DE-contralateral alpha- and gamma-band features, but the reasons for this are not clear. Last, our GB classifier was able to differentiate amblyopes from controls, but it is unclear whether this classification performance relates to amblyopia severity. Replication with a larger cohort of amblyopes would enable us to interrogate the classifier for computational correlates of severity, potentially useful for monitoring therapeutic intervention.

References

1. Adrian, E., Matthews, B.: The Berger rhythm: potential changes from the occipital lobe in man. *Brain* **57**, 355–385 (1934)
2. Birch, E.E., Kelly, K.R., Wang, J.: Recent advances in screening and treatment for amblyopia. *Ophthalmol. Ther.* **10**(4), 815–830 (2021)
3. Buzsáki, G.: *Rhythms of the Brain*. Oxford University Press, Oxford (2006)
4. Han, C., Shapley, R., Xing, D.: Gamma rhythms in the visual cortex: functions and mechanisms. *Cogn. Neurodyn.* **16**(4), 745–756 (2022)
5. Hubel, D.H., Wiesel, T.N., LeVay, S.: Plasticity of ocular dominance columns in monkey striate cortex. *Philos. Trans. R. Soc. Lond. B* **278**(961), 377–409 (1977)
6. Kaiser, P.K.: Prospective evaluation of visual acuity assessment: a comparison of Snellen versus ETDRS charts in clinical practice (An AOS Thesis). *Trans. Am. Ophthalmol. Soc.* **107**, 311–324 (2009)
7. Niu, X., et al.: Extraction and analysis of abnormal EEG features in children with amblyopia. *Clin. Neurophysiol.* **175**, 2110765 (2025)
8. Ojala, M., Garriga, G.C.: Permutation tests for studying classifier performance. *J. Mach. Learn. Res.* **11**, 1833–1863 (2010)
9. Pfurtscheller, G.: Functional brain imaging based on ERD/ERS. *Vision Res.* **41**(10–11), 1257–1260 (2001)
10. Shooner, C., Hallum, L.E., Kumbhani, R.D., Ziemba, C.M., Garcia-Marin, V., Kelly, J.G., Majaj, N.J., Movshon, J.A., Kiorpes, L.: Population representation of visual information in areas V1 and V2 of amblyopic macaques. *Vision Res.* **114**, 56–67 (2015)
11. Wang, C., et al.: The effects of left and right monocular viewing on hemispheric activation. *J. Clin. Exp. Neuropsychol.* **40**(2), 198–204 (2018)



A Critical Survey on the Evaluation of Large Language Models in Clinical Applications

Yasmin Saafi^{1(✉)}, Nizar Omheni², Zayneb Mannai¹, and Ramzi Mahmoudi^{1,3}

¹ Medical Imaging Technology Lab (LTIM-LR12ES06), Faculty of Medicine and Faculty of Sciences, University of Monastir, Monastir, Tunisia
yaasmin.saafi@gmail.com, zeyneb.mannai@gmail.com

² RedCAD Lab, ENIS, University of Sfax, Higher Institute of Computer Sciences and Mathematics, University of Monastir, Monastir, Tunisia
nizar.omheni@redcad.org

³ Gaspard Monge Computer Science Laboratory, Gustave Eiffel University, ESIEE Paris, Noisy-le-Grand, France
ramzi.mahmoudi@esiee.fr

Abstract. LLMs are being implemented for clinical applications, but are met with doubts regarding their reliability, reasonability, and safety. This paper reviews 19 studies on how clinical LLMs are evaluated and is centered around evaluation rather than model performance. We suggest an eight-criterion evaluation framework of technical, reasoning, explainability, and trust components. The findings indicate that the evaluation is poorly structured and is mostly dominated by accuracy, with an evident lack of emphasis on reasoning consistency and trust.

Keywords: LLMs · Clinical AI Evaluation · Trustworthy AI · Explainable AI

1 Introduction

LLMs are becoming prevalent in various parts of the clinical workflow, including clinical decision support, clinical documentation, and triage [7, 18]. However, their widespread adoption raises significant safety concerns. LLMs continue to hallucinate, exhibit irrationality, and generate risky and unsafe suggestions. Evaluation in general is often limited to input/output accuracy and gives only an incomplete picture of system trustworthiness, while generally not addressing explainability or clinical accuracy of reasoning. A lack of standardized protocols for clinical evaluation, varying by different datasets and metrics, makes it hard to compare performances [8]. Trustworthy reasoning and explainability, however, are crucial for adoption into healthcare settings. We introduce a unified framework that combines the technical and trust dimensions of LLMs in clinical applications. In particular, we focus on the clinical accuracy of reasoning,

explainability, and robustness. In this work, we do not provide any evaluation of the models, instead, we review how models are evaluated in the literature, and check whether key evaluation criteria are considered. This paper is structured as follows: Sect. 2 introduces the framework and the study selection, and Sect. 3 discusses the comparative study. Finally, we conclude the paper.

2 Study Selection and Evaluation Framework

We retrieved 19 clinical LLM evaluations (both peer-reviewed papers and pre-prints) covering diagnosis, retrieval, and reasoning systems. This paper focuses not on assessing clinical LLM performance, but rather on assessing their clinical LLM evaluation. We describe a framework of evaluation covering technical, reasoning, explainability, and trust. Most evaluation strategies are non-standardized, focusing purely on accuracy without considering reasoning validity, transparency, safety, or human involvement. Drawing upon principles of clinical AI, our proposed framework focuses on reasoning fidelity, traceability, uncertainty-awareness, and robustness: A more thorough evaluation than current methods of a clinical LLM. These evaluation dimensions are defined below:

1. **Diagnostic Accuracy and Stability:** Correct and stable outputs across similar inputs.
2. **Reasoning Transparency and Traceability:** Interpretable and verifiable reasoning.
3. **Completeness:** Coverage of relevant clinical information.
4. **Clinical Reliability and Readability:** Clarity and interpretability.
5. **Reasoning Fidelity:** Consistency with clinical evidence.
6. **Introspection Quality:** Uncertainty awareness and limitation detection.
7. **Human Agency and Oversight:** Human supervision and validation.
8. **Clinical Robustness and Bias Sensitivity:** Robustness across populations and bias handling.

3 Evaluation-Based Comparative Analysis of LLMs in Clinical Systems

The following table summarizes works applying the proposed framework (see Tables 1 and 2). We notice that accuracy and readability are evaluated most often, while other metrics vary from system to system. Trust is considered to some extent, while introspection and human control are almost absent. Evaluation still relies heavily on the output of the systems.

Table 1. Technical and performance evaluation.

Work	Model	Accuracy	Completeness	Reliability	Robustness
[10]	LLaMA-3	High	High	High	Limited
[2]	GPT-4	High	High	High	Limited
[13]	GPT-4	High	High	High	Limited
[12]	Multi-LLM	High	High	High	Limited
[20]	Multi-LLM	High	Moderate	Moderate	Limited
[3]	GPT-4	High	Low	Moderate	Controlled
[17]	Multi-LLM	High	Moderate	Basic	Controlled
[14]	GPT-4V	High	Moderate	Basic	Controlled
[1]	Multi-LLM	High	Moderate	Basic	Controlled
[15]	Multi-LLM	High	Full	High	Partial
[6]	Multi-LLM	High	Moderate	High	Limited
[16]	Multi-LLM	High	Moderate	High	Limited
[9]	Multi-LLM	High	High	Basic	Controlled
[4]	Multi-LLM	High	Low	Basic	Partial
[8]	GPT-4	High	Low	High	Partial
[11]	Multi-LLM	High	Moderate	Moderate	Partial
[5]	Multi-LLM	High	High	High	Limited
[19]	Multi-LLM	High	Low	Basic	Partial
[21]	Multi-LLM	High	Low	High	Partial

Table 2. Trust Evaluation

Work	Model	Traceability	Fidelity	Introspection	Human Oversight
[10]	LLaMA-3	Limited	Low	Limited	Limited
[2]	GPT-4	Limited	Low	Limited	Limited
[13]	GPT-4	Limited	Low	None	Limited
[12]	Multi-LLM	Limited	Low	None	Limited
[20]	Multi-LLM	Limited	Low	None	Limited
[3]	GPT-4	Partial	Low	None	Limited
[15]	Multi-LLM	High	High	Partial	Full
[6]	Multi-LLM	Limited	Low	None	Limited
[16]	Multi-LLM	Limited	Low	None	Limited
[9]	Multi-LLM	Rare	Low	None	Limited
[4]	Multi-LLM	Rare	Low	None	Limited
[8]	GPT-4	Rare	Low	None	Limited
[11]	Multi-LLM	Partial	Low	None	Limited
[5]	Multi-LLM	Rare	Low	None	Limited
[19]	Multi-LLM	Rare	Low	None	None
[21]	Multi-LLM	Limited	Low	None	Limited

Note: Multi-LLM indicates studies evaluating more than one model within the same experimental setup.

4 Discussion

In this work, we analyzed evaluation practices of clinical LLMs with regard to their quality of reasoning, trustworthiness, and dependability. We showed that current approaches perform well in terms of output quality (accuracy, readability) but fail to capture reasoning, explainability, and clinical trustworthiness of these models. Evaluation remains fragmented and based on the outputs. They have a scarce approach regarding traces of reasoning, uncertainty, and humans-in-the-loop verification. We presented a multi-dimensional framework to standardize clinical LLMs evaluation. This survey covers only a few studies due to the constraints of qualitative works. Future work should broaden the study and evaluate the framework on a large number of studies in real clinical settings.

5 Conclusion

We looked at the evaluation approaches of clinical LLMs with a focus on reasoning, trust, and dependability. We saw that the current approaches have good outcomes on the output quality, such as accuracy and readability, but there are considerable shortcomings when it comes to logical consistency, explainability, and reliability. Evaluation methodology appears to be fragmented and very much focused on the output without considering reasoning traces, introspection, and human in the loop. We propose a multi-dimensional evaluation methodology targeting a standardized and clinically meaningful evaluation of clinical LLMs.

References

1. Carl, N., et al.: Large language model use in clinical oncology. *NPJ Precis. Oncol.* **8**(1), 240 (2024)
2. Chen, X., et al.: Evaluating and enhancing large language models' performance in domain-specific medicine: development and usability study with Docoa. *J. Med. Internet Res.* **26**, e58158 (2024)
3. Goh, E., et al.: Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw. Open* **7**(10), e2440969–e2440969 (2024)
4. Gupta, M., Aizawa, A., Shah, R.: Med-code: medical critique based disagreement evaluation framework. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pp. 112–119 (2025)
5. Ho, C.N., et al.: Qualitative metrics from the biomedical literature for evaluating large language models in clinical decision-making: a narrative review. *BMC Med. Inform. Decis. Making* **24**(1), 357 (2024)
6. Idan, D., et al.: Evaluating the performance of large language models versus human researchers on real world complex medical queries. *Sci. Rep.* **15**(1), 37824 (2025)
7. Jiang, X., Yan, L., Vavekanand, R., Hu, M.: Large language models in healthcare current development and future directions. In: *Generative AI Research* (2023)
8. Joshi, S.: Evaluation of large language models: review of metrics, applications, and methodologies (2025)

9. Kanithi, P.K., et al.: Medic: towards a comprehensive framework for evaluating LLMs in clinical applications (2024). [arXiv:2409.07314](#) arXiv preprint
10. Kuo, S., Tai, S., Lin, H., Chen, R.: Automated clinical trial data analysis and report generation by integrating retrieval-augmented generation (RAG) and large language model (LLM) technologies. *AI* **6**(8), 188 (2025)
11. Lee, J., Park, S., Shin, J., Cho, B.: Analyzing evaluation methods for large language models in the medical field: a scoping review. *BMC Med. Inform. Decis. Making* **24**(1), 366 (2024)
12. Low, Y.S., et al.: Answering real-world clinical questions using large language model based systems. arXiv preprint [arXiv:2407.00541](#) (2024)
13. Markey, N., El-Mansouri, I., Rensonnet, G., van Langen, C., Meier, C.: From RAGs to riches: utilizing large language models to write documents for clinical trials. *Clin. Trials*, 17407745251320806 (2025)
14. Panagoulas, D.P., Palamidis, A., Virvou, M., Tsihrintzis, G.A.: A framework for evaluation and requirement extraction for fine-tuning of large language models in multimodal medical diagnosis. *Knowl. Based Syst.* 113975 (2025)
15. Peer, D., Stabinger, S.: ATA: a neuro-symbolic approach to implement autonomous and trustworthy agents (2025). [arXiv:2510.16381](#) arXiv preprint
16. Shojaei-Mend, H., Mohebbati, R., Amiri, M., Atarodi, A.: Evaluating the strengths and weaknesses of large language models in answering neurophysiology questions. *Sci. Rep.* **14**(1), 10785 (2024)
17. Shool, S., et al.: A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med. Inf. Decis. Making* **25**(1), 117 (2025)
18. Singhal, K., et al.: Large language models encode clinical knowledge. *Nature* **620**(7972), 172–180 (2023)
19. Tekchandani, N., Mukherjee, A., Poonthottam, N., Boussios, S.: Comparative analysis of large language models in dermatological diagnosis: an evaluation of diagnostic accuracy. *Cureus* **17**(9), e92089 (2025)
20. Wang, X., Ye, H., Zhang, S., Yang, M., Wang, X.: Evaluation of the performance of three large language models in clinical decision support: a comparative study based on actual cases. *J. Med. Syst.* **49**(1), 23 (2025)
21. Wei, M.Y., Li, Y.L., Liu, S.Y., Li, G.Y.: Evaluating the competence of large language models in ophthalmology clinical practice: a multi-scenario quantitative study. *Front. Cell Dev. Biol.* **13**, 1704762 (2025)



Multiresolution Fusion Network with Attention for Brain and Breast Cancer Detection

Vagish Kumar¹(✉)  and Souvik Chakraborty^{1,2} 

¹ Department of Applied Mechanics, Indian Institute of Technology Delhi,
New Delhi, India

{vagish.kumar,souvik}@am.iitd.ac.in

² Yardi School of Artificial Intelligence (ScAI), IIT Delhi, New Delhi, India

Abstract. Early and reliable detection of malignant tumors is essential for improving patient outcomes, yet traditional diagnostic workflows remain highly dependent on radiological expertise, which can be both scarce and prone to variability. Deep learning has shown promise in automating medical image analysis; however, many existing models fail to capture the subtle, multi-scale features inherent in complex medical images. In this work, we present a novel multiresolution fusion network with attention for cancer detection, targeting both brain and breast imaging modalities. Our approach combines wavelet-based multiresolution decomposition and kernel-based feature integration to extract rich representations across spatial scales. An attention mechanism is then employed to adaptively emphasize clinically relevant features and capture global contextual dependencies. The proposed method is evaluated on two distinct datasets (MRI scans for brain tumor classification and ultrasound images for breast cancer detection), demonstrating strong generalization across modalities. The model achieves classification accuracies of 98.56% for brain tumors and 93.57% for breast cancer, outperforming state-of-the-art baselines by approximately 2.4 and 8.8% points, respectively. These results highlight the effectiveness of attention-enhanced multiresolution learning for robust, cross-domain cancer diagnosis.

Keywords: Wavelet decomposition · Attention mechanism · Tumor classification · Brain tumor · Breast cancer

1 Introduction

Brain tumors and breast cancer are among the most prevalent malignancies, often diagnosed late due to asymptomatic early stages [1]. MRI and ultrasonography (USG) are primary screening modalities, but accurate interpretation depends on scarce radiological expertise [2]. Deep learning has shown strong potential for automating medical image analysis, yet conventional CNNs use fixed-size kernels and local receptive fields that struggle with the multi-scale

features inherent in tumor imaging [3]. Wavelet decomposition offers principled hierarchical frequency analysis [4], but as a fixed transform it cannot adaptively focus on discriminative regions. Self-attention [5] captures global dependencies but lacks built-in multiresolution structure.

We propose a *Multiresolution Fusion Network with Attention* that combines these complementary strengths. Unlike prior attention–CNN hybrids that operate purely in the spatial domain, the core novelty is to learn parameterized kernels *directly in the wavelet domain* at every decomposition level via element-wise modulation, and then refine the reconstructed feature map with self-attention. This yields noise-robust frequency-localized learning followed by content-adaptive global refinement in a single block.

Contributions.

- An attention-enhanced wavelet block that performs learnable element-wise modulation of wavelet coefficients at every decomposition level, then applies self-attention to the reconstructed feature map, jointly preserving multiresolution localization and enabling adaptive refinement.
- A unified architecture effective across modalities, achieving 98.56% (brain MRI) and 93.57% (breast USG) accuracy and outperforming state-of-the-art by 2.4 and 8.8% points respectively.

2 Related Work

CNN-based brain tumor classifiers [6–8] operate purely in the spatial domain and rely on fixed receptive fields, while breast USG classifiers [9, 10] face additional challenges from speckle noise and operator dependence. Self-attention has been added to CNN backbones for medical imaging tasks [11, 12], but as a post-hoc refinement on noise-mixed spatial features. Our work differs by performing learnable feature modulation *inside* the wavelet domain (separating noise from signal at every scale) before applying attention to the reconstructed representation.

3 Proposed Approach

Given an input feature map, the Discrete Wavelet Transform (DWT) [13] decomposes it into one approximation and three detail sub-bands (horizontal, vertical, diagonal) at each level, halving spatial dimensions while capturing frequency-localized information. The approximation is recursively decomposed for multi-level analysis. We use Daubechies-4 (db4) wavelets for their compact support and texture retrieval capability. We parameterize learnable kernels k_ϕ in the wavelet domain and perform element-wise multiplication with the decomposed coefficients at each level. This enables the network to learn discriminative patterns in a representation where noise and signal are more readily separable, which is particularly valuable for medical images corrupted by acquisition noise.

Let $\mathcal{W}(\cdot)$ and $\mathcal{W}^{-1}(\cdot)$ denote the forward and inverse wavelet transforms. For input feature $y_j(\mathbf{x})$ and learnable wavelet-domain kernel k_ϕ :

$$\tilde{y}_j(\mathbf{x}) = \mathcal{W}^{-1}(k_\phi \cdot \mathcal{W}(y_j(\mathbf{x}))), \quad (1)$$

where $\cdot$ denotes element-wise multiplication applied at each decomposition level. The reconstructed map $\tilde{y}_j$ is then refined with standard scaled dot-product self-attention [5]: query, key, and value projections Q, K, V are produced via 1×1 convolutions, and the spatial attention map $M_s = \text{softmax}(QK^\top / \sqrt{d_k})$ modulates the value (channel) attention map M_c to yield the refined output $a(\mathbf{x}) = M_s \otimes M_c$. The full input image is first uplifted via a 1×1 convolution and then passed through b stacked attention-enhanced wavelet blocks (Fig. 1):

$$y_{j+1}(\mathbf{x}) = \sigma(\mathcal{A}[\mathcal{W}^{-1}(k_\phi \cdot \mathcal{W}(y_j(\mathbf{x}))) + k_s \cdot y_j(\mathbf{x})]), \quad (2)$$

where $\sigma(\cdot)$ is GELU, k_s parameterizes a 1×1 convolution skip connection (which preserves the input representation and stabilizes training), and $\mathcal{A}[\cdot]$ denotes the attention operator. After b blocks the feature volume is flattened and fed to fully-connected classification layers, trained with cross-entropy loss using SGD with momentum and learning rate decay.

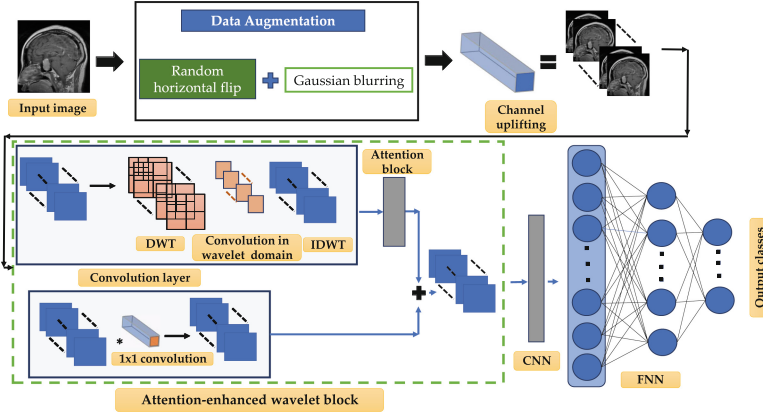


Fig. 1. Proposed architecture: stacked attention-enhanced wavelet blocks, each modulating DWT coefficients element-wise then refining with self-attention.

4 Results

Datasets and Setup. We evaluate on the BUSI dataset [14] (780 breast USG images, 3 classes) and a brain MRI dataset [15] (3,264 scans, 4 classes). All images are resized to 224×224 for parity with the CNN baselines and to keep training tractable on a single GPU; preliminary trials at higher resolution did not

improve accuracy. We report accuracy, precision, recall, and F1-score, and compare against MobileNetV2 [16], EfficientNetV2 [17], and three domain-specific methods [6–8]. Hyperparameters: db4 wavelets with 4 (brain)/6 (breast) decomposition levels, SGD (lr 0.001, momentum 0.9), batch 32, 50 epochs, weight decay 0.005 (brain)/0.001 (breast). Gaussian blurring and random horizontal flipping are used for augmentation; heavier augmentation did not help.

Table 1. Brain tumor classification results (mean $\pm$ std over multiple runs).

Method	Accuracy	Precision	Recall	F1-score
MobileNetV2 [16]	92.26 $\pm$ 1.95	92.22 $\pm$ 1.57	92.63 $\pm$ 1.66	92.10 $\pm$ 1.95
EfficientNetV2 [17]	77.48 $\pm$ 4.38	77.45 $\pm$ 4.15	77.07 $\pm$ 4.36	76.53 $\pm$ 4.61
Ismael et al. [7]	96.11 $\pm$ 1.44	96.21 $\pm$ 1.37	95.96 $\pm$ 1.38	96.07 $\pm$ 1.38
Deepak et al. [6]	95.78 $\pm$ 1.06	95.66 $\pm$ 0.99	95.74 $\pm$ 1.11	95.69 $\pm$ 1.07
Shahin et al. [8]	96.20 $\pm$ 0.71	96.23 $\pm$ 0.58	96.10 $\pm$ 0.96	96.14 $\pm$ 0.64
Proposed	98.56 $\pm$ 0.64	98.93 $\pm$ 0.44	98.83 $\pm$ 0.51	98.91 $\pm$ 0.36

Table 2. Breast cancer classification results (mean $\pm$ std over multiple runs).

Method	Accuracy	Precision	Recall	F1-score
MobileNetV2 [16]	70.42 $\pm$ 7.79	68.07 $\pm$ 4.85	65.91 $\pm$ 5.67	65.99 $\pm$ 5.1
EfficientNetV2 [17]	65.12 $\pm$ 3.49	64.79 $\pm$ 7.88	60.73 $\pm$ 5.78	59.89 $\pm$ 4.69
Ismael et al. [7]	82.22 $\pm$ 3.10	80.51 $\pm$ 3.43	79.80 $\pm$ 4.49	80.08 $\pm$ 3.95
Deepak et al. [6]	84.79 $\pm$ 0.93	83.63 $\pm$ 2.97	82.99 $\pm$ 1.32	82.93 $\pm$ 1.40
Shahin et al. [8]	81.71 $\pm$ 3.94	76.96 $\pm$ 3.02	80.33 $\pm$ 3.16	78.05 $\pm$ 3.01
Proposed	93.57 $\pm$ 0.69	90.92 $\pm$ 1.61	94.16 $\pm$ 1.55	92.52 $\pm$ 1.53

As shown in Tables 1 and 2, the proposed model achieves 98.56% (brain) and 93.57% (breast) mean accuracy, surpassing the prior method by 2.4 and 8.8% points respectively. On brain MRI, our model perfectly classifies meningioma, pituitary, and no tumor; only 5% of glioma samples are misclassified as meningioma, a known clinical difficulty due to similar MRI appearance. On breast USG, per-class accuracies are 100% (normal), 96.88% (malignant), and 90.91% (benign); benign cases occasionally exhibit features similar to either malignant or normal tissue.

The performance gain is because of the wavelet-domain learning: parameterizing kernels element-wise on DWT coefficients at every level lets the network operate in a representation where high-frequency acquisition noise is spatially separated from texture and edges, isolating the discriminative signal before attention is applied.

5 Conclusion

We proposed a multiresolution fusion network combining wavelet-domain learnable kernels with self-attention for medical image classification, achieving 98.56% (brain MRI) and 93.57% (breast USG) accuracy. Both datasets are single-institution; the strong results and absence of gain from heavy augmentation suggest these public benchmarks may not fully reflect real-world clinical variability. Future work includes federated learning across multi-institutional data, continuous wavelet transforms for finer-grained frequency analysis, and interpretability mechanisms to support clinical adoption.

Acknowledgements. V. Kumar acknowledges the Prime Minister’s Research Fellowship (PMRF) from MoE, India. S. Chakraborty acknowledges support from ANRF (CRG/2023/007667) and the Ministry of Ports, Shipping and Waterways (ST-14011/74/MT (356529)).

References

1. Siegel, R.L., et al.: Cancer statistics, 2021. *CA Cancer J. Clin.* **71**(1), 7–33 (2021)
2. McDonald, R.J., et al.: The effects of changes in utilization and technological advancements on radiologist workload. *Acad. Radiol.* **22**(9), 1191–1198 (2015)
3. Rangel, G., et al.: A survey on CNNs and their performance limitations. *J. Sens.* **2024**(1), 2797320 (2024)
4. Mallat, S.G.: A theory for multiresolution signal decomposition. *IEEE TPAMI* **11**(7), 674–693 (1989)
5. Vaswani, A.: Attention is all you need. *NeurIPS* (2017)
6. Deepak, S., Ameer, P.: Brain tumor classification using deep CNN features via transfer learning. *Comput. Biol. Med.* **111**, 103345 (2019)
7. Ismael, S.A.A., et al.: An enhanced deep learning approach for brain cancer MRI images classification. *Artif. Intell. Med.* **102**, 101779 (2020)
8. Shahin, A.I., et al.: MBTFCN: a novel modular fully convolutional network for MRI brain tumor multi-classification. *Expert Syst. Appl.* **212**, 118776 (2023)
9. Karthik, R., et al.: Gaussian dropout based stacked ensemble CNN for classification of breast tumor in ultrasound images. *IRBM* **43**(6), 715–733 (2022)
10. Eroğlu, Y., et al.: CNN-based classification of breast ultrasonography images. *Comput. Biol. Med.* **133**, 104407 (2021)
11. Wu, W., et al.: Scale-aware transformers for diagnosing melanocytic lesions. *IEEE Access* **9**, 163526–163541 (2021)
12. Shaik, N.S., Cherukuri, T.K.: Multi-level attention network for brain tumor classification. *SIViP* **16**(3), 817–824 (2022)
13. Daubechies, I.: *Ten Lectures on Wavelets*. SIAM (1992)
14. Al-Dhabyani, W., et al.: Dataset of breast ultrasound images. *Data Brief* **28**, 104863 (2020)
15. Bhuvaji, S.: Brain Tumor Classification (MRI). Kaggle (2020)
16. Sandler, M.: MobileNetV2: Inverted residuals and linear bottlenecks, pp. 4510–4520. *CVPR* (2018)
17. Tan, M., Le, Q.: EfficientNetV2: Smaller models and faster training. In: *ICML*, pp. 10096–10106. (2021)



Fairness-Aware Time Series Decision Support for Virtual Ward Triage

Basra Jehangir^{1(✉)}, Weizi Vicky Li^{1(✉)}, Dionne Tannetta²,
and Andrew Walden²

¹ Informatics Research Centre, Henley Business School, University of Reading,
Reading, UK

basra@pgr.reading.ac.uk, weizi.li@henley.ac.uk

² Acute Medical Unit, Royal Berkshire NHS Foundation Trust, Reading, UK
dionne.tannetta@royalberkshire.nhs.uk, andrew.walden@nhs.net

Abstract. Virtual wards are increasingly used across the NHS and globally to provide hospital-level care at home for clinically stable patients, but triage between discharge and virtual ward admission is challenging, as poor decisions can lead to deterioration or readmission. Effective triage requires integrating multivariate, time-evolving clinical data, which is difficult with manual heuristics alone. While data-driven decision-support systems can help, they raise concerns about demographic fairness, especially with imbalanced data. This study presents a time-series binary classification framework for virtual ward triage using a Light Gradient Boosting Machine trained on multivariate temporal features, achieving strong performance (accuracy 0.858, AUC 0.939, F1 0.871, precision 0.909, recall 0.835), with comparisons to other models. Initial results showed disparities across ethnic groups, prompting post-hoc fairness adjustments that reduced the equal opportunity gap from 0.112 to 0.075 while maintaining overall performance. These adjustments increased sensitivity for underrepresented groups with a modest rise in false positives, representing a clinically acceptable, safety-focused trade-off. Overall, the work demonstrates that fairness-aware time-series models can support safer and more equitable virtual ward triage in real-world settings.

Keywords: Virtual ward · Clinical decision support · Machine learning in healthcare · Fairness-aware learning · Remote patient monitoring · Light Gradient Boosting Machine

1 Introduction

Virtual wards [1] provide hospital-level care at home for clinically stable patients, helping reduce pressure on acute services while supporting early discharge and admission avoidance. Because patients remain under hospital oversight, deciding who is suitable for a virtual ward is a safety-critical triage task, poor decisions can lead to unsafe discharges or readmissions [5].

These triage decisions are complex, requiring clinicians to integrate diverse and changing patient data. In practice, decisions often rely on clinical judgement and simple scoring tools, which may overlook trends and interactions, leading to inconsistent outcomes [4].

Data-driven decision-support systems could improve this process by better analyzing complex patient information. However, using machine learning in this context raises concerns about fairness, as models trained on routine healthcare data may inherit demographic biases and produce unequal outcomes [3].

To address this, the work proposes a fairness-aware, time-series machine learning framework that predicts, 24 h in advance, whether a patient should remain in a virtual ward or be discharged. Using NHS data, the approach aims to improve safety, reduce readmissions, and ensure more equitable decision-making, while acknowledging broader concerns about algorithmic bias in healthcare.

2 Methodology

A time-series machine learning framework was developed for NHS virtual ward triage to predict 24 h in advance whether patients should be admitted or discharged home, integrating model evaluation with fairness-aware learning. The section outlines the methodology used.

2.1 Data Acquisition

The study used anonymised data from about 7,000 patients at a Southeast England hospital, including time-varying features (vital signs and lab results for time-series modelling) and static features (demographics such as age, sex, and comorbidities), with categorical variables one-hot encoded for model use.

2.2 Model Training and Evaluation

A Light Gradient Boosting Machine (LightGBM) classifier was used to predict, 24 h in advance, whether patients should be admitted to a virtual ward or discharged home, chosen for its strong performance on structured clinical data and ability to capture nonlinear relationships. The learning task is a binary classification problem, with target labels 0 (home discharge) and 1 (virtual ward admission). Time-series physiological data collected 6 hourly were z-score normalised and converted into fixed-length feature vectors by flattening 10-step sequences per admission. Models were trained using an 80:20 admission-level split with early stopping, and compared against XGBoost, logistic regression, and random forest under consistent preprocessing, median imputation, and class imbalance handling. Median imputation may not generalise across subgroups; future work will use stratified or model-based imputation methods. Performance was evaluated on a held-out test set.

2.3 Fairness Analysis

Post-hoc fairness analysis was conducted using the OxonFair framework [2] to assess and mitigate demographic bias without retraining the LightGBM model. The model was wrapped in a Fair Predictor, enabling joint optimisation of accuracy and fairness constraints to reduce disparities across protected groups, such as ethnicity, while preserving overall performance.

Fairness was evaluated on a held-out test set using accuracy, precision, recall, demographic parity(DP), equal opportunity gap (EO), equalized odds gap (ESO), and selection rate (SR) metrics, capturing disparities in overall predictions and outcomes conditional on true labels across groups. Despite class imbalances, this approach supports both accurate and equitable triage decisions.

3 Results

This section assesses the performance and fairness of a virtual ward triage model using patient data (demographics, labs, vitals, comorbidities), presenting classification metrics followed by an ethnic group fairness analysis (Table 1).

Table 1. Performance of predicting virtual ward admission eligibility.

Model	Accuracy	ROC-AUC	PR-AUC	F1 Score	Precision	Recall	Specificity	Log Loss
LightGBM	0.8580	0.9390	0.9594	0.8705	0.9090	0.8352	0.8883	0.3032
XGBoost	0.8530	0.9367	0.9571	0.8678	0.8895	0.8471	0.8607	0.3098
Random Forest	0.8358	0.9198	0.9441	0.8458	0.9092	0.7907	0.8955	0.3939
Logistic Regression	0.7545	0.8337	0.8645	0.7836	0.7866	0.7807	0.7197	0.5186

The analysis compares four classification models. LightGBM performs best overall, achieving the highest accuracy (0.858), AUC (0.939), and F1 score (0.871), with strong balance between precision (0.909) and recall (0.835). XGBoost performs similarly, with accuracy of 0.853 and AUC of 0.937, and attains the highest recall (0.847) but slightly lower precision (0.889). Random Forest achieves competitive accuracy (0.836) but shows greater imbalance between precision and recall, resulting in a lower F1 score (0.846). Logistic Regression underperforms across all metrics, particularly AUC (0.834) and F1 score (0.784). Based on these results, LightGBM is selected for subsequent fairness analysis due to its consistently superior performance.

After applying fairness constraints using the OxonFair framework (Table 2), the model shows a modest drop in accuracy (0.858 to 0.839) and precision (0.909 to 0.860), with improved recall (0.835 to 0.859). The reweighting model instead slightly reduces accuracy (0.854) and recall (0.797), but improves demographic parity (0.194 to 0.174), indicating more balanced predictions across groups. However, this comes with higher equal opportunity (0.112 to 0.178) and equalized

Table 2. Comparison of original and fairness-adjusted models.

Model	Accuracy	Precision	Recall	DP	EO	ESO
Original	0.8580	0.9090	0.8352	0.1943	0.1116	0.0908
Adjusted (Reweighting)	0.8536	0.93823	0.7965	0.1744	0.1782	0.1085
Oxonfair	0.8394	0.8600	0.8589	0.1953	0.0746	0.1437

Table 3. True Positive and False Positive Rates across ethnic groups before and after fairness adjustment.

Group	Original		Adjusted		OxonFair	
	TPR	FPR	TPR	FPR	TPR	FPR
Other	0.636364	0.063380	0.6000	0.0563	0.927273	0.338028
White	0.849655	0.132850	0.8124	0.0772	0.851034	0.137681
Asian	0.857143	0.085714	0.8571	0.0571	0.928571	0.085714
Black	0.857143	0.000000	0.5714	0.000	1.000000	0.444444

odds (0.091 to 0.109) gaps, reflecting increased error-rate disparity. Overall, fairness methods produce different trade-offs depending on the objective

Table 3 reports ethnic group specific True Positive Rates (TPR) and False Positive Rates (FPR) before and after fairness adjustment. Fairness constraints adjust decision thresholds or weights, altering confusion matrix outcomes and derived metrics. After adjustment, TPR increases across groups (e.g., “Other” $0.636 \rightarrow 0.600 \rightarrow 0.927$; “Black” $0.857 \rightarrow 0.571 \rightarrow 1.0$), indicating improved identification of patients requiring virtual ward care. FPR also changes (e.g., “Other” $0.063 \rightarrow 0.056 \rightarrow 0.338$; “Black” $0 \rightarrow 0 \rightarrow 0.444$), reflecting threshold rebalancing, particularly due to small sample size in subgroups (Fig. 1(a)), while remaining stable for Asian patients.

However, the reweighting model is less suitable clinically as it does not prioritise virtual ward allocation effectively and increases error-rate imbalance. In contrast, OxonFair provides a better trade-off by improving recall and supporting safer triage decisions, making it preferable for this application.

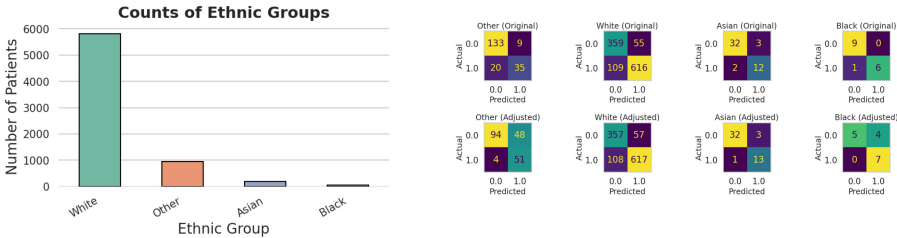


Fig. 1. Model performance and dataset characteristics (a) Imbalance among different ethnic groups, (b) Ethnic group-wise confusion matrix.

Figure 1 (b) summarises ethnic group-specific confusion metrics, showing that fairness interventions improve equity in TPR and selection rate (SR) but increase false positives for some groups, highlighting inherent trade-offs in fairness-constrained models (Table 4).

Table 4. Ethnic group-wise SR comparison between the original model and Oxonfair.

Model	Asian	Black	Other	White
Oxonfair	0.367347	0.750000	0.522843	0.595259
Original	0.285714	0.375000	0.177665	0.624232

The LightGBM model demonstrates strong baseline performance, and fairness constraints improve equity across ethnic groups with minimal accuracy loss. Post-adjustment changes in TP, FP, TN, and FN influence triage decisions: increased TP in minority groups (“Other” and “Black”) improves identification and admission of patients needing virtual ward care, enhancing equitable access. Higher FP rates particularly in these groups mean more patients are considered for virtual wards, prioritising caution and reducing unsafe discharge risk. FN decreases are greatest in the “Other” group (20 to 4), with smaller reductions elsewhere, while TN changes further support a safer, more inclusive approach. Overall, fairness interventions enhance access for minority groups while maintaining patient safety, supporting more equitable and reliable triage decisions.

4 Conclusion

This study shows the feasibility and value of a fairness-aware time-series decision-support framework for NHS virtual ward triage. The LightGBM model achieved strong predictive performance, enabling accurate identification of patients who may benefit from continued remote monitoring while reducing unsafe discharges.

Beyond accuracy, the study emphasises equitable performance in safety-critical settings. Observed ethnic disparities led to post-hoc fairness optimisation, which improved access to virtual ward monitoring for underrepresented groups while maintaining overall model performance. This resulted in increased true positives and a modest rise in false positives, reflecting a clinically acceptable trade-off that prioritises patient safety over premature discharge.

Future work will include external validation across healthcare settings, prospective evaluation of clinical impact, and ongoing monitoring of both performance and fairness as services and patient populations evolve.

Acknowledgments. This work was funded by the UK Engineering and Physical Sciences Research Council (EPSRC, grant number EP/W000652/1) and the Health Innovation Partnership fund of Royal Berkshire NHS Foundation Trust and the University of Reading.

Disclosure of Interests. The authors declare no conflicts of interest.

References

1. Cucurachi, S., Lydon, S., Moens, L.L., Manser, T., O'Connor, P.: Barriers and facilitators to the use of virtual wards: a systematic review of the qualitative evidence. *Int. J. Qual. Health Care* **37**(3), mzaf065 (2025)
2. Delaney, E., Fu, Z., Wachter, S., Mittelstadt, B., Russell, C.: Oxonfair: a flexible toolkit for algorithmic fairness. *Adv. Neural. Inf. Process. Syst.* **37**, 94209–94245 (2024)
3. Hong, C., Liu, M., Wojdyla, D.M., Hickey, J., Pencina, M., Henao, R.: Transbalance: reducing demographic disparity for prediction models in the presence of class imbalance. *J. Biomed. Inform.* **149**, 104532 (2024)
4. Jalilian, A., Sedda, L., Unsworth, A., Farrier, M.: Length of stay and economic sustainability of virtual ward care in a medium-sized hospital of the UK: a retrospective longitudinal study. *BMJ Open* **14**(1), e081378 (2024)
5. Vernon, D., Brown, J.E., Griffiths, E., Nevill, A.M., Pinkney, M.: Reducing readmission rates through a discharge follow-up service. *Future healthc. j.* **6**(2), 114–117 (2019)



When More Data is Not Better: Physiological Drift Explains Transfer Learning Failure in Rare ICU Populations

Michael Poette^{1,2} , Sandrine Mouysset³ , Daniel Ruiz³ ,
Jean-Marc Alliot^{1,3} , Fanny Vardon¹ , and Vincent Minville^{1,2}

¹ Toulouse University Hospital, Toulouse, France

² RESTORE, Toulouse University, Inserm UMR1301, Toulouse, France
`poette.m@chu-toulouse.fr`

³ IRIT, Toulouse University, CNRS UMR5505, Toulouse, France

Abstract. Transfer learning (TL) is frequently proposed as the standard solution for predictive modeling in rare clinical cohorts where data is scarce. In this study, we investigate mortality prediction for patients under Veno-Venous Extracorporeal Membrane Oxygenation (VV-ECMO), a high-stakes but low-volume ICU subpopulation. We compared models trained from scratch on a small ECMO cohort ($n = 148$) against models pre-trained on a large source population of mechanically ventilated patients ($n = 5,731$). Contrary to prevailing assumptions, TL consistently failed to outperform simple models trained on target data only. Using SHAP-based feature importance analysis, we demonstrate that this failure is rooted in a fundamental physiological shift: while mortality in standard respiratory failure is driven by metabolic and hemodynamic exhaustion, outcomes in ECMO patients depend primarily on respiratory mechanics and oxygenation response. Furthermore, strong selection bias in ECMO candidacy neutralizes distinct predictors like age. Our results suggest that in specialized critical care, “more data” can introduce harmful noise if the underlying physiological signatures diverge (concept drift).

Keywords: Transfer Learning · VV-ECMO · SHAP · Concept drift

1 Introduction

The integration of Artificial Intelligence (AI) into Intensive Care Units (ICU) has been largely fueled by the emergence of massive, multicenter Electronic Health Records (EHR) [9]. However, due to the inherent data scarcity affecting niche cohorts, prognostic tools still rely heavily on traditional logistic regression scores derived from limited samples. This gap is particularly striking for patients under Veno-Venous Extracorporeal Membrane Oxygenation (VV-ECMO), an extracorporeal circuit providing blood oxygenation for severe respiratory failure when lungs can no longer ensure gas exchange [6, 8]. These patients represent

an extreme phenotype of clinical complexity in the sense that the initiation of the extracorporeal circuit induces a “*physiological decoupling*” where established correlations between ventilator settings, blood gases, and clinical outcomes are fundamentally altered [4].

Consequently, the potential of advanced machine learning, such as time-series analysis, remains under-explored in this setting despite its theoretical benefits [7]. Transfer Learning (TL) offers a compelling solution, based on the premise that knowledge extracted from a sufficiently large and related source dataset can be successfully adapted to a proximal, specialized target task [1].

In this study, we investigate whether proximity learning paradigms can bridge the performance gap for highly specialized ICU populations. By evaluating a representative panel of state-of-the-art AI architectures, we explore the efficiency of transfer learning in adapting knowledge from general mechanically ventilated patients to the specific complexities of ECMO care.

2 Methods

Data and Population. This retrospective study used EHR data from Toulouse University Hospital (2014–2024). We defined a **Source Domain** (D_V , $n = 5,731$) of ventilated patients and a **Target Domain** (D_E , $n = 148$) of VV-ECMO patients. The observation window was the first 24 h following the initiation of support (intubation for D_V , ECMO initiation for D_E). The primary outcome was 28-day all-cause mortality.

Feature Engineering. A total of 19 input features were included: (1) *Demographics*: Age, sex; (2) *Vital Signs*: Heart rate, peripheral oxygen saturation (SpO_2), temperature, and blood pressure (systolic, mean, diastolic); (3) *Laboratory values*: Lactate, pH, bicarbonate (HCO_3^-), partial pressure of arterial oxygen (PaO_2), partial pressure of arterial carbon dioxide (PaCO_2), mixed venous oxygen saturation (SvO_2), and capillary glucose; (4) *Respiratory mechanics*: Fraction of inspired oxygen (FiO_2), Positive End Expiratory Pressure (PEEP), and plateau pressure; (5) *Neurology*: Patient consciousness based on nursing bedside assessments binarized as a boolean feature. We structured these features into two distinct representations: (1) **Aggregated Features**, which encompass the mean, standard deviation, minimum, and maximum of physiological signals, alongside static variables, for traditional machine learning models; and (2) **Sequential Tensors** (N, T, F), specifically designed for Long Short-Term Memory (LSTM) architectures.

Experimental Design. We compared RF [2], XGBoost [3], and LSTM [10] models using class weighting to manage imbalance. Following baseline validation on D_V , three strategies were evaluated: (1) Direct Training on target data (2) Zero-Shot transfer, and (3) Fine-Tuning, implemented as warm-start tree addition (RF), leaf-value refresh (XGBoost), and shallow unfreezing of the last LSTM layer plus classifier head. Performance (AUROC) was assessed across 100

independent runs using a 70/30 stratified split. Feature importance was analyzed using SHAP values derived from RF models. Full code, fine-tuning hyperparameters, and extended cohort statistics are available at https://github.com/mpoette/transfer_ecmo to ensure reproducibility.

3 Results

Patient Characteristics. Despite similar severity scores (SAPS II), the populations diverge on some aspects (Table 1) : the ECMO cohort is younger (52.5 vs 58.6 years) and exhibits more severe respiratory failure (higher PEEP/ FiO_2) but greater hemodynamic stability (lower Lactate).

Table 1. Main Patient Characteristics. Data represent mean $\pm$ SD or counts (%).

Variable	Ventilation ($n = 5,731$)	VV-ECMO ($n = 148$)
Age (years)	58.6 ± 16.3	52.5 ± 13.1
SAPS II Probability	0.60 ± 0.30	0.60 ± 0.30
In-hospital Mortality	1,325 (23.1%)	25 (16.9%)
Lactate (mmol/L)	2.5 ± 2.1	1.9 ± 1.6
FiO_2 (%)	45.4 ± 15.0	56.2 ± 17.8
PEEP (cmH ₂ O)	6.8 ± 2.5	11.5 ± 3.8

Performance. TL provided no benefit (Table 2). Indeed, best AUROC on ECMO was achieved by models trained from scratch (RF: 0.67 ± 0.07 , LSTM: 0.67 ± 0.09). Fine-tuning failed to surpass these baselines.

Table 2. AUROC Performance (100 seeds). Best results on Target Dataset (D_E) by model in bold.

Model	Source (D_V)	Target Performance (D_E)		
	Baseline	<i>Direct</i>	<i>Zero-Shot</i>	<i>Fine-Tuning</i>
XGBoost	0.75 ± 0.01	0.63 ± 0.08	0.59 ± 0.08	0.61 ± 0.11
Random Forest	0.74 ± 0.01	0.67 ± 0.07	0.60 ± 0.09	0.67 ± 0.07
LSTM	0.75 ± 0.01	0.67 ± 0.09	0.63 ± 1.10	0.66 ± 0.10

Feature Shift. SHAP analysis was conducted on the RF models, as they demonstrated superior predictive performance among non-deep learning algorithms. This analysis revealed a negligible rank correlation ($\rho = 0.095$) between feature importance in the source and target datasets. The respective feature rankings

are illustrated in Fig. 1. In the source population, mortality prediction is primarily driven by metabolic markers (Lactate, Rank 1; HCO_3^- , Rank 2). Conversely, respiratory mechanics dominate the ECMO cohort (FiO_2 , Rank 1; SpO_2 , Rank 2). Notably, some traditional predictors lose their prognostic value in the specialized cohort; for instance, Age drops from Rank 5 in the source population to Rank 18 in the ECMO group.

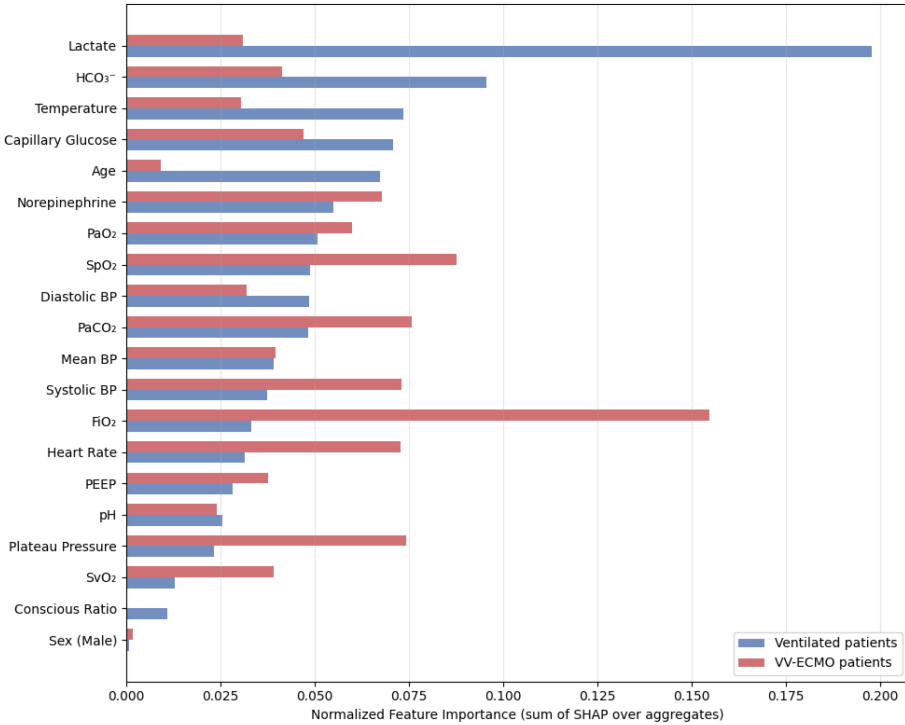


Fig. 1. Feature importance shift between source and target populations. SHAP-based feature importance comparing Random Forest models trained on mechanically ventilated patients (blue, $n = 5,731$) versus VV-ECMO patients (red, $n = 148$). Importance values represent the sum of mean absolute SHAP values across temporal aggregates (mean, SD, min, max) for each physiological variable.

4 Discussion and Conclusion

In our study, transfer learning (TL) failed to improve mortality prediction for VV-ECMO patients despite a clinically proximal source. While constrained by the small target cohort, these results highlight a gap between clinical proximity and statistical similarity. The initiation of an extracorporeal circuit induces a “physiological decoupling”: the mortality signal shifts from metabolic exhaustion in standard respiratory failure to respiratory mechanics and oxygenation

response in ECMO. This concept drift [5] suggests that “more data” can introduce harmful noise when physiological signatures diverge. Compounding this, selection bias in ECMO candidacy homogenizes the cohort, neutralizing predictors such as age. For rare ICU subpopulations, small domain-specific models currently outperform large pre-trained architectures, underscoring the need for AI methodologies dedicated to rare clinical conditions [11].

Acknowledgments. This work was supported by the Occitanie region (France) as part of the “IA pour la Santé” program.

Disclosure of Interest. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Badawi, M., Abushanab, M., Bhat, S., Maier, A.: Review of zero-shot and few-shot ai algorithms in the medical domain (2024). <https://doi.org/10.48550/arXiv.2406.16143>. [arXiv:2406.16143](https://arxiv.org/abs/2406.16143)
2. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
3. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794 (2016). <https://doi.org/10.1145/2939672.2939785>, [arXiv:1603.02754](https://arxiv.org/abs/1603.02754)
4. Fresiello, L., Hermens, J.A., Pladet, L., Meuwese, C.L., Donker, D.W.: The physiology of venoarterial extracorporeal membrane oxygenation - a comprehensive clinical perspective. *Perfusion* **39**(1_suppl), 5S–12S (2024). <https://doi.org/10.1177/02676591241237639>
5. Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., Zhang, G.: Learning under concept drift: a review. *IEEE Trans. Knowl. Data Eng.*, 1–1 (2018). <https://doi.org/10.1109/TKDE.2018.2876857>, [arXiv:2004.05785](https://arxiv.org/abs/2004.05785)
6. Neumann, E., et al.: Predictors associated with mortality of veno-venous extracorporeal membrane oxygenation therapy. *J. Thor. Dis.*, **15**(5), 2389–2401 (2023). <https://doi.org/10.21037/jtd-22-1273>
7. Pladet, L., et al.: Clinical decision support for extracorporeal membrane oxygenation: will we fly by wire? *Perfusion* **38**(1 Suppl), 68–81 (2023). <https://doi.org/10.1177/02676591231163688>
8. Rilinger, J., et al.: A mortality prediction score for patients with veno-venous extracorporeal membrane oxygenation (VV-ECMO): the PREDICT VV-ECMO score. *ASAIO J.* **70**(4), 293 (2024). <https://doi.org/10.1097/MAT.0000000000002088>
9. Sun, H., Kang, M., Zhang, H., Jia, J., Wang, Q.: Machine learning for predicting mortality in intensive care unit patients: a prognostic performance systematic review and meta-analysis. *Nurs. Crit. Care* **30**(6), e70206 (2025). <https://doi.org/10.1111/nicc.70206>
10. Vennerød, C.B., Kjærran, A., Bugge, E.S.: Long short-term memory RNN (2021). <https://doi.org/10.48550/arXiv.2105.06756>. <http://arxiv.org/abs/2105.06756>, [arXiv:2105.06756](https://arxiv.org/abs/2105.06756)
11. Zhu, M., Liu, Y., Luo, Z., Zhu, T.: Bridging data gaps of rare conditions in ICU: a multi-disease adaptation approach for clinical prediction. *NPJ Digit. Med.*, **9**(1), 7 (2026). <https://doi.org/10.1038/s41746-025-02176-y>



Informative Missingness to Generate Irregular Clinical Time Series

Hadi Mehdizavareh^{1(✉)}, Gabriele Santangelo², Giovanna Nicora²,
Simon Lebech Cichosz¹, Arianna Dagliati², Arijit Khan^{1,3},
and Riccardo Bellazzi^{1,2}

¹ Aalborg University, Aalborg, Denmark
mhme@cs.aau.dk

² University of Pavia, Pavia, Italy

³ Bowling Green State University, Bowling Green, OH, USA

Abstract. Laboratory tests in electronic health records are collected irregularly, and the absence of a test order can be as informative as the measurement itself. Such missingness reflects clinicians' decisions and patient physiology, making it important to model it directly rather than treat it as a preprocessing artifact. Here we present a diffusion-based approach for generating clinical time series that jointly models laboratory values and their observation patterns using the public Data Analytics Challenge on Missing data Imputation (DACMI) benchmark derived from MIMIC-III. To preserve realistic sampling, we align chart times into 4-hour intervals and segment admissions into 7-day windows, producing trajectories that pair each lab value with a corresponding observation indicator. Standard transformations and normalization are applied to stabilize training. Our method extends the TimeDiff framework to learn continuous lab values and discrete missingness patterns through complementary diffusion objectives. Experiments show that the generated data closely match real patient trajectories across individual lab distributions and joint value-missingness embeddings, demonstrating that diffusion models can capture clinically meaningful dependencies between patient physiology and clinicians' testing behavior under MNAR-like (missing-not-at-random) missingness. These preliminary results indicate that our model can serve as an initial component toward developing clinical foundation models. By producing synthetic priors that preserve key physiology-missingness relationships, this work motivates the subsequent training of Prior-Data Fitted Networks capable of leveraging informative missingness, which we will investigate in the extended work.

Keywords: Clinical time series · Diffusion models · Informative missingness · Synthetic EHR generation

1 Background

Electronic health records (EHRs) record irregular sequences of measurements shaped by physiology, patient behavior, and clinical decisions [8]. As an example, labs are rarely collected on a fixed schedule, and the absence of testing

can itself reflect severity or clinical intent. Prior work shows that lab missingness is often informative rather than random and can differ across sites and care settings [6, 10]. For generative modeling, this means that producing realistic synthetic lab trajectories requires matching not only lab values, but also when tests occur, how they appear in bursts, and how sampling relates to the underlying values. Many clinical time-series pipelines handle missingness mainly through imputation, aiming to densify sparse records before modeling [3, 4]. The DACMI challenge [5], built from MIMIC-III [2], offers a public benchmark for imputation methods with ground truth for 13 common blood labs. However, success on imputation benchmarks does not necessarily show that a generator captures the joint process of values and observations under real EHR sampling. This is most critical in the Missing-Not-At-Random (MNAR) setting, where whether a lab is measured depends on unobserved or partially observed clinical state, making missingness tightly coupled to the data-generating mechanism. Diffusion models are state-of-the-art option for time-series generation [1, 9]. TimeDiff [7] is among the first diffusion approaches tailored to EHR sequences and supports mixed continuous and binary variables. Still, much prior evaluation focuses on near-regular signals (e.g., vitals) and reports missingness mainly through marginal rates, leaving open whether models reproduce irregular lab sampling bursts and their coupling with values—an issue that can persist even when univariate distributions look correct.

Contributions. We present a generative framework designed as the preliminary step for developing clinical foundation models, providing high-fidelity synthetic priors required to train Prior-Data Fitted Networks [12]. Our contributions are: (i) A diffusion-based synthetic lab generation pipeline on a public benchmark that targets irregular labs and informative missingness, generating both lab values and their observation indicators jointly to support downstream inference. (ii) We modify the numerical (Gaussian) loss in TimeDiff [7] to exclude truly missing lab values, treating missingness as absence rather than noisy values. (iii) We present a practical preprocessing and windowing protocol designed to preserve MNAR-like structure while remaining compatible with diffusion training. (iv) We evaluate distributional similarity using univariate lab distributions, joint embeddings over values and masks, and additional similarity metrics for model selection to assess realism in both physiology and sampling behavior.

2 Materials and Methods

Dataset. We used the publicly available DACMI imputation challenge dataset [5], derived from MIMIC-III, which provides a standardized train/test split designed for fair comparison of methods on sparse clinical time series. From each admission, we extracted 13 laboratory variables and constructed a corresponding missingness mask directly from the raw tables, where each lab at each time step is marked as observed (0) or missing (1). This yields a paired representation of physiology (continuous lab values) and sampling behavior (binary

indicators) that is suitable for studying informative missingness, especially in the MNAR-like regime where ordering and frequency of tests reflect clinical decisions. Models development and selection are based on the DACMI training split.

Setup and Preprocessing. We discretized `charttime` (minutes since admission) into 4-hour intervals to minimize artificial missingness, creating 42-step windows (7 d). Admissions exceeding this horizon were segmented into consecutive non-overlapping windows, yielding 16,580 training samples from 8,267 admissions. To approximate Gaussianity, features underwent log or Box-Cox transformations followed by z-score standardization. The model input consists of 26 channels (13 standardized values and 13 binary masks). In addition to the values and binary mask trajectories, we used *delta conditioning* to encode observation recency. The delta conditioning provides an auxiliary conditioning tensor to the denoising network at each diffusion step. Therefore, the model still generates only the lab values and mask variables, but is guided by this extra context during denoising. For each lab and time index, the delta value represents the number of steps since the most recent observation. Specifically, at each time index, we computed the number of time steps that must be traversed backward from the current position to reach the latest time point at which the lab is observed (mask = 0). This yields a (42, 13) matrix per window, normalized to [0, 1) by division by 42.

Experimental Setup and Sampling. We trained the model for 200,000 steps using the Adam optimizer (learning rate 8×10^{-5} , batch size 64). To balance our two objectives, we applied a weight of $\lambda = 10^{-5}$ to the discrete missingness loss relative to the continuous value loss, adopting the notation from TimeDiff [7]. For evaluation, we generated 16,000 synthetic windows (42 steps $\times$ 26 channels) by iteratively denoising random noise. This sample size matches the real dataset, allowing us to compare the synthetic distributions and observation patterns against the ground truth.

3 Results and Discussions

Figure 1, panels A–B compare real and synthetic univariate distributions for two representative labs with different shapes. Hemoglobin (A) is fairly concentrated with a mild right tail, and the synthetic density closely matches the real histogram around the main mode and across the upper tail. White Blood Cells count (B) is markedly more right-skewed with a long tail driven by occasional high values; despite this more challenging setting, the synthetic samples follow the real distribution well in the high-density region and preserve the overall tail profile.

Figure 1 panel C compares the overall distribution of real data with the synthetic ones. The embedding was computed using UMAP with a DTW-based distance metric, for all the 26 channels (13 values and 13 masks). The substantial overlap between the two distributions indicates that the generative approach

produces high-quality synthetic samples, without generating instances outside the support of the real data.

Table 1. Per-feature comparison across 13 labs values: missingness prevalence (mask=1), univariate EMD on lab-value marginals, and DTW trajectory distance.

Lab	Missingness prevalence			EMD (univariate)			DTW (trajectory)		
	Real	Synth w/ Δ	Synth w/o Δ	R-R w/ Δ	R-S w/o Δ		R-R w/ Δ	R-S w/o Δ	
Chloride	0.78	0.72	0.74	0.06	0.09	0.11	37.71	37.63	38.09
Potassium	0.78	0.72	0.74	0.03	0.06	0.09	38.44	38.27	38.48
Bicarb	0.78	0.72	0.74	0.07	0.14	0.14	37.90	37.80	38.13
Sodium	0.78	0.72	0.74	0.05	0.12	0.09	37.91	37.77	38.20
Hematocrit	0.80	0.75	0.77	0.06	0.06	0.03	38.34	38.83	38.82
Hemoglobin	0.81	0.76	0.78	0.02	0.11	0.07	38.42	38.97	39.00
MCV	0.81	0.76	0.78	0.03	0.16	0.20	37.73	38.00	38.12
Platelets	0.81	0.76	0.77	0.08	0.20	0.24	37.72	37.62	37.83
WBC count	0.81	0.76	0.77	0.08	0.07	0.08	37.88	38.01	38.15
RDW	0.81	0.76	0.78	0.04	0.09	0.13	37.53	37.53	37.88
BUN	0.78	0.72	0.74	0.08	0.10	0.15	36.41	36.70	37.10
Creatinine	0.78	0.72	0.74	0.02	0.06	0.08	35.99	35.87	36.40
Glucose	0.78	0.72	0.74	0.03	0.07	0.04	37.51	37.43	37.72
Mean	0.79	0.74	0.76	0.05	0.10	0.11	37.65	37.72	37.99

Table 1 summarizes per-feature realism across the 13 lab variables. It reports (i) missingness prevalence (fraction of time points with mask=1) in real windows and in synthetic windows generated with and without delta conditioning, (ii) univariate Earth Mover’s Distance (EMD) [11] between lab-value marginals using Real–Real as a same-domain reference and Real–Synthetic for each generator, and (iii) Dynamic Time Warping (DTW) trajectory distances computed per feature on a random subsample of $N = 100$ windows. Both EMD and DTW are computed in a missingness-aware manner by restricting comparisons to time points marked as observed according to the real or generated masks.

Overall, delta conditioning yields slightly better agreement with the real data: the mean EMD and DTW are lower than the corresponding no-conditioning setting and closer to the Real–Real reference. In contrast, the synthetic missingness prevalence is consistently below the observed prevalence, suggesting that the model still underestimates the frequency of unobserved measurements and that the missingness generation process could be improved. Finally, we note that these results should be interpreted as preliminary evidence that the model captures important aspects of the coupled physiology–sampling process, rather than as a complete validation of downstream clinical utility. In the extended work, we evaluate how the resulting synthetic data can be used to train Prior-Data Fitted Networks for imputation and forecasting under distribution shifts, while explicitly leveraging informative missingness.

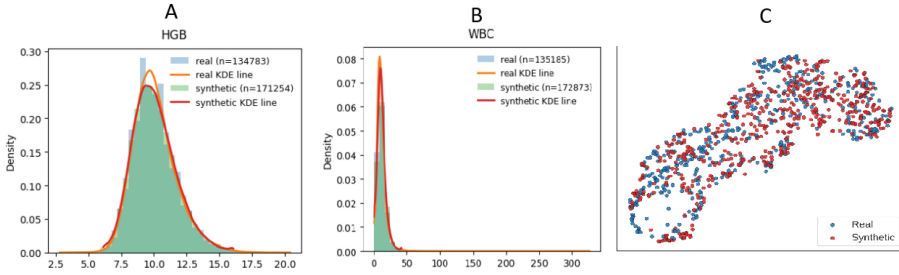


Fig. 1. Real vs. synthetic comparisons: (A) Hemoglobin and (B) White Blood Cells count univariate densities; (C) UMAP embedding of windows using all 26 channels (labs + masks), computed on a random subsample of 500 real and 500 synthetic windows.

4 Conclusions

This preliminary work shows how TimeDiff can be adapted to generate laboratory trajectories while jointly modeling informative missingness on a public MIMIC-III-derived benchmark. With practical preprocessing, the model produces synthetic windows that closely match real data in marginal lab distributions and in the joint structure of values and missingness masks. Our results suggest that diffusion models are a promising foundation for realistic synthetic lab generation under MNAR-like missingness, and they motivate future work that leverages the learned generative prior for downstream inference and forecasting in foundation models.

Acknowledgment. Hadi Mehdizavareh and Arijit Khan acknowledge support from the Novo Nordisk Foundation grant NNF 22OC0072415.

References

1. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS), pp. 6840–6851. Curran Associates Inc, Red Hook, NY (2020)
2. Johnson, A.E.W., et al.: MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016)
3. Kazijevs, M., Samad, M.D.: Deep imputation of missing values in time series health data: a review with benchmarking. *J. Biomed. Inform.* **144**, 104440 (2023)
4. Li, J., Yan, et al.: Imputation of missing values for electronic health record laboratory data. *NPJ Digit. Med.*, **4**(1), 14 (2021)
5. Luo, Y.: Evaluating the state of the art in missing data imputation for clinical data. *Briefings Bioinform.*, **23**(1), bbab489 (2022)
6. Tan, A.L.M., et al.: Informative missingness: what can we learn from patterns in missing laboratory data in the electronic health record? *J. Biomed. Inform.* **139**, 104306 (2023)

7. Tian, M., et al.: Reliable generation of privacy-preserving synthetic electronic health record time series via diffusion models. *JAMIA Open* **5**(4), ooac229 (2022)
8. Xiao, C., Choi, E., Sun, J.: Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.* **25**(10), 1419–1428 (2018)
9. Yang, Y., et al.: A survey on diffusion models for time series and spatio-temporal data. *ACM Comput. Surv.*, **58**(8), 196:1–196:39 (2026)
10. Cichosz, S.L., et al.: Biases in glucose metrics are directly related to low coverage of continuous glucose monitoring: insights from diverse populations. *Diabetes Tech. Ther.* **28**(2), 180–184 (2026)
11. Farahani, M.A., et al.: Time-series forecasting in smart manufacturing systems: an experimental evaluation of the state-of-the-art algorithms. *Robot. Comput. Int. Manuf.* **95**, 103010 (2025)
12. Müller, S.: Transformers can do Bayesian inference. In: *International Conference on Learning Representations (ICLR)* (2022)



Logic-Guided Interpretable Hospital Readmission Risk Modeling Using Italian Administrative Healthcare Data

Marina Andrić^{1(✉)}, Gianluca Apriceno², Kevin Gelmini³, and Mauro Dragoni²

¹ Free University of Bozen-Bolzano, Bolzano, Italy
`marina.andric@unibz.it`

² Digital Health & Wellbeing Research Center, Fondazione Bruno Kessler, Trento, Italy

`{apriceno,dragoni}@fbk.eu`

³ University of Trento, Trento, Italy
`kevin.gelmini@studenti.unitn.it`

Abstract. Hospital readmissions are a major burden on healthcare systems and result from complex interactions among heterogeneous clinical factors. While machine learning can model such interactions, interpretability is essential for clinical use. We propose a logic-guided learning framework for readmission prediction based on Logic Tensor Networks (LTN), integrating first-order logical rules into supervised learning using Italian administrative healthcare data. The rules encode clinically motivated relationships among chronic conditions, prior utilization, and patient stability. Training combines classification loss with rule-satisfaction constraints, enabling interpretable predicate representations. Results across cross-validation runs show competitive performance compared to baseline models, while the learned predicates provide clinically meaningful and interpretable structure for readmission risk.

Keywords: Hospital readmission · Administrative healthcare data · Neuro-symbolic AI · Explainability

1 Introduction

A hospital readmission is a new admission within a fixed period (commonly 30 days) after discharge and is widely used as a quality indicator due to its association with worse patient outcomes, higher costs, and financial penalties. In Italy, the National Health Service provides universal coverage through a regional structure; however, reductions in inpatient capacity have not been matched by equivalent strengthening of community care, despite an ageing population and rising chronic disease burden [5]. This study focuses on 30-day readmissions for heart failure (HF), ischemic heart disease (IHD), chronic obstructive pulmonary

GitHub: <https://github.com/IDA-FBK/Hospital-Readmission-Prediction>.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
P. Andreev et al. (Eds.): AIME 2026, LNAI 16749, pp. 204–209, 2027.
https://doi.org/10.1007/978-3-032-30813-9_38

disease (COPD), pneumonia (PNA), and stroke (S), using pseudo-anonymized Italian administrative data [1]. Readmission prediction has been studied using statistical and machine learning methods, often relying on interpretable models such as logistic regression and decision trees, but these approaches struggle to integrate structured clinical knowledge while remaining interpretable. Neuro-symbolic AI addresses this by combining neural learning with symbolic reasoning. Logic Tensor Networks (LTN) [3, 6] allow first-order logic constraints to be embedded in learning and have been applied to event recognition [4], energy prediction [2], and diabetes prediction [13]. We extend LTN-based logic-guided learning to readmission prediction by encoding clinically motivated relationships between chronic conditions, prior hospitalizations, and patient stability as logical constraints within supervised learning. This promotes knowledge-consistent predictions while preserving neural flexibility. Results show competitive performance with standard baselines and improved interpretability through clinically meaningful predicate structures.

2 Dataset

2.1 Data Source and Cohort Construction

This study used publicly available administrative healthcare records collected between Jan 2019 and Apr 2024 by a local healthcare authority in Piedmont, including hospital discharge records, outpatient, emergency and medication data [1]. Patients with HF, IHD, and COPD were identified following Gini [9], while PNA and S were identified using Cascini [7] and Hyeraci [10], based on ICD-9-CM codes and medication prescriptions. We constructed a hospitalization-level dataset of all inpatient admissions for eligible patients between Jan 2019 and Apr 2024. Episodes were preprocessed to merge overlapping or consecutive stays and ensure completeness. Each admission was treated as a potential *index stay* eligible for 30-day follow-up. We excluded hospitalizations ending after Mar 1, 2024 to ensure complete follow-up, as well as cases lacking ≥ 12 months of prior history or post-discharge records. Additional exclusions included day-hospital admissions, non-relevant departments (e.g., internal medicine, cardiology, pulmonology, neurology), and hospitalizations without target conditions. The outcome was defined as a binary indicator of all-cause readmission within 30 days of discharge. The final cohort included 5,590 hospitalizations from 3,678 patients: HF (1,010), IHD (1,675), COPD (598), PNA (1,234), and S (291). The overall readmission rate was 7.71% (hospitalization-level) and 9.41% (patient-level), with the highest rates for HF (10.17%; 12.97%) and lowest for PNA (7.00%; 7.94%), with intermediate values for IHD, COPD, and S. These results are broadly consistent with prior Italian studies: S and IHD align with [12], while HF and PNA are slightly lower than [14], likely due to cohort and data differences.

2.2 Features Selection

We characterized clinical complexity and prior healthcare utilization using literature-informed features derived from data available up to the end of the

index hospitalization. Multimorbidity is a known predictor of readmission in Italian cohorts [15]. Comorbidity burden was measured as the number of Charlson conditions identified from ICD-9 codes in the preceding 12 months (median 1, IQR 0–2). Medication burden at discharge was defined as the number of distinct ATC-coded drugs (median 0, IQR 0–5), excluding vitamins, minerals, and dermatological agents, in line with evidence linking polypharmacy to readmission risk in older populations [8]. Prior healthcare utilization was captured via the number acute inpatient admissions in the previous 12 months (median 0, IQR 0–1) and emergency department visits in the 30 days before admission (median 0, IQR 0–1) [12]. Additional binary indicators included antidiabetic, antihypertensive, and diuretic use, as well as prior hospitalizations for renal disease, S, and coronary heart disease. Index admissions occurred mainly in general medicine (41.68%) and cardiovascular (37.6%), followed by rehabilitation and frailty (9.64%), neurovascular (3.43%), high-acuity care (3.26%), renal and metabolic (2.08%) and other (2.31%). Length of stay was excluded due to inconsistent evidence regarding its association with readmission risk [8].

3 LTN for Hospital Readmission Prediction

Clinically motivated risk factors for hospital readmission were encoded using atomic and composite predicates. Atomic predicates represent individual clinical attributes, including diagnoses, medication exposure, and prior healthcare utilization, defined as either binary indicators (e.g., disease presence) or fuzzy variables (e.g., comorbidity and medication burden), with thresholds learned within the LTN framework (Sect. 4). Composite predicates capture higher-level risk profiles such as low recent healthcare utilization, clinical stability, cardiopulmonary disease, and cardiometabolic risk, and are used to structure clinically informed relationships for learning. **Risk-reducing (RR)** rules encode protective patterns, e.g., a stable profile with low comorbidity, low medication burden, and no recent utilization implies reduced risk, $\text{StableProfile}(x) \rightarrow \neg \text{Readmit}(x)$. **Risk-enhancing (RE)** rules capture established risk patterns, such as multiple prior admissions, $\text{PriorInpatient}_{\geq 2}(x) \rightarrow \text{Readmit}(x)$, high comorbidity with recent emergency visits, $\text{HighComorb}(x) \wedge \text{PriorED}_{\geq 1}(x) \rightarrow \text{Readmit}(x)$, or heart failure with high medication burden, $\text{HF}(x) \wedge \text{HighMeds}(x) \rightarrow \text{Readmit}(x)$. **Consistency constraints (CC)** enforce logical coherence, ensuring mutual exclusivity between burden levels, $\text{HighMeds}(x) \rightarrow \neg \text{LowMeds}(x)$ and $\text{HighComorb}(x) \rightarrow \neg \text{LowComorb}(x)$.

4 Experimental Setting

The dataset was split at the patient level into training (80%) and test (20%) sets, stratified by the presence of at least one 30-day readmission. The training set included 2,941 patients (H: 4,471 hospitalizations), and the test set 737 patients (H: 1,119 hospitalizations). Decision thresholds were selected via 5-fold cross-validation on the training set, optimizing F_1 -score under a minimum

recall constraint of 0.60. Final evaluation was performed on a held-out test set, averaged over five random seeds. LTN predicates use thresholded clinical features calibrated via grid search over clinically meaningful ranges under the same recall constraint. Configurations with degenerate or uninformative predicates were excluded based on rule-antecedent prevalence. The final stable setting used $\tau_{\text{med}}^{\text{high}} = 4$, $\tau_{\text{med}}^{\text{low}} = 2$, $\tau_{\text{com}}^{\text{high}} = 2$, $\tau_{\text{com}}^{\text{low}} = 0$, $k = 2$ for $\text{PriorInpatient}_{\geq k}$, and $m = 1$ for $\text{PriorED}_{\geq m}$. We jointly optimize a supervised loss and logic regularization, $\mathcal{L} = \mathcal{L}_{\text{sup}} + \beta \mathcal{L}_{\text{logic}}$. A neural network baseline (NN) is obtained by setting $\beta = 0$ while keeping architecture, preprocessing, optimizer (Adam), learning rate (10^{-3}), batch size, and early stopping fixed. The logic weight β was tuned in $[0, 0.3]$ via cross-validation under the same recall constraint. Candidates were pre-screened with short runs and fully trained for 60 epochs with early stopping (patience = 6, $\Delta = 5 \times 10^{-4}$), using a linear warm-up over the first 10 epochs. The final value $\beta = 0.30$ was selected based on PR-AUC across five seeds. We compare against logistic regression (LR), random forest (RF), and multilayer perceptron (MLP) [11], using identical feature sets from administrative data. Performance is evaluated on the held-out test set using accuracy, precision, recall, F₁-score, ROC-AUC, and PR-AUC.

5 Results

Table 1 reports predictive performance for 30-day hospital readmission. The logic-guided LTN achieves the best overall trade-off, outperforming the NN baseline and other models in F1-score (0.198) and precision (0.117), while maintaining competitive recall (0.651). RF yields the highest recall (0.702) but with substantially lower precision, reflecting a higher false positive rate. Overall, the LTN shifts the operating point toward better balance between precision and recall. AUC-ROC is comparable between LTN and NN (0.678 vs. 0.679), indicating similar ranking ability, while PR-AUC is slightly higher for LTN (0.145), consistent with improved performance under class imbalance.

Table 1. Predictive performance on 30-day readmission (mean $\pm$ std).

Model	Acc	Prec	Rec	F1	AUC-ROC	PR-AUC
LTN	0.612$\pm$0.024	0.117$\pm$0.004	0.651 $\pm$ 0.023	0.198$\pm$0.005	0.678 $\pm$ 0.006	0.145$\pm$0.005
NN	0.605 $\pm$ 0.030	0.114 $\pm$ 0.006	0.641 $\pm$ 0.024	0.193 $\pm$ 0.008	0.679$\pm$0.006	0.144 $\pm$ 0.004
MLP	0.594 $\pm$ 0.023	0.109 $\pm$ 0.009	0.634 $\pm$ 0.059	0.186 $\pm$ 0.015	0.659 $\pm$ 0.037	0.133 $\pm$ 0.015
LR	0.594 $\pm$ 0.006	0.105 $\pm$ 0.002	0.605 $\pm$ 0.021	0.179 $\pm$ 0.004	0.652 $\pm$ 0.008	0.130 $\pm$ 0.003
RF	0.503 $\pm$ 0.028	0.098 $\pm$ 0.005	0.702$\pm$0.017	0.172 $\pm$ 0.008	0.650 $\pm$ 0.002	0.140 $\pm$ 0.006

From an interpretability perspective, learned logical constraints are largely consistent with observed data. CC are almost perfectly satisfied with negligible

violations. RE rules based on prior admissions, emergency visits, and comorbidity burden show high satisfaction (> 0.90), confirming recent healthcare utilization as a strong risk signal. HF-related rules exhibit similarly stable behavior across comorbidity, medication burden, and cardiometabolic interactions. The most frequently activated rule, involving cardiopulmonary patients with low burden but recent emergency use, shows lower satisfaction and higher violation rates, suggesting greater clinical heterogeneity in this subgroup. In contrast, the risk-reducing rule is less frequently triggered but strongly associated with non-readmission, consistent with its protective role.

6 Conclusions

Predicting hospital readmission is challenging due to heterogeneous factors and class imbalance. The proposed approach achieves performance comparable to administrative-data models, with F1-score and PR-AUC within reported ranges, while providing an interpretable representation of clinical knowledge. Atomic and composite predicates encode clinically meaningful factors (e.g., multimorbidity and cardiopulmonary disease), enabling transparent mapping from inputs to risk. Limitations include the absence of sociodemographic variables and the lack of distinction between planned and unplanned readmissions, partially mitigated by restricting to ordinary inpatient stays. Generalizability may be constrained by single-region data and the absence of clinical measurements. Predicate thresholds were clinically plausible and robust, while logical rules are heuristic and literature-driven, requiring clinician validation. Overall, the framework balances competitive performance with interpretability. Future work will focus on external validation, temporal modeling, and clinician-in-the-loop refinement.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Andrić, M., et al.: Healthcare utilization data from piedmont, Italy: an Administrative dataset of care episodes (2019–2024). *Data Brief* **63** (2025)
2. Apriceno, G.: End-to-End Neuro-Symbolic Approaches for Event Recognition (2023)
3. Badreddine, S.: Logic tensor networks. *Artif. Intell.* **303** (2022)
4. Badreddine, S., et al.: Interval logic tensor networks. *CoRR* **abs/2303.17892** (2023)
5. de Belvis, A., et al.: Italy: Health System Review. *Health Systems in Transition* (2022), review of Italy’s regionalized National Health Service and performance
6. de Campos Souza, P.V., et al.: Enhancing logical tensor networks: integrating uninorm-based fuzzy operators for complex reasoning. In: *Proceedings of the Neural-Symbolic Learning and Reasoning - 18th International Conference. NeSy 2024, Barcelona, Spain, September 9–12, 2024, Proceedings, Part II*, pp. 68–79. Springer, *Lecture Notes in Computer Science* (2024)

7. Cascini, S., et al.: Pneumonia burden in elderly patients: a classification algorithm using administrative data. *BMC Infectious Diseases* **13**(1) (2013)
8. Franchi, C.: Risk factors for hospital readmission of elderly Patients. *Eur. J. Intern. Med.* **24**(1) (2013)
9. Gini, R., et al.: Chronic disease prevalence from italian administrative databases in the valore project: a validation through comparison of population estimates with general practice databases and national survey. *BMC Public Health* **13**(1) (2013)
10. Hyeraci, G., et al.: A systematic review of case-identification algorithms based on Italian healthcare administrative databases for three relevant diseases of the cardiovascular system: acute myocardial infarction, ischemic heart disease, and stroke. *Epidemiologia Prevenzione* **43**(4S2) (2019)
11. Maisonnave, M.: Explainable machine learning to identify risk factors for unplanned hospital readmissions in nova Scotian hospitals. *Comput. Biol. Med.* **190** (2025)
12. Mercurio, G.: A novel risk score predicting 30-day hospital re-admission of patients with acute stroke by machine learning model. *Eur. J. Neurol.* **31**(3) (2023)
13. Mondal, S.: A logic tensor network-based neurosymbolic framework for explainable diabetes prediction. *Appl. Sci.* **15**(21) (2025)
14. Tellini, M.: The revolving door syndrome in internal medicine: a study on 11,846 subjects discharged from all internal medicine departments of tuscany with diagnosis of heart failure and pneumonia. *Res. Dev. Med. Med. Sci.* **6**, 163–175. B P International (2023)
15. Trevisan, C.: Hospital length of stay and 30-day readmissions in older people: their association in a 20-year cohort study in Italy. *BMC Geriatr.* **23**(1) (2023)



DST-GAN: Dempster-Shafer Based Fusion for Multimodal Oversampling of Healthcare Data

Kevin Fee¹(✉) , Mohammed Hasanuzzaman¹ , Suneil Jain³ ,
Ross G. Murphy² , and Anna Jurek-Loughrey¹

¹ School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, Ulster, Northern Ireland, UK

kfee04@qub.ac.uk

² The Centre for Genomic Medicine, Ulster University, Coleraine, Ulster, Northern Ireland, UK

³ Johnston Cancer Research Centre, Queen's University Belfast, Belfast, Ulster, Northern Ireland, UK

Abstract. Multimodal healthcare data often have small population sizes and are imbalanced both within modalities (class imbalance) and across modalities (dimensionality imbalance), leading to biased fusion where minority classes are underrepresented and high-dimensional modalities dominate. We propose a Generative Adversarial Network (GAN) with a differentiable component based on Dempster-Shafer Theory (DST). Our DST-GAN is a DST-based multimodal generative framework that fuses modalities to form an equitable joint representation and generates high-quality unimodal samples from it. On five real-world datasets, DST-GAN improves unimodal and multimodal fusion classification over four strong baselines, demonstrating more balanced modality contributions as well as better and more reliable downstream performance.

Keywords: Multimodal · Oversampling · Data Fusion · Uncertainty

1 Introduction

We address two key challenges in small-sample biomedical datasets: (1) **intra-modality class imbalance**, where minority disease samples are underrepresented [1]; and (2) **inter-modality dimensionality imbalance**, where high-dimensional data dominates fused representations [2]. Clinical data typically has far fewer features than imaging or omics data, making truly multimodal fusion difficult. Standard oversampling [3] fails due to synthetic modality misalignment [4], missing modalities, and cross-modal structural differences [5]. We propose a DST-based oversampling uncertainty framework for equitable multimodal fusion that preserves cross-modal correlations and prevents low-dimensional (e.g., clinical) modalities from being overshadowed. Our contributions are: **DST-guided**

multimodal fusion - DST-based reweighting to prevent high-dimensional dominance; **Uncertainty-aware generation** - generating high-fidelity unimodal samples from the learned joint representation.

2 Related Work

To address intra-modality class imbalance, deep generative methods (DGMs) such as GANs [6] can oversample a class-imbalanced dataset. While GANs generate high-fidelity unimodal synthetic samples, they struggle in the multimodal setting due to inter-modality dimensionality imbalance. Multimodal classifiers can be used instead. Notable examples include MMPareto [2], which adjusts fusion and unimodal gradients via Pareto optimization; ReconBoost [7], which disentangles modality-specific gradients; and methods that re-balance modality contributions during training [8]. Multimodal data fusion can be used to address inter-modality dimensionality imbalance and is typically subdivided into early (concatenation), intermediate (encode then fuse), or late (aggregate predictions) fusion. Uncertainty modeling, such as Embracing Aleatoric Fusion (EAU) [4] weights modalities by their estimated noise, down-weighting unreliable ones while prioritizing less noisy signals. However, this approach relies on Bayesian fusion, which requires fixed priors and can be computationally heavy.

3 Methods

3.1 Preliminaries

To enable robust multimodal fusion under class-imbalanced modalities, we adopt a differentiable DST. It represents aleatoric uncertainty across modalities, ensuring all modalities contribute to fusion without fixed priors [9]. Our implementation assigns each modality a Wasserstein distance (real vs. synthetic distribution; smaller distance = lower uncertainty). Dempster’s rule fuses these distances, and the pignistic transform [9] produces weights that down-weight uncertain or conflicting modalities, preventing high-dimensional dominance.

3.2 Architecture Overview

DST-GAN applies DST in both the generator and discriminator of a GAN: the generator fuses modality encodings into an equitable joint representation $\hat{F}$ via cross-attention and a DST fusion layer, while the adversarially trained discriminator uses DST fusion to re-weight modality-specific Wasserstein-based losses. The entire architecture is illustrated in Fig. 1.

Generator: Cross-Attention and DST Fusion. We compute modality-wise cross-attention using the lowest-dimensional modality as reference to reduce complexity. Wasserstein distances W_i between each latent encoding and its cross-attended version are mapped to DST masses m_i and fused via Dempster’s rule, then projected back to the input dimensionality. The generators objective is

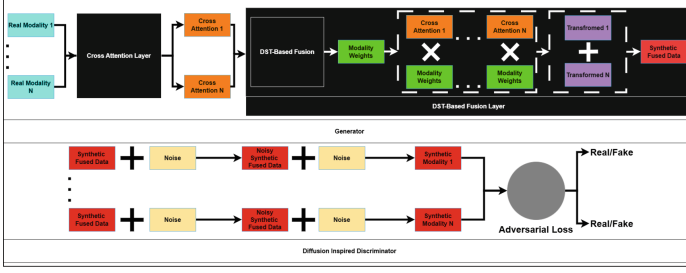


Fig. 1. DST-GAN architecture with DST fusion, cross-attention, and a diffusion-inspired discriminator.

the Wasserstein distance between synthetic $\hat{F}$ and real concatenated features: $L_G = W(\hat{F}, \text{concat}(M_1, \dots, M_N))$.

Discriminator: Diffusion-Inspired. Our discriminator injects modality-specific noise(uniform vs. Gaussian) to stabilize training. It includes a head per modality. The synthetic unimodal Wasserstein distances are aggregated via DST fusion to reduce conflicting evidence. As in the generator loss, each unimodal representation’s Wasserstein distance is combined via DST fusion, supporting lower epistemic (model) uncertainty:

$$L_D = \text{DSTFusion}(W(\hat{M}_1, M_1), \dots, W(\hat{M}_N, M_N)).$$

4 Results

4.1 Datasets

We include five real-world datasets (four biomedical, one human activity recognition(HAR)). Each biomedical dataset contains at least two of Clinical(C), Radiomics(R) or Genomics(G) modalities. Our biomedical datasets are listed as Dataset Name (Number of Samples, Modality/Number of Features, class imbalance): Prostate-FASTMAN (184,C/6,R/154,G/417 DEGs; 0.10), Prostate-PRAD (328, C/38, G/650 DEGs; 0.18), Breast-BCBM (91, R/80, G/220 DEGs; 0.03), Breast-BRCA (453,C/9,G/164;0.49). HAR-MEx contains 30 subjects, ~ 6300 points [10] with three modalities, Accelerometer on thigh and wrist (ACT and ACW) and pressure mat(PM) with 180 features each and a class imbalance of 0.18. All data sets exhibit intra-modality class imbalance. HAR-MEx isolates this effect without dimensionality imbalance (equal feature dimensions); Breast-BRCA tests the converse (mild class imbalance with dimension disparity).

4.2 Experimental Settings

To mitigate optimistic evaluation, we use patient-level leave-one-out cross-validation (LOOCV) and external validation on Prostate datasets with similar modalities. We report $F1$ (precision-recall trade-off under imbalance) and the Matthews correlation coefficient (MCC; class separability from the full confusion matrix).

Evaluation of Synthetic Data Quality. We compute multiple complementary metrics of synthetic data quality: Wasserstein distance for distributional fidelity; average Manhattan distance between samples before and after oversampling for realism; and Cross-Correlation Analysis (CCA) for multimodal alignment.

Table 1. Biomedical Classification Results with early fusion (EF), EAU Fusion, MMPareto, ReconBoost, and DST-GAN. Scores in bold illustrate the best-performing model for each modality combination.

Dataset	Input	Output	EF-GAN		EAU-GAN (2024)		MMPareto (2024)		ReconBoost (2024)		DST-GAN	
			F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC
Prostate-Fastman	R+G	R	0.90	0.83	0.98	0.97	0.00	0.00	0.47	0.00	0.98	0.97
		G	0.98	0.97	0.98	0.97	0.00	0.00	0.08	0.00	0.98	0.97
		F	0.98	0.97	0.98	0.97	0.05	-0.62	0.40	0.04	0.98	0.97
	C+G	C	0.22	-0.003	0.26	0.15	0.08	0.00	0.47	0.00	0.86	0.76
		G	0.97	0.94	0.97	0.94	0.47	0.00	0.08	0.00	0.98	0.97
		F	0.94	0.89	0.95	0.91	0.05	-0.62	0.45	0.12	0.98	0.97
	C+R	C	0.00	0.00	0.26	0.15	0.00	0.00	0.47	0.00	0.85	0.72
		R	0.97	0.94	0.97	0.94	0.47	-0.04	0.08	0.00	0.98	0.97
		F	0.97	0.94	0.97	0.94	0.46	-0.07	0.38	-0.01	0.98	0.97
	C+R+G	C	0.18	-0.02	0.21	0.11	0.23	-0.26	0.08	0.00	0.87	0.77
		R	0.97	0.94	0.97	0.94	0.47	0.00	0.08	0.00	0.98	0.97
		G	0.97	0.94	0.97	0.94	0.47	0.00	0.47	0.00	0.98	0.97
		F	0.97	0.94	0.97	0.94	0.46	-0.05	0.42	0.11	0.98	0.97
Prostate-PRAD	C+G	C	0.84	0.74	0.69	0.52	0.45	-0.06	0.45	0.00	0.75	0.59
		G	1.00	1.00	0.74	0.57	0.42	-0.14	0.13	0.00	0.85	0.74
		F	1.00	1.00	0.68	0.48	0.45	-0.06	0.43	0.002	0.95	0.90
Breast-BCBM	R+G	R	0.24	0.10	0.88	0.49	0.00	0.00	0.49	0.06	1	1
		G	1	1	0.48	-0.01	0.00	0.00	0.48	-0.01	1	1
		F	0.93	0.86	0.57	0.17	0.57	0.17	0.57	0.17	0.93	0.86
Breast-BRCA	C+G	C	0.30	-0.31	0.31	-0.37	0.86	0.73	0.40	0.00	0.37	-0.24
		G	0.19	-0.60	0.20	-0.53	0.37	-0.18	0.40	0.00	0.21	-0.57
		F	0.30	-0.31	0.18	-0.59	0.89	0.79	0.45	-0.0	0.62	0.24
HAR-MEx	ACT+ACW	ACT	0.50	0.39	0.63	0.46	0.00	0.00	0.45	0.25	0.64	0.47
		ACW	0.75	0.42	0.65	0.38	0.00	0.00	0.13	0.10	0.52	0.34
		F	0.14	0.08	0.69	0.52	0.15	0.13	0.24	0.00	0.84	0.82
	ACT+PM	ACT	0.33	0.30	0.00	0.00	0.00	0.00	0.45	0.25	0.59	0.43
		PM	0.70	0.44	0.65	0.42	0.00	0.00	0.13	0.10	0.59	0.43
		F	0.59	0.44	0.63	0.48	0.15	0.13	0.30	0.19	0.68	0.49
	ACW+PM	ACW	0.67	0.38	0.69	0.40	0.29	0.17	0.45	0.25	0.68	0.39
		PM	0.54	0.32	0.56	0.43	0.16	0.11	0.13	0.10	0.58	0.44
		F	0.59	0.47	0.65	0.46	0.25	0.08	0.31	0.10	0.66	0.48
	ACT+ACW+PM	ACT	0.51	0.35	0.00	0.00	0.00	0.00	0.45	0.25	0.81	0.44
		ACW	0.65	0.38	0.00	0.00	0.00	0.00	0.31	0.12	0.79	0.43
		PM	0.40	0.21	0.00	0.00	0.00	0.00	0.13	0.08	0.77	0.46
		F	0.13	0.05	0.60	0.20	0.05	0.00	0.30	0.08	0.65	0.38

4.3 Classification Results

We use patient-level LOOCV with differential expression on genomics training splits. DST-GAN is compared against a GAN using early fusion (EF) and EAU fusion instead of DST, and against MMPareto/ReconBoost as imbalance classifiers. In Table 1, EF-GAN and EAU-GAN suffer mode collapse (high W_i , low CCA, low diversity), while DST-GAN achieves the best trade-off (low W_i , high CCA, diverse samples). DST-GAN improves F1/MCC on 4/5 datasets (19/22 configurations, 86% vs. EF/EAU), leads MMPareto/ReconBoost in 19/22 cases (86%), and is best or tied on HAR in 11/13 (85%). Performance is weaker on Breast-BRCA, where near class balance (0.49) limits DST-GAN’s uncertainty advantage. Perfect scores on Breast-BCBM (n=91, imbalance 0.03) are supported by perfect MCC, patient-level LOOCV, and cross-dataset Prostate validation, confirming generalisability. Wilcoxon signed-rank tests confirm DST-GAN significantly improves multimodal (biomedical: $p = 0.01$; HAR: $p = 0.01$) and unimodal (biomedical: $p = 0.004$; HAR: $p = 0.04$) classification.

Ablation Study. We ablate the full DST-GAN by removing one component at a time (cross-attention, DST-fusion, diffusion-inspired discriminator). Removing any component either decreases performance or matches the best-performing ablation, confirming each component’s contribution.

5 Conclusion

DST-GAN addresses multimodal imbalance via DST-guided fusion, cross-modal attention, and a diffusion-inspired discriminator. It reduces high-dimensional dominance and achieves consistent gains across five real-world datasets under small, imbalanced settings. Limitations include robustness to missing modalities and GAN training challenges (hyperparameter sensitivity, mode collapse). Future work will include PR-AUC for class-imbalanced evaluation. Clinically, DST-GAN improves minority-class generation and supports more reliable multimodal AI decision-making.

References

1. Fee, K.: Towards a biological evaluation framework for oversampling (BEFO) gene expression data. *J. Biomed. Inform.* (2025)
2. Wei, Y., Hu, D.: MMPareto: boosting multimodal learning with innocent unimodal assistance. In: *Proceedings of the 41st ICML* (2024)
3. Mingyang, D.: An oversampling method for multi-class imbalanced data based on composite weights. *PLoS ONE* (2021)
4. Gao, Z.: Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In: *Proceedings of the 2024 IEEE/CVF Conference on CVPR* (2024)
5. Zhang, L.: EFVAE: efficient federated variational autoencoder for collaborative filtering. In: *Proceedings of the 33rd ACM CIKM* (2024)
6. Goodfellow, I., et al.: Generative adversarial nets. In: *Proceedings of the Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc. (2014)

7. Cong, H.: Reconboost: boosting can achieve modality reconciliation. In: Proceedings of the 41st International Conference on Machine Learning (2024)
8. Qingyuan, J.: Rebalanced multimodal learning with data-aware unimodal sampling. arXiv (2025)
9. Fei, L., et al.: A novel method to determine basic probability assignment in Dempster–Shafer theory and its application in multi-sensor information fusion. *Int. J. Distrib. Sens. Netw.* (2019)
10. Wijekoon, A., Wiratunga, N., Cooper, K.: MEx: Multi-modal Exercises Dataset for Human Activity Recognition (2019)



A Structured Residual Refinement Transformer Framework for Long-Horizon Healthcare Demand Forecasting with Causal and Seasonal Signals

Guillem Hernández Guillaumet^{1,2} , Joan Samper Vila¹ ,
and Beatriz López Ibáñez¹ 

¹ eXiT Research Group, Universitat de Girona (UdG), 17003 Girona, Catalunya, Spain

² Assistance strategy management, Hospital Germans Trias i Pujol (ICS), 08916 Badalona, Catalunya, Spain
guillemhg98@gmail.com

Abstract. Reliable long-horizon forecasting of diagnosis-specific healthcare demand is essential for resource planning, yet remains difficult due to non-stationarity and seasonal effects. We propose a contextual residual Transformer framework that separates intrinsic demand dynamics from cross-diagnostic and calendar-driven corrections through staged residual refinement. A univariate Transformer is initially used to model the target demand time series; subsequent residual stages correct its forecasts using causally selected diagnostic covariates and future-known seasonal features. Diagnostic predictors are selected with the CCLR-DL framework, which combines lagged regression and Granger-causality testing to identify temporally meaningful dependencies. We evaluate the method on large-scale Catalan primary-care data comprising 6,575 daily observations from 2007–2025, and over one billion clinical encounters. Results across horizons up to 365 d show that causally guided residual refinement improves accuracy and long-horizon stability while preserving interpretable diagnostic and seasonal correction mechanisms.

Keywords: Healthcare demand forecasting · Transformers · Residual learning · Causal inference · Seasonality

1 Introduction and Related Work

Accurate diagnosis-specific forecasting of healthcare demand is essential for clinical decision support, staffing, scheduling, and capacity planning. However, reliable long-horizon forecasting remains difficult because healthcare activity is shaped by non-stationarity, cross-diagnostic interactions, seasonal effects, and population-level changes. In this setting, forecasts must remain accurate over horizons ranging from days to months while remaining interpretable for operational decision-making.

Neural forecasting models can capture nonlinear temporal dynamics and multivariate dependencies. Recurrent and convolutional-recurrent approaches such as LSTNet [1] model both short- and long-term temporal patterns, while Transformer-based models use self-attention to connect distant time steps directly. Efficient variants such as Informer [6] and Log-Linear Attention [2] further address scalability for long sequences. Nevertheless, long-horizon healthcare demand cannot depend solely on the choice of model architecture; diagnostic covariates and seasonal patterns must also be integrated in a manner that maintains both robustness and interpretability.

A common strategy is to concatenate target histories, related diagnoses, and calendar variables into a single multivariate input. Although effective, this can obscure which signals correct forecast errors and may entangle heterogeneous sources of variation. Causal and contextual modelling provides a more structured alternative. In time series, Granger-causality-based analysis identifies predictors whose past values improve prediction beyond the target history [1]. The CCLR-DL framework [3] combines cross-correlation, lagged regression, and Granger-causality-based selection for healthcare demand forecasting, showing that statistically guided feature selection can improve neural models while retaining interpretability.

Motivated by this, we propose a causal-contextual residual Transformer framework for long-horizon healthcare demand forecasting. In this work, ‘causal’ refers to Granger-style temporal predictive dependence, not to clinical causality: selected covariates are interpreted as temporally informative predictors whose past values help explain future residual errors. The model first learns a univariate forecast of the target diagnosis and then applies residual corrections using selected diagnostic covariates and future-known seasonal features. This design makes three contributions: (i) a staged residual formulation that separates intrinsic demand dynamics from contextual corrections; (ii) the integration of g-causal diagnostic covariates with Transformer forecasting; and (iii) an evaluation on large-scale diagnosis-specific primary-care demand with horizons up to one year.

2 Methodology

We formulate diagnosis-specific healthcare demand forecasting as a multi-step prediction problem. Let $\{y_t\}_{t=1}^T$ denote the historical demand series for a target activity or diagnosis of length T . Given a lookback window L and a forecasting horizon H , the objective is to predict $\mathbf{y}_{t+1:t+H}$ from the recent target history $\mathbf{y}_{t-L+1:t}$ and contextual information. We differentiate between historical diagnostic covariates $\mathbf{X}_{t-L+1:t} \in \mathbb{R}^{L \times D}$, which are available only up to time t (with dimensionality D), and future-known seasonal features $\mathbf{S}_{t+1:t+H} \in \mathbb{R}^{H \times K}$, such as weekday, holiday, and calendar indicators (with dimensionality K). The forecasting task can be formulated as

$$\hat{\mathbf{y}}_{t+1:t+H} = f_{\theta}(\mathbf{y}_{t-L+1:t}, \mathbf{X}_{t-L+1:t}, \mathbf{S}_{t+1:t+H}). \quad (1)$$

Rather than learning this mapping with a single fully multivariate model, we propose a contextual residual Transformer framework. The model first estimates

a univariate forecast of the target series and then applies successive residual corrections using g-causal diagnostic features and seasonal information, as shown in Fig. 1. Let $\hat{\mathbf{y}}_{t+1:t+H}^{(0)}$ be the univariate Transformer forecast obtained from the target history alone. At refinement stage k , a residual model estimates a correction $\hat{\mathbf{r}}_{t+1:t+H}^{(k)}$ and updates the forecast as

$$\hat{\mathbf{y}}_{t+1:t+H}^{(k)} = \hat{\mathbf{y}}_{t+1:t+H}^{(k-1)} + \hat{\mathbf{r}}_{t+1:t+H}^{(k)}. \quad (2)$$

In our experiments, two refinement stages are used. The first corrects systematic errors using lagged diagnostic covariates selected by the CCLR-DL framework [3]. This selection combines lagged regression and Granger-causality-based testing to identify temporally informative predictors, but these relations should be interpreted as predictive temporal dependencies rather than proof of clinical causation. The second stage corrects remaining residual structure using future-known seasonal and calendar features. This separation avoids information leakage and supports stage-level interpretability: diagnostic covariates explain cross-diagnostic residual corrections, whereas seasonal features explain recurrent calendar-driven corrections.

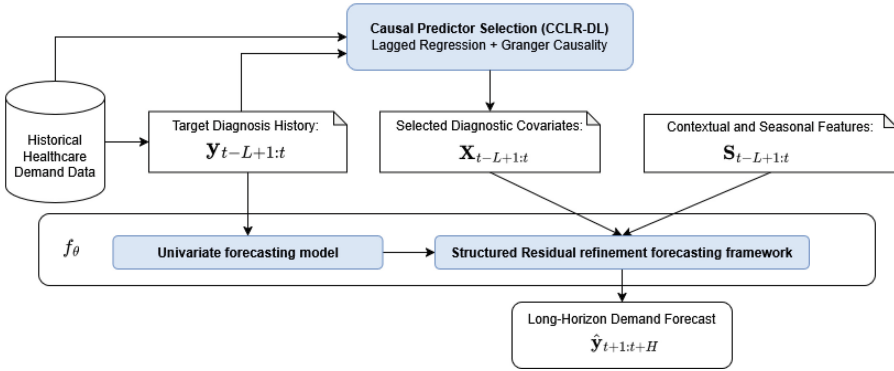


Fig. 1. Overview of the proposed causal-contextual residual forecasting framework.

All forecasting stages use the same encoder-only Transformer backbone. Inputs are sequence-normalized, projected to the model dimension, and augmented with sinusoidal positional encodings [5]. The encoder contains two Transformer blocks with eight attention heads, GELU feed-forward layers, dropout regularization, and a two-layer MLP regression head that produces horizon-specific forecasts. For longer lookback windows, temporal average pooling is applied before the regression head to reduce dimensionality. Hyperparameters for the proposed model and neural baselines were selected by grid search on the validation set using the same train-validation-test split, forecasting horizons, and search budget.

The proposed methodology is evaluated on a representative subset of large-scale healthcare activity time series from the Catalan public primary care system, comprising 2,736 activity codes over 6,575 daily time steps from 2007 to 2025 and more than 1 billion clinical encounters.

Table 1. Final-stage performance averaged over target series. Best values for each metric and horizon are shown in bold.

Model	MAE			RMSE			WAPE		
	7	60	365	7	60	365	7	60	365
Informer	6405.57	8351.45	8526.83	16582.06	19912.05	20527.66	8.83	16.78	19.45
LogFormer	4311.60	8223.74	8881.74	7854.15	19671.91	20827.18	8.50	16.34	19.82
LSTNet	3724.73	10433.90	9808.78	7035.59	20371.83	18841.44	8.17	21.13	21.45
Ours	4294.20	7533.08	8467.11	7731.90	18797.53	19961.38	8.38	16.33	18.24

We evaluate short- and long-term forecasting horizons up to 365 d using MAE, RMSE, and WAPE. MAE is used to quantify average absolute demand error, RMSE to capture sensitivity to larger deviations, and WAPE to compare relative error across diagnoses with different activity volumes. The proposed model is compared against representative neural long-horizon forecasting baselines: LSTNet [4], LogFormer [2], and Informer [6].

3 Results and Discussion

Table 1 reports final-stage performance after diagnostic and seasonal residual refinement, averaged across the target time series. LSTNet achieves the lowest error at the 7-day horizon, confirming that recurrent architectures remain competitive for short-term forecasting. However, its performance decreases at longer horizons. The proposed model obtains the best MAE and WAPE at both 60 and 365 d, and the best RMSE at 60 d, indicating stronger medium- and long-horizon performance. At 365 d, LSTNet obtains the lowest RMSE, while the proposed model remains best under MAE and WAPE, showing that conclusions depend on the operational meaning of each metric.

The residual refinement strategy improves performance mainly at longer horizons. Adding diagnostic covariates reduces MAE by 6.32% and 15.19% at 60 and 365 d, respectively, while seasonal refinement provides additional reductions of 6.54% and 4.53%. This suggests that diagnostic covariates help correct medium- and long-range structural errors, whereas seasonal features capture recurrent calendar-driven variation.

From an operational perspective, MAE and WAPE are especially relevant because they express planning error in absolute and relative terms. Lower errors can support more reliable staffing, appointment capacity, and resource allocation.

However, the current framework does not distinguish between the costs of under- and overestimation, which remains an important extension for deployment.

The framework provides stage-level interpretability. The univariate stage captures the target trajectory, the diagnostic stage identifies cross-diagnostic correction signals, and the seasonal stage isolates future-known calendar effects. These signals should be interpreted as predictive temporal associations rather than clinical causality.

An ablation against a no-causal variant confirms the value of diagnostic residual refinement. The full pipeline improves final-stage MAE by 0.88%, 9.43%, and 2.71% at 7, 60, and 365 d, with the largest gain at the medium horizon.

Several limitations remain. The comparison is restricted to neural forecasting baselines, while statistical approaches such as seasonal ARIMAX or dynamic harmonic regression should be considered in future work. In addition, gains may vary across diagnoses, with larger benefits expected for highly seasonal or high-variability series.

4 Conclusion

We proposed a contextual residual Transformer framework for long-horizon healthcare demand forecasting. The model first estimates diagnosis-specific demand from the target history and then applies structured residual corrections using temporally selected granger-causal related diagnostic covariates and future-known seasonal features. Experiments on large-scale Catalan primary-care data show improved medium- and long-horizon accuracy while providing a stage-level decomposition of diagnostic and calendar-driven corrections. These findings support staged residual refinement as a practical strategy for stable and explainable healthcare demand forecasting.

Acknowledgments. This work was supported by the Secretary of Universities and Research of the Department of Business and Knowledge, Generalitat de Catalunya, under grant 2021 SGR 01125. AI-based tools were used only for language editing; the authors take full responsibility for the scientific content.

Disclosure of Interests. The authors declare no competing interests.

References

1. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**(3), 424–438 (1969)
2. Guo, H., Yang, S., Goel, T., Xing, E.P., Dao, T., Kim, Y.: Log-linear attention (2025)
3. Hernández Guillaumet, G., López Seguí, F., Vidal Alaball, J., López, B.: CCLR-DL: a novel statistics and deep learning hybrid method for feature selection and forecasting healthcare demand. *Comput. Methods Programs Biomed.* **272**, 109057 (2025)

4. Lai, G., Chang, W.C., Yang, Y., Liu, H.: Modeling long- and short-term temporal patterns with deep neural networks (2018)
5. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008 (2017)
6. Zhou, H., et al.: Informer: beyond efficient transformer for long sequence time-series forecasting. Proc. AAAI Conf. Artif. Intell. **35**(12), 11106–11115 (2021)



A Deep Learning Framework for Vision-Based Intraocular Pressure Monitoring Using Microfluidic Contact Lenses

Justin Jacob¹(✉) , Tanzila Afrin² , Yong Jun Lai² ,
and Michael Greenspan¹

¹ Electrical and Computer Engineering, Ingenuity Labs, Kingston, Canada
{justin.j,michael.greenspan}@queensu.ca

² Mechanical and Materials Engineering, Queen's University, Kingston, Canada
{tanzila.afrin,lai}@queensu.ca

Abstract. Continuous monitoring of intraocular pressure (IOP) is important for the management of glaucoma, but current clinical measurements are limited to intermittent in-clinic assessments that fail to capture short-term fluctuations. Microfluidic contact lenses provide a passive sensing mechanism by converting pressure-induced deformation into observable fluid displacement.

We propose a system-level deep learning framework for automated interpretation of microfluidic signals for IOP monitoring. In the absence of a physical lens prototype, training data are generated by simulating microfluidic channels over eye images. The model achieved a mean IoU of 0.94 and Dice score of 0.97, making it a feasible IOP monitoring system.

Keywords: Intraocular pressure monitoring · Glaucoma · Microfluidic contact lens · Deep learning · Medical image segmentation

1 Introduction

Glaucoma is a leading cause of blindness and is strongly associated with elevated IOP [9]. Goldmann Applanation Tonometry (GAT) remains the clinical gold standard for IOP measurement despite the development of alternative techniques. These methods are limited to in-clinic measurements that are influenced by corneal thickness and biomechanics [10]. IOP exhibits short-term and diurnal fluctuations that are not captured during routine clinical visits and may be predictive of disease progression [6]. This highlights the need for frequent or continuous monitoring outside clinical settings. While existing evidence is primarily based on conventional techniques, contact lens-based approaches introduce an interface with the ocular surface, although clinical observations suggest that contact lenses do not alter IOP levels or measurements [2].

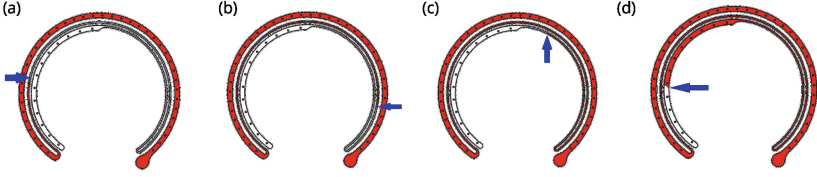


Fig. 1. Simulated microfluidic channel response under increasing IOP. The arrow indicates the extent of fluid displacement. (a) Low IOP with minimal fluid displacement in the sensing channel. (b) Moderate IOP showing partial fluid advancement along the channel. (c) Elevated IOP with substantial fluid displacement. (d) High IOP demonstrating near-maximal fluid propagation within the sensing channel.

Existing wearables use embedded electronics that result in calibration drift and power constraints, while microfluidic systems provide a passive alternative [5]. Whereas, current microfluidic approaches focus on hardware design and heuristic image processing [1,3], while learning-based interpretation remains largely unexplored.

2 Proposed System

The proposed system uses a microfluidic contact lens in which IOP variations induce measurable fluid displacement within a sensing channel. These changes are captured using passive imaging and analyzed using an automated computer vision pipeline. The users would wear the contact lens throughout the day and capture periodic images that allow longitudinal monitoring of fluid displacement under the guidance of the clinician.

Contact Lens Design and Fluid Channels. The proposed lens contains a circular sensing channel and a reservoir channel aligned with the iris. Variations in IOP induce fluid displacement within the sensing channel, providing a visual indicator for pressure estimation. The linear distribution of displaced fluid within the channel is considered to monotonically relate to IOP and is used as a proxy. The extent of fluid displacement along the sensing channel can thus be quantified and mapped to IOP through calibration with reference measurements. Figure 1 depicts the fluid position within the channel at different pressure levels.

Dataset Generation. A physical prototype is under development, and so training was performed using simulated microfluidic channels over eye images to approximate realistic imaging conditions, including variability in appearance, illumination and channel visibility (via data augmentation). This enables the model to learn fluid-like patterns under realistic varied ocular imaging conditions. The iris center and radius were estimated using a localization model to geometrically align a microfluidic channel template with the iris boundary (Fig. 2 b). Fluid displacement was simulated by progressively filling the sensing channel from a fixed seed point over a specified fill range. Eyelid masks were used to

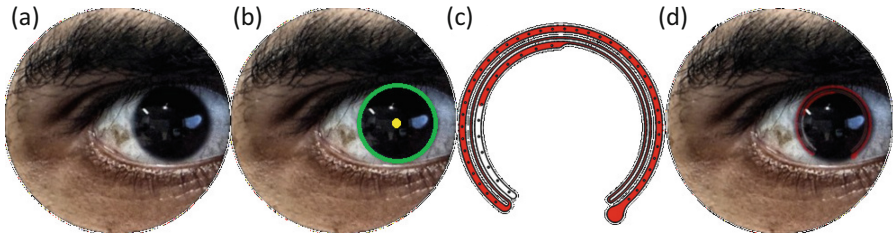


Fig. 2. Simulated microfluidic channel generation pipeline. (a) Input eye image. (b) Predicted iris center (yellow) and radius (green). (c) Enlarged version of the simulated microfluidic channel overlay used for training. (d) Warped microfluidic channel template geometrically aligned to the estimated iris boundary. (Color figure online)

suppress channel regions in occluded areas, and the rendered channel and fluid masks were saved as paired image–mask samples. This process generated 70,933 simulated image–mask pairs (Fig. 2 c,d).

Image Acquisition. Data was captured using a smartphone under controlled conditions. Test data was collected from the authors with their permission to be used for research and publication. The iris was located first in the images and then low quality images were filtered out followed by the segmentation stage.

Fluid Segmentation. Fluid displacement was estimated using a U-Net with a ResNet-18 encoder pretrained on ImageNet-1K. The model outputs binary masks of fluid regions used to estimate displacement along the sensing channel. A lightweight architecture was selected to balance segmentation accuracy and computational efficiency, providing a favorable trade-off between accuracy and computational efficiency for potential mobile deployment.

3 Experimental Setup

Datasets. The IIT Delhi Iris Database (IITD v1.0) contains grayscale iris images captured under controlled conditions, while UBIRIS.v2 provides color images acquired in more unconstrained settings. Both datasets were used to train the iris localization and eyelid segmentation models [4, 8]. A custom dataset (200 images, 30 videos) was augmented to 8,520 images using geometric and photometric transformations, then used to generate 70,933 simulated samples with varying fluid displacement levels (80:10:10 split).

Models and Training. The fluid segmentation model was a U-Net encoder–decoder network with a ResNet-18 encoder, trained to perform binary semantic segmentation of the fluid region inside the microfluidic sensing channel (fluid vs. background). Input images were resized to 512×512 . The network outputs per-pixel logits and is optimized using weighted cross-entropy to address class imbalance. Training was performed with the Adam optimizer ($\text{lr} = 10^{-3}$) for 67 epochs. The learning rate was reduced on plateau (factor = 0.5, patience = 3).

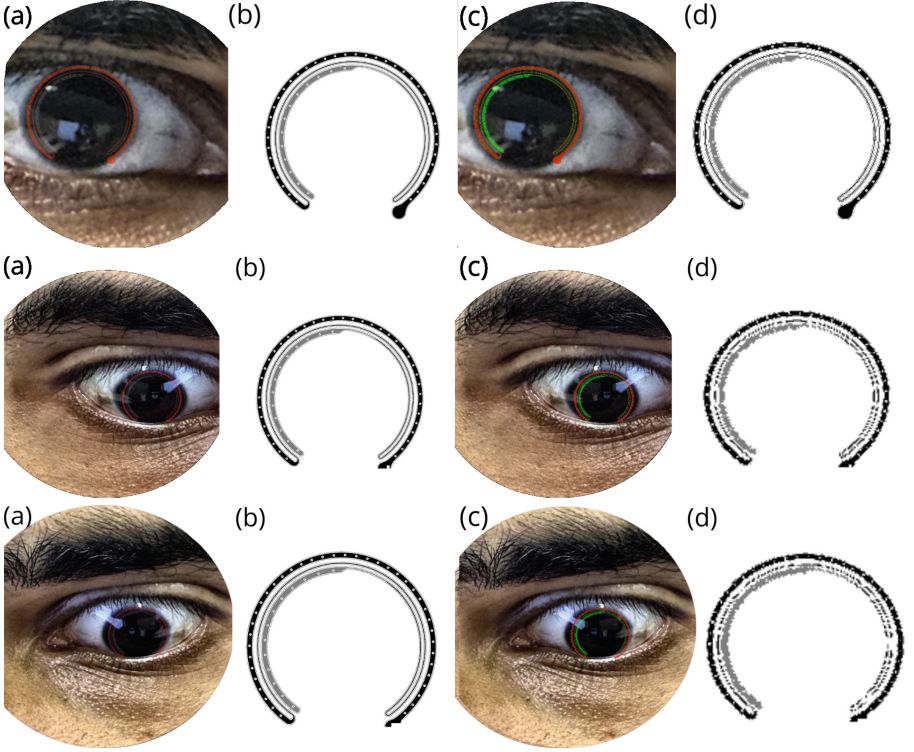


Fig. 3. Fluid segmentation results on the test set. Each row shows (a) input image, (b) ground-truth mask, (c) predicted mask overlaid on the input and (d) predicted mask. The first row shows a successful case ($\text{IoU} = 0.97$), the middle row a moderate case ($\text{IoU} = 0.91$) and the last row a challenging case ($\text{IoU} = 0.88$).

Two U-Net models were trained to segment the upper and lower eyelids which can occlude the fluid. A binary ResNet-18 classifier was trained for image quality assessment, to filter out low-quality frames, such as when a blink closes the eyelid. MediaPipe Face Mesh [7] was used for region cropping, followed by a ResNet-18 model to estimate iris center and radius. The ResNet-18 regression model predicts normalized (c_x, c_y, r) parameters, after being trained on manually annotated iris centers and radii. All ResNet-18 models used were pre-trained on ImageNet-1K. Segmentation was evaluated using Dice and IoU (on the foreground class), image quality assessment using standard classification metrics and iris localization using Mean Absolute Error and Euclidean distance.

4 Results and Discussion

The model achieved a mean IoU of 0.94 and Dice score of 0.97, indicating accurate and reliable recovery of fluid regions under realistic imaging conditions.

These segmented regions provide the basis for estimating relative fluid displacement, which serves as a proxy for IOP. Therefore, the segmentation accuracy directly influences the reliability of displacement estimation and underpins the overall effectiveness of the proposed IOP monitoring framework.

Inspection of the predicted segmentation visualizations confirmed that the model accurately captured fluid displacement patterns along the sensing channel under varying conditions, as shown in Fig. 3. Successful cases showed precise boundary alignment and smooth fluid contours, while failure cases were primarily associated with extreme reflections or near-complete eyelid occlusion. Representative visualizations are included to illustrate both typical successes and challenging scenarios. The supporting models functioned as intended, with image quality assessment filtering unusable frames, eyelid segmentation preventing unrealistic channel overlap and iris localization providing stable and accurate estimates.

Limitations. Training and evaluation are performed on simulated data, as a physical contact lens prototype is not yet available. The relationship between fluid displacement and absolute IOP remains uncalibrated, and this work focuses on segmentation rather than direct IOP estimation. While simulations provide a controlled environment for evaluating robustness, they do not fully capture all real-world variations, including distortions and optical artifacts.

5 Conclusion and Future Work

This paper presents a learning-based framework for IOP estimation using a microfluidic contact lens. The system enables passive, vision-based estimation of fluid displacement from eye images. Future work will focus on validation using data acquired from a physical microfluidic contact lens prototype, as well as subject-level generalization, long-term repeatability and establishing a quantitative relationship between fluid displacement and absolute IOP values.

Acknowledgments. This study was supported by the Canadian Institutes of Health Research (CIHR) under grant number 202409PJT-525974-CMZ-CEDA-215779.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Aravind, M., et al.: Microfluidic contact lens: fabrication approaches and applications. *Microsyst. Nanoeng.* **11**, 59 (2025)
2. Bozgöl, P., Cankaya, C., Doganay, S., et al.: The influence of soft contact lenses on the intraocular pressure measurement. *Eye* **26**(2), 278–282 (2012)
3. Choi, J., et al.: Soft, skin-mounted microfluidic systems for measuring secretory fluidic pressures. *Lab Chip* **17**, 2572–2580 (2017)
4. Kumar, A., Passi, A.: Iris matchers for personal authentication. *Pattern Recogn.* (2010)

5. Leonardi, M., et al.: A soft contact lens with a mems strain gage embedded for intraocular pressure monitoring. In: *Transducers*, pp. 1043–1046 (2003)
6. Liu, T., et al.: Intraocular pressure fluctuations in glaucoma. *Adv. Ophthalmol. Pract. Res.* (2025)
7. Lugaresi, C., et al.: Mediapipe: a framework for building perception pipelines. In: *CVPR Workshops* (2019)
8. Proença, H., et al.: Ubiris.v2: a visible wavelength iris image database. *IEEE Trans. Pattern Anal. Mach. Intell.* (2010)
9. Qi, T., et al.: Understanding glaucoma. *Klin. Monatsbl. Augenheilkd.* (2025)
10. Zeppieri, M., Gurnani, B.: Applanation tonometry. In: *StatPearls*. StatPearls Publishing (2023)



Effectiveness of Artificial Intelligence in Cardiovascular Disease Prevention: A Systematic Review

Chetan Shirvankar¹ , Lily Puterman-Salzman² , Josep Iglesias-Grau³ ,
Hervé Tchala Vignon Zomahoun⁴ , and Samira Abbasgholizadeh Rahimi¹

¹ McGill University, Montreal, QC, Canada
chetan.shirvankar@mail.mcgill.ca, samira.rahimi@mcgill.ca

² Université de Montreal, Montreal, QC, Canada

³ Montreal Heart Institute, Montreal, QC, Canada

⁴ Université Laval, Quebec City, QC, Canada

Abstract. This systematic review evaluates the effectiveness of Artificial Intelligence (AI) in preventing cardiovascular disease (CVD).

We conducted a systematic review using Joanna Briggs Institute (JBI) methodology, with the protocol registered under CRD42022369548 on PROSPERO. We included studies on all populations who receive or provide care and used AI for risk prediction and prevention of CVD. We searched seven databases from 1999 till October 2022. Two reviewers independently screened the identified records and extracted data from the included studies. We conducted a subgroup analysis on studies that compare AI to standard care. We assessed the risk of bias using a Modified IJMEDI and PROBAST tool, and data are synthesized in narrative form, considering various study aspects. This review adheres to PRISMA guidelines for systematic reviews and meta-analyses. After screening 7,505 identified records, 266 articles were included in the review. Risk prediction and early disease diagnosis were the most common prevention modalities. The subgroup analysis of 16 studies that compared these interventions with standard care included 51,687,627 participants. The Framingham Risk Score was the most used comparator, reported in 8 of the 16 studies (50%). Findings from identified studies suggest that AI, specifically Machine learning models such as Neural Networks, significantly improved the performance of clinical tasks such as risk prediction and diagnosis, suggesting that AI-based models may enhance cardiovascular disease prevention. Further high-quality prospective studies and real-world implementation research are needed before routine clinical adoption.

Keywords: Cardiovascular Disease · Systematic Review · Clinical Decision Support Systems · Artificial Intelligence · Machine Learning

1 Introduction

Cardiovascular diseases (CVDs) are the leading cause of death and the primary cause of disability worldwide [1]. In 2022, the Global Burden of Diseases Study reported 19.7 million deaths due to CVD [2]. These trends emphasize the urgent need for enhanced

cardiovascular prevention and care globally. In recent decades, significant risk factors for CVD have been identified, and practical and well-tolerated treatments exist [3].

Artificial intelligence (AI) has become a popular approach to enhance risk identification [4]. While the diagnostic accuracy of this technology may not achieve 100%, its integration with the clinician's expertise has the potential to enhance overall performance [5]. Current clinical guidelines for primary prevention of CVD emphasize identifying asymptomatic patients who may benefit from preventive measures (e.g., statin therapy) based on their predicted risk [6]. Our study aims to systematically review and synthesize the literature on AI-enabled tools and interventions for CVD prevention, focusing on the following research questions: RQ1: How is AI used in the prevention of cardiovascular disease? RQ2: How does AI compare to existing risk scores to prevent cardiovascular disease?

2 Methods

We conducted this systematic review following the methodology by Joanna-Briggs Institute (JBI) [7] and reported according to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist [8]. The study protocol was developed and registered on the International Prospective Register of Systematic Reviews (PROSPERO) under the ID CRD42022369548. We used the PICO (population, intervention, comparator, outcome) framework [9] to develop the inclusion and exclusion criteria. The included studies had to be written in English and published between 1999 and October 2022.

Population. We included any individual receiving or providing care. This includes four key groups: (1) the general population, (2) patients receiving care, (3) healthcare professionals delivering care, and (4) health managers overseeing care delivery.

Intervention. Any AI intervention used for risk prediction and prevention of CVD. We included all AI methods, such as machine learning methods.

Comparator. We included studies with or without a comparator as standard care or other intervention. Head-to-head comparisons were also considered for inclusion.

Outcomes. We looked for studies reporting the effectiveness of preventing CVD by early risk prediction, healthcare monitoring, or intervention delivery.

Exclusion Criteria. Studies focusing on rheumatic heart disease, congenital cardiac disease, and cardiac surgery were excluded because they primarily address disease management or procedural/surgical care rather than primary or secondary prevention of cardiovascular disease. Additionally, articles centered on diagnosing CVD, such as using ECG or PPG, rather than early detection, were omitted due to their limited role in disease prevention. Studies not written in English were also excluded.

An information specialist developed the search strategy in collaboration with our team members. The search strategy was first conducted in Ovid-MEDLINE. After two rounds of revisions, a comprehensive search of seven electronic databases (Ovid-MEDLINE, Embase, CINAHL, Cochrane CENTRAL, Web of Science, IEEE Xplore,

Scopus) was performed in October 2022. We used EndNote 20 to import all references and to deduplicate articles. We performed the screening in two stages using the software Rayyan [10]. Titles and abstracts were screened for potentially eligible studies. In the next stage, full-text reports were retrieved and further assessed for eligibility according to the specified inclusion criteria. Each stage was blinded and independently reviewed by two authors (CS and LPS), and a third senior author (SAR) resolved any conflicts. We further performed a sub-group analysis on studies that compare AI to standard care, such as existing risk score systems. We developed an extraction form which was guided by the Evidence-Based Practice for Improving Quality (EPOC) and the recommendations outlined in Stevens et al., “Recommendations for Reporting Machine Learning Analyses in Clinical Research.” [11] Two reviewers (CS and LPS) performed data extraction and quality assessment. A final data extraction template was prepared after discussions with expert team members.

To answer the second question in this review, we conducted a thorough data extraction from studies ($n = 16$) that used a clinical comparator. Due to the lack of existing tools to assess studies involving AI, risk of bias (RoB) was assessed during data extraction of studies involving AI using the Prediction model Risk of Bias Assessment Tool (PROBAST) and components of the comprehensive checklist for the assessment of medical AI studies [12]. This assessment tool systematically evaluated the RoB in six domains, including participants, predictors, outcome, analysis, AI data, and AI Model validation. We performed a narrative and qualitative data synthesis to summarize the characteristics of the included studies, considering study design, context, and participant characteristics. Supplementary materials, including the full search strategy are available at: <https://github.com/rahimi-s-lab/AI-CVD-Systematic-Review>.

3 Results

A total of 12,149 records were found using the search strategy. A total of 266 studies met the selection criteria, of which 16 articles were sub-grouped for further analysis.

This systematic review reports on the sub-group analysis of 16 studies which used a clinical comparator. These studies collectively included 51,687,627 participants. Sample sizes ranged from 500 to over 6 million (median 28,185). Studies were conducted in North America ($n = 5$), Europe ($n = 6$), Asia ($n = 5$), Australia ($n = 1$), and across multinational datasets ($n = 2$). All 16 studies were retrospective or used research datasets. None reported deployment in routine clinical care. Medical tasks included CVD risk prediction, stroke prediction, ASCVD prediction, CAD risk prediction, MACE and bleeding-event prediction in ACS, and automated coronary artery calcium (CAC) scoring. The Framingham Risk Score was the most frequently used comparator ($n = 8/16$, 50%), followed by SCORE/SCORE2 ($n = 3$), ACC/AHA Risk Calculator ($n = 2$), CHA2DS2-VASc ($n = 2$), QRISK3 ($n = 2$) and TIMI Risk Score ($n = 2$) (Fig. 1).

The most commonly applied AI approaches were neural networks (including attention-based deep neural networks and quantum neural networks), gradient-boosted trees and boosting variants (XGBoost, GBM, AdaBoost), random forests, and support vector machines. Several studies employed ensemble or super-learner architectures combining multiple algorithms. Risk of bias varied across the six domains. Most concerns

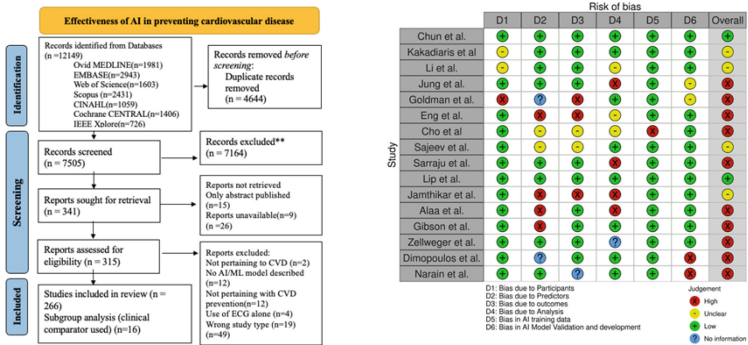


Fig. 1. PRISMA flow diagram illustrating the literature search and study selection process(left) and the risk of bias assessment of included studies(right)

clustered around participants, analysis, and model validation domains, reflecting reliance on internal cross-validation and limited external testing.

4 Discussion

Our analysis revealed a significant trend: machine learning models consistently outperformed conventional methods such as the Framingham Risk Score (FRS), ACC/AHA Risk Calculator, and CHA2DS2-VASc score in various settings.

Most models were developed using retrospective or single-cohort data, which may limit the generalizability of the findings. Reported performance gains were often attributed to flexible algorithms (e.g., gradient-boosted trees and neural networks) capable of modelling nonlinear relationships and complex interactions between variables. Moreover, integrating novel predictors such as socioeconomic factors and lifestyle parameters expanded risk stratification beyond traditional clinical markers, supporting a more holistic view of cardiovascular risk. This pattern aligns with broader reviews of AI in cardiovascular medicine, which describe machine learning’s capacity to extract non-linear signals from large, multimodal datasets that conventional regression-based scores cannot fully capture [4].

AI in prevention focuses more on risk prediction and early detection than on delivering actual interventions. Among studies with a clinical comparator, reported improvements over standard care varied widely (2.71–62.07%), indicating substantial context dependence across populations, outcomes, and evaluation frameworks. While AUROC, calibration, and NRI were commonly reported, measures critical for imbalanced prediction tasks (e.g., sensitivity, specificity, recall, precision, F1-score) were frequently absent, restricting assessment of clinical utility.

5 Conclusion

This review shows that AI, particularly machine learning models, generally outperforms traditional risk calculators for predicting cardiovascular disease across diverse settings, with higher accuracy, sensitivity, and specificity. However, most evidence is limited

to retrospective, single-center studies with internal validation only, raising concerns about generalizability and real-world performance. Future work must prioritize external validation, explainable models, prospective trials, and implementation-focused studies that demonstrate impact on patient outcomes and healthcare delivery.

References

1. Wang, H., et al.: Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the global burden of disease study 2015. *Lancet*. **388**(10053), 1459–1544 (2016)
2. Vaduganathan, M., et al.: The Global Burden of Cardiovascular Diseases and Risk: a Compass for Future Health, pp. 2361–2371. American College of Cardiology Foundation Washington DC (2022)
3. Visseren, F.L.J., et al.: ESC Guidelines on cardiovascular disease prevention in clinical practice. *Eur. J. Prev. Cardiol.* **29**(1), 5–115 (2021)
4. Jankauskas, S.S., et al.: Artificial intelligence in cardiovascular medicine: a giant step in personalized medicine? *J. Pers. Med.* **16**(4), 192 (2026)
5. Miller, D.D., et al.: Artificial intelligence in medical practice: the question to the answer? *Am. J. Med.* **131**(2), 129–133 (2018)
6. Arnett, D.K., et al.: 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American college of Cardiology/American heart association task force on clinical practice guidelines. *Circulation*. **140**(11), e596–e646 (2019)
7. Moola, S., et al.: JBI Manual for Evidence Synthesis. Systematic Reviews of Etiology and Risk. JBI (2020)
8. Page, M.J., et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. **372**, n71 (2021)
9. Richardson, W.S., et al.: The well-built clinical question: a key to evidence-based decisions. *ACP J. Club*. **123**(3), A12–A13 (1995)
10. Ouzzani, M., et al.: Rayyan-a web and mobile app for systematic reviews. *Syst. Rev.* **5**(1), 210 (2016)
11. Stevens, L.M., et al.: Recommendations for reporting machine learning analyses in clinical research. *Circ. Cardiovasc. Qual. Outcomes*. **13**(10), e006556 (2020)
12. Cabitza, F., et al.: The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical AI studies. *Int. J. Med. Inform.* **153**, 104510 (2021) Elsevier



Rehabilitation Engagement Prediction in Pediatric Patients with Neuromotor Disorders: An Adaptive Hierarchical Bayesian Approach

Simone Costantini^{1,2} , Emilia Biffi², Giuseppe Andreoni^{2,3},
Anna Bianchi¹, and Fabio Storm²

¹ Department of Electronics Information and Bioengineering, Politecnico di Milano, Milan, Italy

simone.costantini@polimi.it

² Scientific Institute, IRCCS “E. Medea”, Bosisio Parini, Italy

³ Department of Design, Politecnico di Milano, Milan, Italy

Abstract. Patient engagement is essential for improving clinical outcomes in pediatric neuromotor rehabilitation. Physiological computing offers a quantitative aid for its assessment, but clinical implementation is limited by intrinsic noise in therapist-reported labels and high inter-subject physiological variability. The study aims to predict absolute Engagement State and relative Engagement Dynamics in 31 underage participants (21 with neuromotor disorders, 10 neurotypical). Multimodal data (heart rate variability, electrodermal activity, and acceleration) and PRIME-SP engagement labels were collected during technology-assisted rehabilitation. Following feature extraction, confident learning reduced noise in therapist annotations, and hierarchical Bayesian models were trained and tested under cold start and adaptive learning evaluation paradigms. Results proved the non-destructive effectiveness of adaptive learning, achieving F1-Scores of 0.75 and 0.79 for Engagement State and Dynamics, respectively. The study marks a first step towards the development of a decision support system for real-time engagement assessment.

Keywords: Rehabilitation Engagement · HBM · Affective Computing

1 Introduction

High-intensive, technology-assisted training is crucial to promote neuroplasticity and improve motor recovery in pediatric neuromotor rehabilitation, but its intrinsic repetitive nature can lead to boredom and disengagement [12]. This is critical, as rehabilitation engagement is essential for enhanced clinical outcomes [6]. Currently, engagement assessment heavily relies on psychometric scales, which are discrete and subjective tools, and fail to capture dynamic fluctuations of engagement [2]. Physiological computing may offer a valid continuous

and quantitative alternative [4], although its deployment in clinics presents two major challenges. The first depends on the inter-subject variability of physiological patterns, which leads to a significant generalization gap in pooled models or data scarcity in subject-dependent models [4]. The second addresses ground truth reliability: labels rely on the therapist’s subjective interpretation of the patient’s internal state, risking the embedding of human biases in the final predictions [9]. The present work aims at objectively predicting rehabilitation engagement. A Machine Learning (ML) framework based on Hierarchical Bayesian Models (HBMs) [5] was implemented to manage inter-subject variability, leveraging partial pooling. Furthermore, a confident learning approach [9] was exploited to handle noisy annotations before training. Finally, clinical interpretability was provided through a post-hoc explainability analysis.

2 Materials and Methods

Experimental Design and Target Definition. The study involved 31 participants: 10 neurotypical underage subjects and 21 pediatric patients with neuromotor disorders due to cerebral palsy or acquired brain injury. The study was approved by the Ethical Committee Lombardia 2 (L2-246) and registered at ClinicalTrials.gov (NCT06993389). Participants underwent 1–3 robot-assisted rehabilitation sessions using either the Lokomat (Hocoma AG, Volketswil, Switzerland) or GRAIL (Motek Medical, The Netherlands) systems. Each session consisted of a 10-minute rest phase followed by 3–6 5-minute gamified exercises (exergames). Multimodal data, including Heart Rate Variability (HRV) from electrocardiographic (ECG) signal, electrodermal activity (EDA) and 3-axis acceleration signals, were acquired using Shimmer3 sensors (Shimmer Sensing, Dublin, Ireland). Therapists filled the PRIME-SP questionnaire [7] after each exergame. To derive a continuous target from the questionnaires, an exploratory factorial analysis [3] was applied to the four PRIME-SP dimensions. The extracted single continuous latent score was then binarized for two targets. For **Engagement State** (absolute intensity), exergames whose score exceeded the neutral point were labeled High Engagement, otherwise Low Engagement. For **Engagement Dynamics** (in-session fluctuation), a differential latent score relative to the preceding exergame was computed ($\Delta(t) = \text{Score}(t) - \text{Score}(t - 1)$). Negative variations were labeled Disengagement, whereas stable or positive changes as Maintenance.

Data Processing and Refinement. HRV, EDA and acceleration signals were preprocessed through standard filtering pipelines to limit noise and artifacts. Following transient removal, signals in the rest phase or related to exergames were segmented into 3-minute windows. From each window, a comprehensive set of 49 literature-driven temporal, spectral, and non-linear features was extracted [1, 10, 11]. To align feature representation with classification targets, exergames features were normalized by subtracting subject-specific rest values for the Engagement State, and by subtracting the mean value of the preceding

exergame for the Engagement Dynamics. To mitigate labels noise, a confident learning approach was implemented [9]. A Logistic Regression model, selected for its robustness against overfitting, was trained using Leave-One-Subject-Out (LOSO) cross-validation to estimate the inconsistency between features and labels (contrary evidence probability). To align window-level predictions with exergame-level labels, predictions were aggregated through temporal majority voting, thus averaging probabilities across all windows inside an exergame to obtain a single exergame-level probability. Exergames with contrary evidence probability between 0.7 and 0.9 were pruned, while labels were inverted when the probability exceeded 0.9. Final datasets consisted of 989 samples for Engagement State and 756 for Engagement Dynamics.

Machine Learning Framework and Explainability. The ML framework was nested within an outer LOSO cross-validation loop to test exclusively on unseen subjects and run independently for both classification targets. First, features were z-score normalized using statistical parameters fitted on the training set. To identify the most discriminative features, a linear discriminant analysis classifier was trained through an inner 5-fold cross-validation loop to optimize its shrinkage parameter. Feature selection was then driven by SHapley Additive Explanations (SHAP) [8], selecting the top-10 features ranked by mean absolute SHAP values. Next, leveraging partial pooling to balance generalization and personalization, a HBM was applied to estimate population-level fixed effects, representing the global impact of features, while also modeling subject-specific random intercepts to adjust for individual variability. The framework was evaluated under two scenarios simulating distinct stages of the patient’s clinical life-cycle: **Cold Start** (CS), simulating a new patient arrival by computing predictions for the held-out subject using exclusively the $N - 1$ population model priors; and **Adaptive Learning** (AL), simulating longitudinal adaptation. In AL, the HBM was initialized with the population priors and sequentially updated after every exergame on the subject’s history, progressively adjusting the parameters specifically for that patient. Finally, for trustworthiness and explainability, a global HBM was fitted on the entire dataset to estimate the magnitude and direction of the features’ influence on engagement predictions.

3 Results and Discussion

The HBM-based framework achieved satisfactory classification performance (Table 1), proving its capability to generalize over a heterogeneous pediatric cohort. Comparing the two evaluation paradigms, AL consistently exceeded CS performance for Engagement State, and reached comparable performance for Engagement Dynamics, although no statistically significant differences were found, probably due to the limited number of update steps available per subject (between 2 and 17 exergames), restricting the HBM opportunities to diverge from the priors. Nevertheless, the AL paradigm proved to be non-destructive, preserving or improving CS performance even with few data points, which is

highly desirable in longitudinal treatments. Noteworthy, Engagement Dynamics showed slightly higher performance compared to Engagement State, potentially suggesting that the inference of the fluctuation of engagement was inherently more feasible than estimating its absolute state [2].

Table 1. LOSO Cross-Validation ($N = 31$) performance metrics (Mean (SD)).

Target	Paradigm	F1-Score	Precision	Recall	ROC AUC
Engagement State	CS	0.72 (0.29)	0.72 (0.32)	0.76 (0.24)	0.77 (0.26)
	AL	0.75 (0.27)	0.76 (0.28)	0.77 (0.23)	0.73 (0.24)
Engagement Dynamics	CS	0.78 (0.21)	0.81 (0.20)	0.79 (0.21)	0.82 (0.23)
	AL	0.79 (0.22)	0.80 (0.21)	0.80 (0.21)	0.81 (0.23)

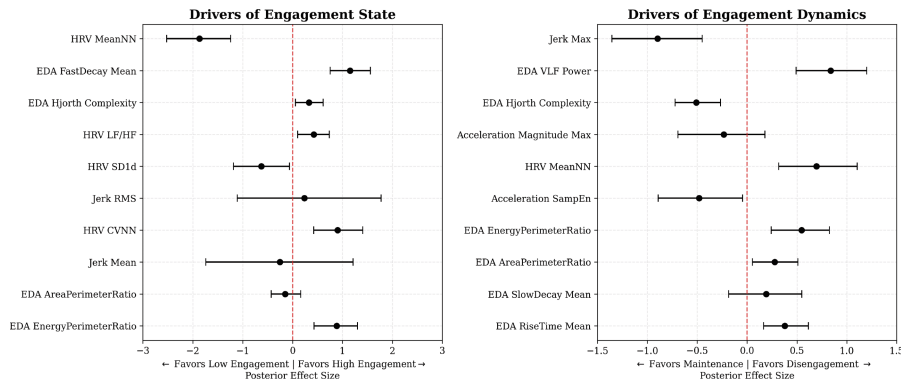


Fig. 1. Explainability analysis showing the fixed effects coefficients of the HBM.

Additionally, the HBM framework offered a trustworthy alternative to black-box approaches [5]. As shown in Fig. 1, Engagement State inference was primarily driven by autonomic features. A shift towards sympathetic dominance (negative coefficient of HRV MeanNN, positive LF/HF) and vagal withdrawal (negative SD1d) correlated high engagement to increased arousal [11]. Likewise, positive coefficients for EDA energy-perimeter ratio and Hjorth complexity linked higher engagement to greater sudomotor activity and sympathetic arousal [1]. Conversely, Engagement Dynamics was also driven by kinematic features reflecting motor behavior: maximum jerk magnitude and sample entropy of acceleration both revealed negative coefficients, working as protective factors against disengagement. High values of jerk (i.e., rate of change of acceleration) and entropy

(i.e., movement complexity) reflected rapid and adaptive motor adjustments [10], which favored engagement maintenance, whereas a shift towards lethargic or monotonous patterns likely acted as precursors to disengagement. Furthermore, autonomic features indicating parasympathetic relaxation significantly increased the likelihood of disengagement [11].

4 Conclusions

The study proved the feasibility of adaptive HBMs to infer rehabilitation engagement state and dynamics, balancing population-level generalization with subject-specific personalization. The explainability analysis identified key motor and autonomic drivers, resulting in trustworthy measures for clinical practice. Despite the restricted number of exergames per participant, the study sets the stage for an objective clinical decision support system, aiding therapists in tailoring rehabilitation strategies to both psychological and motor needs of patients.

Acknowledgments. This work was funded by the Italian Ministry of Health (Ricerca Corrente 2025-2026 awarded to Dr. E. Biffi), and partially funded by the Italian Ministry of University and Research (Doctoral Scholarship awarded to S. Costantini).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Boucsein, W.: *Electrodermal Activity*. Springer, New York (2012). <https://doi.org/10.1007/978-1-4614-1126-0>
2. Bright, F.A.S., McPherson, K.M.: A conceptual review of engagement in health-care and rehabilitation. *Disabil. Rehabil.* **37**(8) (2015). <https://doi.org/10.3109/09638288.2014.933899>
3. Fabrigar, L.R., Strahan, E.J.: Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods* **4**(3) (1999). <https://doi.org/10.1037/1082-989X.4.3.272>
4. Fairclough, S.H.: Fundamentals of physiological computing. *Interact. Compute.* **21**(1–2) (2009). <https://doi.org/10.1016/j.intcom.2008.10.001>
5. Gelman, A., Hill, J.: *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge (2006). <https://doi.org/10.1017/CBO9780511790942>
6. Graffigna, G., Claudio, A.: Promozione del patient engagement in ambito clinico-assistenziale per le malattie croniche: raccomandazioni dalla prima conferenza di consenso italiana. *Recenti Progressi in Medicina* **108**(11) (2017). <https://doi.org/10.1701/2812.28441>
7. King, G., Chiarello, L.A.: Youth engagement in pediatric rehabilitation: service providers' perceptions in a real-time study of solution-focused coaching for participation goals. *Phys. Occup. Therapy Pediatr.* **38**(5) (2018). <https://doi.org/10.1080/01942638.2017.1405864>
8. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*, vol. 30 (2017)

9. Northcutt, C., Chuang, I.: Confident learning: estimating uncertainty in dataset labels. *J. Artif. Intell. Res.* **70** (2021). <https://doi.org/10.1613/jair.1.12125>
10. Sejdić, E., Brach, J.S.: A comprehensive assessment of gait accelerometry signals. *IEEE Trans. Biomed. Eng.* **61**(8) (2014). <https://doi.org/10.1109/TNSRE.2013.2265887>
11. Shaffer, F., Zerr, C.L.: A healthy heart is not a metronome. *Front. Psychol.* **5** (2014). <https://doi.org/10.3389/fpsyg.2014.01040>
12. Zhong, B., Zhang, M.: Bringing psychological strategies to Robot-Assisted physiotherapy for enhanced treatment efficacy. *Front. Neurosci.* **13** (2019). <https://doi.org/10.3389/fnins.2019.00984>



Explainable AI Identifies High Retrospective Prodromal Symptom Burden in Parkinson's Disease

Dante Trabassi¹ (✉) , Stefano Filippo Castiglia¹ , Roberto De Icco² ,
Cristina Tassorelli² , Alberto Ranavolo³ , and Mariano Serrao¹

¹ Department of Medical and Surgical Sciences and Biotechnologies, Sapienza University of Rome, Latina, Italy

dante.trabassi@uniroma1.it

² Department of Brain and Behavioral Sciences, University of Pavia, Pavia, Italy

³ Department of Occupational and Environmental Medicine, Epidemiology and Hygiene, INAIL, Rome, Italy

Abstract. Characterizing heterogeneity in Parkinson's disease (PD) related to prodromal symptom burden may provide insight into early disease mechanisms. Wearable gait analysis offers a scalable approach to capture subtle motor alterations. We developed an interpretable machine learning framework to identify high *retrospective* prodromal symptom burden (≥ 3 symptoms) in 275 individuals with PD using single-sensor trunk accelerometry. A three-stage feature selection pipeline identified five key predictors: multiscale entropy (MSE) along three axes, vertical improved harmonic ratio (iHR_v), and body weight. A Random Forest classifier achieved $ROC_{AUC} = 0.84$, $PR_{AUC} = 0.86$, F1-score = 0.76. Class imbalance was addressed via CTGAN-based augmentation applied within the training set. Explainability analyses consistently identified increased medio-lateral gait complexity (MSE_{ml}) and reduced iHR_v as dominant contributors.

These findings define an interpretable gait phenotype associated with retrospective prodromal burden in PD. While not reflecting true prodromal prediction, this framework provides mechanistic insight into early motor alterations and supports scalable digital phenotyping.

Keywords: Parkinson's disease · wearable sensors · digital biomarkers · gait analysis · explainable artificial intelligence · retrospective prodromal burden

1 Introduction

Parkinson's disease (PD) involves a prolonged preclinical phase characterized by the emergence of non-motor symptoms such as hyposmia, REM sleep behavior disorder, depression, and constipation [1, 2]. While individual symptoms have limited specificity, their accumulation is often used as a proxy of increased prodromal burden. However, in established PD, such burden is retrospectively defined and does not necessarily reflect true prodromal disease progression, but rather a reconstructed symptom profile within diagnosed patients.

In parallel, subtle alterations in gait dynamics may arise early and reflect underlying multisystem involvement. Wearable inertial sensors provide scalable access to trunk-derived metrics capturing rhythmicity, symmetry, variability, and complexity of locomotion [3, 4]. Entropy-based and harmonic-based features have shown sensitivity to disease-related motor alterations and fall risk in PD [5, 6].

Despite these advances, the relationship between objective gait biomarkers and retrospective prodromal symptom burden within PD remains poorly characterized, particularly in the absence of interpretable models linking biomechanical features to clinically defined symptom burden. In addition, methodological challenges such as class imbalance and limited interpretability constrain the clinical utility of machine learning models in this setting.

In this study, we propose an interpretable machine learning framework using single-sensor gait data to identify high retrospective prodromal symptom burden (≥ 3 symptoms) in PD. This threshold was defined within the present dataset to capture a subgroup with higher cumulative symptom load and is therefore interpreted in a descriptive rather than predictive or causal sense. Specifically, we aim to: (i) identify a minimal and stable biomechanical signature associated with high symptom burden; (ii) evaluate a data-driven balancing strategy to address class imbalance; and (iii) provide transparent model interpretation through explainability analysis.

2 Methods

2.1 Cohort and Outcome Definition

We analyzed 275 individuals with idiopathic Parkinson's disease (Hoehn & Yahr I–IV) recruited at two Italian centers (Latina and Pavia). All participants performed a standardized 30-m walk at self-selected speed using a lumbar inertial sensor (L5, 100 Hz). Assessments were conducted under stable medication conditions.

Four non-motor symptoms (hyposmia, REM sleep behavior disorder, depression, constipation) were retrospectively extracted from clinical records. Participants were classified as high retrospective prodromal burden (PB+) if presenting ≥ 3 symptoms, and low burden (PB−) otherwise. This threshold was derived through exploratory evaluation of symptom-count configurations to balance class prevalence and stability. As such, it reflects a data-driven analytical definition and is not intended as a clinically validated cutoff, nor as evidence of prospective prodromal identification.

2.2 Gait Features and Dimensionality Reduction

From trunk acceleration signals, 35 biomechanical and clinical variables were extracted, including spatiotemporal parameters, harmonic-based indices (HR, iHR), recurrence metrics (DET), and multiscale sample entropy (MSE) along three axes. One artifact-free trial per subject (≥ 20 gait cycles) was retained.

A three-stage feature selection pipeline was applied. First, low-variance and highly correlated variables ($|r| > 0.90$) were removed, followed by ANOVA filtering ($p < 0.05$). Recursive feature elimination with Random Forest and Gini importance stability analysis were then used to identify robust predictors.

Five features consistently emerged: MSE (medio-lateral, antero-posterior, vertical), vertical improved harmonic ratio (iHR_v), and body weight.

2.3 Data Balancing and Classification

The dataset was split into training (80%) and testing (20%) sets using stratified sampling. Synthetic data augmentation was applied exclusively to the training set to avoid information leakage. A Conditional Tabular GAN (CTGAN) was used to balance classes to a 1:1 ratio. Synthetic data quality was assessed through distributional comparison (Kolmogorov–Smirnov tests), correlation structure preservation (mean absolute error = 0.082), and discriminator performance (AUC = 0.30), indicating high realism. A Random Forest classifier was trained with hyperparameters optimized via five-fold cross-validation. Performance was evaluated using ROC-AUC, PR-AUC, F1-score, and Brier score, with 1000-bootstrap confidence intervals.

To contextualize model performance, simpler baseline models (logistic regression and class-weighted Random Forest) were evaluated under the same train/test splits.

2.4 Model Explainability

Model interpretability was assessed using SHAP values to quantify global and individual feature contributions. Nonlinear effects were explored through partial dependence and interaction analysis (Friedman H-statistic).

A shallow surrogate decision tree (depth = 3) was used to approximate model behavior and extract interpretable decision rules.

3 Results

3.1 Cohort Characteristics

The cohort included 275 individuals with Parkinson's disease, of whom 110 (40%) were classified as high retrospective prodromal burden (PB+) and 165 (60%) as low burden (PB−). Groups were comparable in age, disease duration, motor severity (UPDRS III, Hoehn & Yahr), and medication exposure, indicating that differences were not driven by global disease stage. PB+ individuals showed lower age at onset and lower body weight, but no difference in fall history.

3.2 Model Performance and Data Balancing

The feature selection pipeline identified five stable predictors: multiscale entropy along three axes (MSE_{ml} , MSE_{ap} , MSE_v), vertical improved harmonic ratio (iHR_v), and body weight.

In the original dataset, the Random Forest achieved ROC-AUC = 0.796 and F1 = 0.624. After CTGAN-based balancing, performance improved to ROC-AUC = 0.838, PR-AUC = 0.863, and F1 = 0.760.

Synthetic data validation confirmed preservation of data structure, with no significant distributional differences, correlation matrix MAE = 0.082, and low discriminability between real and synthetic samples (AUC = 0.30).

Baseline models (logistic regression and class-weighted Random Forest) showed consistently lower performance across all evaluation metrics, supporting the use of a nonlinear classifier.

3.3 Explainability and Biomechanical Phenotype

SHAP analysis identified MSE_{ml} as the dominant predictor, with higher values strongly associated with PB+. Reduced iHR_v further contributed, indicating altered vertical rhythmicity and symmetry. Interaction analysis revealed a synergistic effect between MSE_{ml} and iHR_v , suggesting that lateral gait complexity and vertical symmetry jointly characterize the high-burden phenotype. A surrogate decision tree (depth = 3, fidelity = 0.84) confirmed this structure, with primary stratification driven by MSE_{ml} and subsequent splits involving gait speed, MSE_v , and body weight (Fig. 1). Overall, the model identifies a consistent biomechanical phenotype characterized by increased lateral gait complexity and reduced rhythmic symmetry, independent of global motor severity.

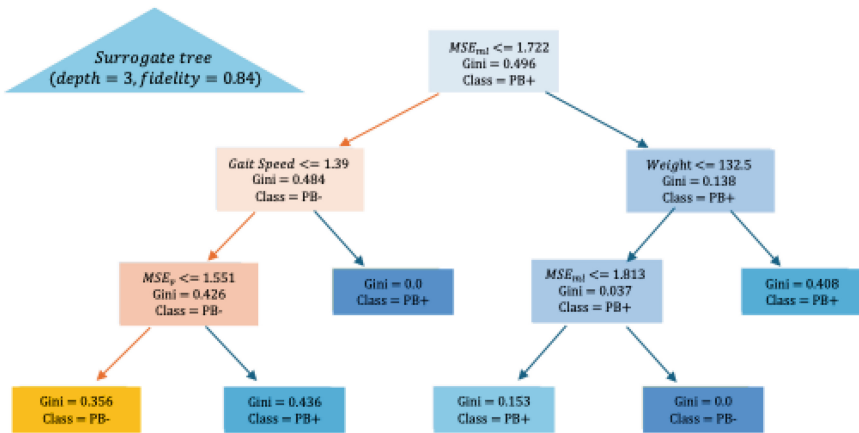


Fig. 1. Interpretable shallow decision tree (depth = 3; fidelity = 0.84) approximating the ensemble model.

4 Discussion

This study identifies a reproducible gait pattern associated with high *retrospective* prodromal symptom burden in Parkinson’s disease using single-sensor data and an interpretable machine learning framework.

The dominant contribution of medio-lateral multiscale entropy indicates increased irregularity of trunk dynamics and altered lateral control in individuals with higher

symptom burden. This pattern emerged independently of global motor severity, supporting the interpretation that it reflects phenotypic heterogeneity rather than disease stage. Reduced vertical harmonic ratio provided complementary information, suggesting impaired rhythmicity and symmetry. Together, these features delineate a multi-dimensional disruption of locomotor control.

The outcome definition requires careful interpretation. The ≥ 3 symptom threshold was derived through data-driven optimization, introducing circularity and potential inflation of performance. Accordingly, this target should be considered an analytical construct rather than a clinically validated cutoff. Moreover, “prodromal” refers here to retrospectively coded symptom burden in diagnosed patients and does not imply prospective identification of prodromal PD.

From a methodological perspective, CTGAN-based balancing improved sensitivity, but its added value relative to simpler approaches remains uncertain. The inclusion of body weight may reflect systemic or confounding effects rather than a direct biomechanical driver.

Several limitations must be acknowledged. The cross-sectional design precludes causal inference, and external validation is required to assess generalizability. Retrospective symptom coding may also introduce misclassification.

In summary, this work supports the existence of an interpretable gait phenotype associated with retrospective prodromal burden in PD, while highlighting the need for prospective validation and cautious interpretation of data-driven targets.

Acknowledgments. The activities conducted at the IRCCS Mondino Foundation were supported by a Research Grant from the Italian Ministry of Health.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Kalia, L.V., Lang, A.E.: Parkinson’s disease. *Lancet*. **386**(9996), 896–912 (2015)
2. Berg, D., et al.: MDS research criteria for prodromal Parkinson’s disease. *Mov. Disord.* **30**(12), 1600–1611 (2015)
3. Schalkamp, A.-K., et al.: Wearable movement-tracking data identify Parkinson’s disease years before clinical diagnosis. *Nat. Med.* **29**(8), 2048–2056 (2023)
4. Zijlstra, W., Hof, A.L.: Assessment of spatio-temporal gait parameters from trunk accelerations during human walking. *Gait Posture*. **18**(2), 1–10 (2003)
5. Costa, M., et al.: Multiscale entropy analysis of human gait dynamics. *Physica A: Stat. Mech. Appl.* **330**(1-2), 53–60 (2003)
6. Ihlen, E.A.F., et al.: Improved prediction of falls in community-dwelling older adults through phase-dependent entropy of daily-life walking. *Front. Aging Neurosci.* **10**, 44 (2018)



A Bayesian Latent-Threshold Framework for Fall Occurrence in Parkinson's Disease

Dante Trabassi¹ , Stefano Filippo Castiglia¹ , Roberto De Icco² ,
Cristina Tassorelli² , Alberto Ranavolo³ , and Mariano Serrao¹ 

¹ Department of Medical and Surgical Sciences and Biotechnologies, Sapienza University of Rome, Latina, Italy

dante.trabassi@uniroma1.it

² Department of Brain and Behavioral Sciences, University of Pavia, Pavia, Italy

³ Department of Occupational and Environmental Medicine, Epidemiology and Hygiene, INAIL, Rome, Italy

Abstract. Falls in Parkinson's disease (PD) are commonly modeled as direct outcomes of isolated gait features, although wearable measurements are noisy indicators of broader motor-control processes. We propose a Bayesian latent-threshold framework in which fall occurrence is modeled as the probabilistic manifestation of an unobserved gait instability construct. In 269 individuals with Parkinson's disease performing a 30-m walk with a single L5-mounted inertial sensor, trunk-derived features were grouped into lower-trunk kinematics, rhythmicity–recurrence, and neuromotor complexity domains. Within-domain principal component analysis provided three domain-specific axes, which were integrated in a Bayesian hierarchical measurement model to estimate subject-specific latent instability with posterior uncertainty. Fall occurrence was then modeled using a Bayesian probit threshold formulation. Latent instability showed a positive association with fall occurrence (posterior probability $\beta > 0 = 0.964$), and the inferred threshold ($\tau \approx 0.32$) identified a transition region where small changes in instability produced the largest changes in fall probability. Counterfactual contrasts showed baseline-dependent risk reductions, maximal near the threshold. This framework reframes wearable fall-risk modeling from feature-level prediction toward uncertainty-aware latent digital biomarker inference.

Keywords: Parkinson's disease · falls · Bayesian modeling · digital biomarkers

1 Introduction

Falls are a major source of morbidity in Parkinson's disease (PD), yet individual fall susceptibility remains difficult to characterize [1, 2]. Wearable-sensor studies commonly estimate fall risk through direct associations between selected gait features and fall history [3, 4]. However, observable biomechanical metrics are noisy and partial manifestations of broader motor-control processes rather than independent causal entities.

Gait instability may instead be viewed as an unobserved construct emerging from coordinated lower-trunk kinematics, rhythmic regulation, and neuromotor complexity. Latent-variable models are well suited to infer such constructs from multiple imperfect indicators while explicitly accounting for measurement uncertainty [5, 6].

In addition, fall risk may not increase linearly with impairment, but may rise sharply once instability crosses a transition region. Motivated by this perspective, we propose a Bayesian latent-threshold framework in which trunk-derived wearable features are first summarized into domain-specific representations and then integrated into a hierarchical latent model. Fall occurrence is subsequently modeled through a probabilistic threshold function.

Using data from 269 individuals with PD performing a standardized 30-m walk with a single lumbar inertial sensor, we evaluate whether latent gait instability provides an interpretable and uncertainty-aware representation of fall susceptibility.

2 Methods

2.1 Cohort and Feature Construction

We analyzed 269 individuals with idiopathic Parkinson's disease recruited at two Italian centers (Latina and Pavia). Participants performed a standardized 30-m straight walk at self-selected speed while wearing a single lumbar inertial measurement unit positioned at L5 (100 Hz). Fall occurrence was defined as at least one fall in the previous year. All assessments were performed in the clinically defined ON-medication state, i.e., during the expected symptomatic benefit of dopaminergic therapy.

The analytical workflow consisted of feature extraction, domain reduction, latent instability estimation, and threshold-based fall modeling. Trunk acceleration signals were grouped *a priori* into three biomechanical domains: lower-trunk kinematics (pelvic tilt, obliquity, rotation), rhythmicity–recurrence (harmonic ratios, recurrence determinism), and neuromotor complexity. The latter was quantified using multiscale sample entropy computed from trunk acceleration time series during the walking trial.

Within each domain, standardized features were summarized using principal component analysis (PCA). The first principal component (PC1) was retained to obtain one interpretable domain-level representation per construct and avoid over-parameterization. PC1 explained 52.9% of variance in lower-trunk kinematics, 30.8% in rhythmicity–recurrence, and 61.1% in neuromotor complexity.

2.2 Bayesian Latent Instability Model

Let y_{ij} denote the PC1 score of subject i in domain j . The three domain-specific scores were modeled as noisy indicators of a common latent gait instability variable s_i :

$$y_{ij} \sim \text{Normal} \left(\lambda_j \cdot s_i, \psi_j^2 \right)$$

where λ_j are domain loadings and ψ_j^2 are domain-specific residual variances. For identifiability, the loading of lower-trunk kinematics was fixed to 1. Individual latent scores

were linked through a population distribution:

$$s_i \sim \text{Normal} \left(0, \sigma_s^2 \right)$$

Posterior inference was performed using Hamiltonian Monte Carlo. For each subject, latent instability was summarized by posterior mean $\mathbf{m}_i$ and posterior variance $\mathbf{v}_i$.

2.3 Threshold and Counterfactual Analysis

Fall occurrence was modeled through a Bayesian probit threshold model:

$$P(\text{fall}_i = 1 \mid s_i, \tau) = \Phi(s_i - \tau)$$

where Φ is the standard normal cumulative distribution function and τ is the population-level instability threshold.

Measurement uncertainty was propagated using $s_i \sim \text{Normal}(m_i, v_i)$.

Counterfactual contrasts for hypothetical instability reductions Δs were computed as:

$$\Delta P_i(\Delta s) = \Phi(s_i - \Delta s - \tau) - \Phi(s_i - \tau)$$

Posterior uncertainty was summarized using 95% highest density intervals (HDI).

3 Results

3.1 Latent Instability Structure and Association with Falls

Within-domain principal component analysis yielded one dominant axis per biomechanical domain. PC1 explained 52.9% of variance in lower-trunk kinematics, 30.8% in rhythmicity–recurrence, and 61.1% in neuromotor complexity. Pairwise correlations among the three domain-specific axes were low ($|\text{r}| \leq 0.15$), supporting complementary rather than redundant information.

In the Bayesian measurement model, rhythmicity–recurrence showed a positive loading on latent instability (posterior mean 0.15, 95% highest density interval excluding zero), whereas neuromotor complexity showed weaker contribution. Posterior latent scores were continuous and well identified across participants (mean posterior SD ≈ 0.10). Latent instability showed a positive probabilistic association with fall occurrence. A one-standard deviation increase corresponded to an odds ratio of approximately 1.4, with posterior probability $P(\beta > 0) = 0.964$.

3.2 Threshold Effects and Counterfactual Contrasts

The Bayesian probit model estimated an instability threshold of $\tau \approx 0.32$ (95% HDI [0.10, 0.53]). Fall probability crossed 50% near this level and exhibited maximal slope in its vicinity (Fig. 1), identifying a transition region where small instability changes corresponded to disproportionately large risk changes.

Counterfactual contrasts revealed baseline-dependent benefits under hypothetical instability reductions. For $\Delta s = 0.3$, the maximal posterior mean reduction in fall probability was approximately -0.12 and occurred near the inferred threshold. At the individual level, 73.9% of participants showed an absolute posterior mean reduction $\geq 5\%$.

Posterior predictive checks showed close agreement between observed fall prevalence (0.392) and posterior predictive mean prevalence (0.389), indicating adequate calibration.

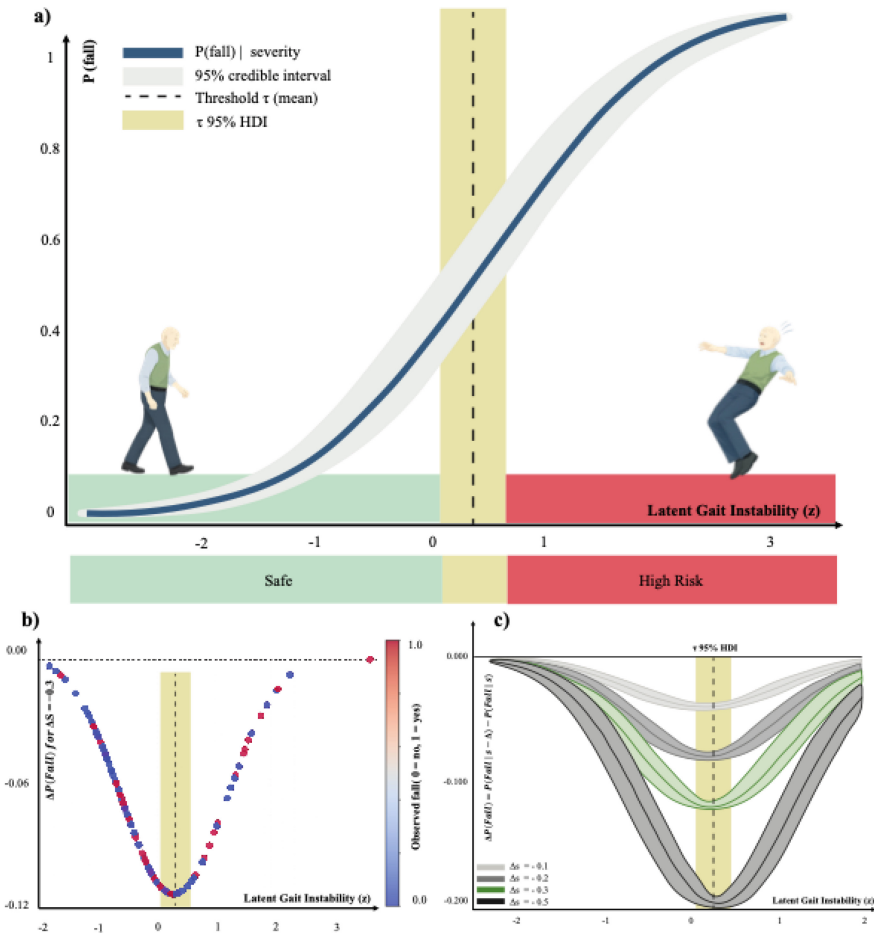


Fig. 1. Threshold structure and baseline-dependent counterfactual effects. (a) Posterior mean probit curve linking latent gait instability to fall probability, with 95% credible interval and inferred threshold τ (mean and HDI). (b) Individual-level change in fall probability for $\Delta s = 0.3$ across baseline instability. (c) Posterior counterfactual contrasts for multiple Δs values, showing maximal sensitivity near the inferred threshold.

4 Discussion

This study supports the interpretation of fall susceptibility in PD as a thresholded manifestation of an underlying latent gait instability process inferred from wearable trunk biomechanics. Rather than relying on isolated gait metrics, the proposed framework integrates complementary biomechanical domains into a single uncertainty-aware latent construct associated with fall occurrence.

The estimated threshold suggests that fall risk may not increase uniformly across the instability continuum. Instead, a transition region emerges in which relatively small changes in latent instability correspond to disproportionately large changes in predicted fall probability. This may help explain why patients with similar observable motor severity can experience different fall outcomes. Counterfactual analyses further showed that hypothetical instability reductions yielded the greatest expected benefit near the inferred threshold.

Several limitations should be acknowledged. Fall history was retrospectively assessed and dichotomized, introducing potential recall bias and preventing modeling of fall frequency. The study was cross-sectional and does not establish temporal causality. In addition, the numerical value of the inferred threshold may depend on the current acquisition protocol (30-m walk, single L5-mounted sensor, ON-medication state) and should therefore be externally validated in independent cohorts.

In conclusion, a Bayesian latent-threshold formulation reframes wearable fall modeling in Parkinson's disease from feature-level prediction toward structured probabilistic inference. If prospectively validated, this approach may support uncertainty-aware digital biomarkers and personalized fall-risk stratification.

Acknowledgments. The activities conducted at the IRCCS Mondino Foundation were supported by a Research Grant from the Italian Ministry of Health (Current Research 2025–2027).

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bloem, B.R., et al.: Falls and freezing of gait in Parkinson's disease: a review of two interconnected, episodic phenomena. *Mov. Disord.: Off. J. Mov. Disord. Soc.* **19**(8), 871–884 (2004)
2. Fasano, A., et al.: Falls in Parkinson's disease: a complex and evolving picture. *Mov. Disord.* **32**(11), 1524–1536 (2017)
3. Horak, F.B., Mancini, M.: Objective biomarkers of balance and gait for Parkinson's disease using body-worn sensors. *Mov. Disord.* **28**(11), 1544–1551 (2013)
4. Weiss, A., et al.: Does the evaluation of gait quality during daily life provide insight into fall risk? A novel approach using 3-day accelerometer recordings. *Neurorehabil. Neural Repair.* **27**(8), 742–752 (2013)
5. Bollen, K.A.: *Structural Equations with Latent Variables*. John Wiley & Sons (1989)
6. McElreath, R.: *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC (2018)



Ontology-Aligned Clinical NER in Emergency Medicine Using Weak Supervision

Amila Kugic¹(✉) , Adrian Schiegl², Alistair Tiefenbacher², Lukas Seper²,
Stefan Schulz¹ , and Markus Kreuzthaler¹

¹ Institute for Medical Informatics, Statistics and Documentation, Medical
University of Graz, Graz, Austria

{amil.kugic,stefan.schulz,markus.kreuzthaler}@medunigraz.at

² XUND Solutions GmbH, Vienna, Austria

{schiegl,tiefenbacher,seper}@xund.ai

Abstract. This study presents a weakly supervised framework for clinical named entity recognition (NER) in emergency medicine, leveraging ontology-aligned representations to enhance downstream modeling. From a corpus of 2,000 de-identified emergency department clinical notes, silver-standard annotations were generated using an automated annotator that assigned SNOMED CT concept labels to extracted spans. These labels were subsequently mapped to six custom clinical entity categories, i.e., XUND types, through a predefined SNOMED CT mapping table. We implemented a sequence-based clinical NER approach using the domain-specific language model medBERT.de to classify entities according to the derived XUND taxonomy. Our findings demonstrate the effectiveness of this approach, achieving a macro-averaged F1-measure of 0.91 across all six XUND types. These results suggest that ontology-driven weak supervision may help support effective NER systems in emergency medicine.

Keywords: Natural Language Processing · Electronic Health Records · Machine Learning

1 Introduction and Motivation

Clinical documentation in the emergency department is difficult due to unpredictable arrivals and variable acuity [1]. Emergency medicine providers must capture reasoning and outcomes beyond billing, including changes in condition, decisions, disposition, and instructions [15]. Documentation has grown longer due to multiple factors, such as efforts to enhance billing, mitigating malpractice risks, as well as leveraging electronic health record (EHR) tools. Large language model-assisted documentation may support this process but introduces risks such as bias, hallucinations, privacy issues, and reduced clinical specificity [6].

Therefore, NLP methods such as clinical NER require scalable labeling strategies, where ontology-backed silver-standard annotations can complement costly gold-standard labels. Prior work surveyed biomedical semantic annotation tools and their use for retrieval and EHR usability [9]. Multiple studies indicate that silver-standard corpora (SSC) can be competitive with gold-standard corpora (GSC), especially in cross-domain settings, and that SSC can support transfer learning benefits [7, 10]. For phenotype-oriented concept recognition, performance varies by EHR tools and is limited by issues such as boundary errors, abbreviations, and missing context [13]. Ontology-driven weak supervision and SSC-based transformer modeling have shown promising results in entity and phenotype tasks, including multilingual settings [5, 8, 12, 16].

In this paper, we evaluate automated annotation of German clinical discharge summaries and train a sequence-model-based classifier on silver-standard annotations using ontology-based SNOMED CT mapping tables for entity classification. This study addresses the following research questions: (i) Can silver-standard labeling combined with SNOMED CT mapping create a semantically coherent representation of clinical narratives suitable for downstream clinical reasoning tasks? (ii) What performance can be expected from such a workflow, and what are the implications of those performances?

2 Data and Methods

Initial Dataset. The initial dataset consisted of 2,000 de-identified German clinical documents. These were extracted from an Austrian hospital network provider running an i.s.h.med-based clinical information system called open-MEDOCS. The documents were all extracted from a department aimed at internal medicine and neurological cases in emergency situations.

Clinical Entity Types. The following six clinical entity types, named “*XUND types*”, were chosen for NER due to their relevance in emergency medicine by domain experts: Qualifier, MedicalCondition, Procedure, Location, HealthFactor, and Medication. Qualifiers modify or specify another entity, e.g., “*stechender*” (sharp) in “*stechender Schmerz*” (sharp pain). MedicalCondition refers to a disease, disorder, or clinically relevant sign or symptom, e.g., “*Appetitlosigkeit*” (loss of appetite). Procedures denote a medical intervention performed for diagnostic, therapeutic, or preventive purposes, e.g., “*Appendektomie*” (appendectomy). Locations indicate the anatomical site of a condition or health-related factor, e.g., “*Gehirn*” (brain) as the location of “*Demenz*” (Dementia). HealthFactor describes a behavioral, environmental, social, or personal factor that influences health, e.g., “*Raucher*” (smoker). Medication is a pharmaceutical drug or active substance used for treatment or prevention, e.g., “*Metformin*”.

Semantic Type Alignment. Annotation layers were defined in INCEPTION [3] and instantiated via a REST connection to the on-premise Averbis Health Discovery (AHD) pipeline using DKPro Cassis [11]. AHD is a UIMA-based clinical text-mining platform [4]. AHD-generated SNOMED annotations

were mapped to XUND types using a guideline-driven mapping table so that each span received both a SNOMED CT code and an XUND label. The mapping table covered all 368,268 active SNOMED CT concepts.

SNOMED CT GIT. The SNOMED CT Graz Interface Terminology (SNOGIT) links German clinical expressions, including jargon and short forms, to SNOMED CT to support text mining, coding, and semantic annotation [14]. It contains more than 8 million entries mapped semi-automatically to active SNOMED CT concepts to broaden coverage beyond standard terminology content.

Silver Standard Corpus. With the application of SNOGIT, the AHD pipeline generated silver-standard annotations that were converted to BIO (beginning-inside-outside) labels for the six XUND types. Texts were split on document level into 80% training, 10% validation, and 10% test set. No additional text pre-processing was applied.

Language Modeling. We fine-tuned medBERT.de [2] in a single fine-tuning run per XUND type. The German clinical BERT-based model was pre-trained on approximately 4.7 million clinical and health-related texts for sentence-level NER (max 512 tokens). Training used BIO-formatted SSC labels for 50 epochs (NVIDIA TITAN V; learning rate 2×10^{-5} , batch size 64, weight decay 0.01; training runtime 9.25 h per entity type). Evaluation was conducted at exact entity-level matching of the BIO labeling format. Performance was reported as macro-averaged precision, recall, and F1-measure over six XUND types, using the curated SSC as ground truth, similar to prior SSC work [5, 7, 16].

3 Results and Discussion

The SSC contained 180,372 annotations across 2,000 documents. Table 1 summarizes NER performance, with a macro-averaged F1-measure of 0.91 on the test set and only small differences across XUND types, ranging from 0.89 to 0.92. Although experimental setups differ, this result was consistent with prior SSC and GSC findings (0.86 vs. 0.88 in Fries et al. [5]) and within the range reported for SSC-based transformer models (0.90 in Wang et al. [16] and up to 0.79 in MedProcNER [12]). Error analysis showed problems leading back to lexical variations and terminology coverage, similar to Oellrich et al. [13].

System Limitation. Ontology alignment was limited by SNOMED CT coverage; rare, colloquial, or institution-specific expressions might have been missed. SNOGIT mitigated this issue for common local variants. The rule-based mapping of SNOMED to XUND labels could be rigid; however, because all active concepts in the reference terminology were included, overall NER coverage was likely only minimally affected. Our results on SSC are provisional and tied to the silver-labeling process, while further GSC evaluations are needed.

Table 1. NER results on the test set with medBERT.de language model.

XUND type	Precision	Recall	F1-score	Support
Qualifier	0.91	0.93	0.92	6,677
MedicalCondition	0.88	0.90	0.89	3,814
Procedure	0.89	0.93	0.91	2,713
Location	0.90	0.89	0.89	1,484
HealthFactor	0.93	0.91	0.92	890
Medication	0.89	0.94	0.91	300
macro-avg	0.90	0.92	0.91	

4 Conclusion and Outlook

An automated SNOMED CT-based clinical NER pipeline was developed using a silver-standard corpus and a fine-tuned medBERT.de model to identify six custom XUND entity types. The results suggest that ontology-driven weak supervision can support NER systems in emergency medicine and serve as a pre-annotation step to reduce manual labeling effort. Future work will focus on relation extraction and annotation of GSC to measure SSC-GSC performance gaps, enabling privacy-preserving, on-premise deployments.

Acknowledgments. This work was supported by the Austrian Research Promotion Agency (Österreichische Forschungsförderungsgesellschaft (FFG)) through the PRE-MEDICAL project (Predicting Patient Outcomes in Emergency Departments with Causal Machine Learning) (grant number 4365236 and number 5136377).

Disclosure of Interests. The authors declare no competing interests.

References

1. Blome, A., Yu, D., Lu, X., Schreyer, K.E.: Pitfalls of extensive documentation in the emergency department. *Ochsner J.* **20**(3), 299–302 (2020). <https://doi.org/10.31486/toj.19.0108>
2. Bressem, K.K., et al.: medBERT.de: a comprehensive German BERT model for the medical domain. *Expert Syst. Appl.* **237**, 121598 (2024). <https://doi.org/10.1016/j.eswa.2023.121598>
3. Eckart De Castilho, R., Klie, J.C., Gurevych, I.: Integrating INCEpTION into larger annotation processes. In: Hernandez Farias, D.I., Hope, T., Li, M. (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Miami, Florida, USA, pp. 110–121. Association for Computational Linguistics (2024). <https://doi.org/10.18653/v1/2024.emnlp-demo.12>. <https://aclanthology.org/2024.emnlp-demo.12/>
4. Ferrucci, D., Lally, A.: Accelerating corporate research in the development, application and deployment of human language technologies. In: *SEALTS 2003: Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture*

- of Language Technology Systems, Morristown, NJ, USA, pp. 67–74. Association for Computational Linguistics (2003). <https://doi.org/10.3115/1119226.1119236>
5. Fries, J.A., et al.: Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nat. Commun.* **12**(1), 2017 (2021). <https://doi.org/10.1038/s41467-021-22328-4>
6. Genes, N., Sills, J., Heaton, H.A., Shy, B.D., Scofi, J.: Addressing note bloat: solutions for effective clinical documentation. *J. Am. Coll. Emerg. Physicians Open* **6**(1), 100031 (2025). <https://doi.org/10.1016/j.acepjo.2024.100031>
7. Giorgi, J.M., Bader, G.D.: Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics* **34**(23), 4087–4094 (2018). <https://doi.org/10.1093/bioinformatics/bty449>. Place: Oxford, England
8. Greco, K.F., et al.: A weakly supervised transformer for rare disease diagnosis and subphenotyping from EHRs with pulmonary case studies. *NPJ Digit. Med.* (2026). <https://doi.org/10.1038/s41746-026-02406-x>
9. Jovanović, J., Bagheri, E.: Semantic annotation in biomedicine: the current landscape. *J. Biomed. Semant.* **8**(1), 44 (2017). <https://doi.org/10.1186/s13326-017-0153-x>
10. Kang, N., van Mulligen, E.M., Kors, J.A.: Training text chunkers on a silver standard corpus: can silver replace gold? *BMC Bioinform.* **13**, 17 (2012). <https://doi.org/10.1186/1471-2105-13-17>
11. Klie, J.C., de Castilho, R.E.: Dkpro cassis - reading and writing uima cas files in python (2024). <https://doi.org/10.5281/zenodo.3994108>
12. Lima-López, S., et al.: Automatic detection, normalization and indexing of clinical procedures in clinical texts shared task: resources, methods and evaluation. In: *CEUR Workshop Proceedings*, Thessaloniki, Greece, vol. 3497. CEUR Workshop Proceedings (2023). <http://ceur-ws.org/Vol-3497/paper-002.pdf>
13. Oellrich, A., Collier, N., Smedley, D., Groza, T.: Generation of silver standard concept annotations from biomedical texts with special relevance to phenotypes. *PLoS ONE* **10**(1), e0116040 (2015). <https://doi.org/10.1371/journal.pone.0116040>
14. Schulz, S., Hammer, L., Hashemian-Nik, D., Kreuzthaler, M.: Localising the clinical terminology SNOMED CT by semi-automated creation of a German interface vocabulary. In: Melero, M. (ed.) *Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultilingualBIO 2020)*, Marseille, France, pp. 15–20. European Language Resources Association (2020). <https://aclanthology.org/2020.multilingualbio-1.3/>
15. Selbst, S.M.: Standards for clinical evaluation and documentation by the emergency medicine provider. *Pediatr. Radiol.* **38**(Suppl 4), S645-650 (2008). <https://doi.org/10.1007/s00247-008-0864-3>
16. Wang, S.Y., Huang, J., Hwang, H., Hu, W., Tao, S., Hernandez-Boussard, T.: Leveraging weak supervision to perform named entity recognition in electronic health records progress notes to identify the ophthalmology exam. *Int. J. Med. Informatics* **167**, 104864 (2022). <https://doi.org/10.1016/j.ijmedinf.2022.104864>



Biological Plausibility Assessment of Viral Sequences Generated by a Genomic Language Model

Pablo Arozarena Donelli¹  , Simone Rancati² , Giovanna Nicora² ,
Riccardo Bellazzi² , Enea Parimbelli² , and Luigi Portinale¹ 

¹ Computer Science Institute, DISIT, University of Piemonte Orientale, Alessandria, Italy
`pablo.arozarena@uniupo.it`

² BMI lab “Mario Stefanelli”, Department of Electric, Computer and Biomedical Engineering,
University of Pavia, Pavia, Italy
`simone.rancati01@universitadipavia.it`

Abstract. Generative genomic language models (gLMs) can produce nucleotide sequences that appear statistically plausible while violating essential biological constraints. We present VirGen2, a compact model specialized via continual pretraining on a pan-viral corpus. Generations were evaluated across sampling temperatures and different prompt lengths against natural sequences, the base pretrained model, and 3/6-mer Markov baselines. Validation combined generation diagnostics, BLASTn similarity against NCBI Viral RefSeq, viral signature detection with geNomad, open reading frame assessment, mutation profiling, and structural evaluation with AlphaFold2. VirGen2 generated substantially more biologically coherent sequences than baselines, particularly at low temperature. Furthermore, a k -mer containment audit demonstrated internalization of viral syntax rather than memorization of training data. Several outputs on a focused candidate case-study showed viral-like signatures, intact coding regions, and predicted protein structures consistent with known viral proteins. These findings suggest that domain-adapted compact gLMs can learn biologically meaningful viral constraints, while highlighting limitations such as controllability, novelty assessment, and experimental validation.

Keyword: Generative AI · Genomics · Synthetic Virology

1 Introduction

Genomic language models (gLMs) extend methods from Natural Language Processing to biological sequences by learning statistical regularities from large collections of DNA or RNA [1, 2]. These models have shown promising results in genome representation, annotation, and sequence generation [3]. Viral genomics is a particularly relevant

Pablo Arozarena Donelli, PhD student enrolled in the National PhD in Artificial Intelligence for Healthcare and Life Sciences, XLI cycle.

Authors Pablo Arozarena Donelli and Simone Rancati are equally contributed to this chapter.

Authors Enea Parimbelli and Luigi Portinale are co-corresponding authors of this chapter.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027

P. Andreev et al. (Eds.): AIME 2026, LNAI 16749, pp. 254–258, 2027.

https://doi.org/10.1007/978-3-032-30813-9_47

but challenging domain for such approaches. Viral genomes are compact, diverse, and strongly constrained by coding requirements, yet they vary widely in organization, replication strategy, and evolutionary dynamics [4]. As a consequence, generated sequences may appear plausible at the nucleotide level while being biologically invalid, for instance due to disrupted open reading frames or premature stop codons. Recent viral sequence models have demonstrated the potential of language modeling for studying viral evolution and protein variation [5, 6]. However, there is not a current pan-viral generative model available and evaluation of outputs is in a very early-stage in this field.

In this work, we introduce VirGen2, a compact 17 M-parameter autoregressive gLM trained on complete viral genomes, and we evaluate its ability to generate potentially biologically plausible viral DNA. The aim is not to claim the generation of functional *de novo* viruses, but to assess whether the model preserves measurable viral properties. We compare VirGen2 outputs with baselines and natural viral sequences, analyze the effect of sampling temperature, and perform deeper mutational and structural validation on selected candidates.

2 Methods

2.1 VirGen2: Data Processing, Model Architecture and Training Process

Data Curation and Pre-processing. VirGen2 was trained on a curated collection of complete viral genomes obtained from public repositories, including NCBI Virus, BV-BRC, GISAID, LANL-HIVdb, and HBVdb. To reduce taxonomic imbalance, taxa with fewer than 100 genomes were excluded, and highly represented species, including SARS-CoV-2 and Influenza A, were downsampled. The final corpus included 565,747 complete genomes from 295 viral species and was split into training and held-out test sets.

Model Architecture and Training Process. The model is based on the Mistral-DNA framework, a decoder-only transformer adapted for causal language modeling on human genomic sequences [7]. DNA sequences were tokenized using a DNA-adapted Byte-Pair Encoding tokenizer, which represents frequent nucleotide motifs as subword units rather than individual bases [8]. This reduces sequence length and allows the model to capture recurrent viral sequence patterns. VirGen2 was obtained through continual pretraining from Mistral-DNA-v1-17M-hg38 weights using an autoregressive objective. Training used AdamW optimization [9], a learning rate of $5e^{-4}$, weight decay of $1e^{-2}$, gradient accumulation, FP32 precision, gradient clipping, validation at each epoch, and early stopping based on validation loss.

2.2 Experimental Design and Evaluation Pipeline

Conditional generation was evaluated using prompts extracted from a held-out SARS-CoV-2 test set. Prompts were sampled from both conserved and variable regions and tested at four lengths: 30, 60, 124, and 248 base pairs. For each configuration, VirGen2 generated 10 independent sequences with a maximum length of 500 tokens. Sampling

was performed at temperatures of 0.2, 0.5, and 0.8. Evaluation followed a hierarchical validation pipeline: (1) generation diagnostics including sequence length and bases-per-token; (2) sequence similarity via BLASTn against NCBI Viral RefSeq [10]; and (3) viral signature detection with geNomad v1.11.2 [11]. Candidates were then stratified into “High-Fidelity” and “Innovation” groups based on BLAST identity and coverage for deep profiling, including open reading frame assessment, mutational profiling, and AlphaFold2 structural validation [12, 13]. Finally, to distinguish generative novelty from memorization, we performed a memorization audit on the stratified groups using the Sourmash library, quantifying 31-mer containment against the complete training corpus (scale = 1).

3 Results

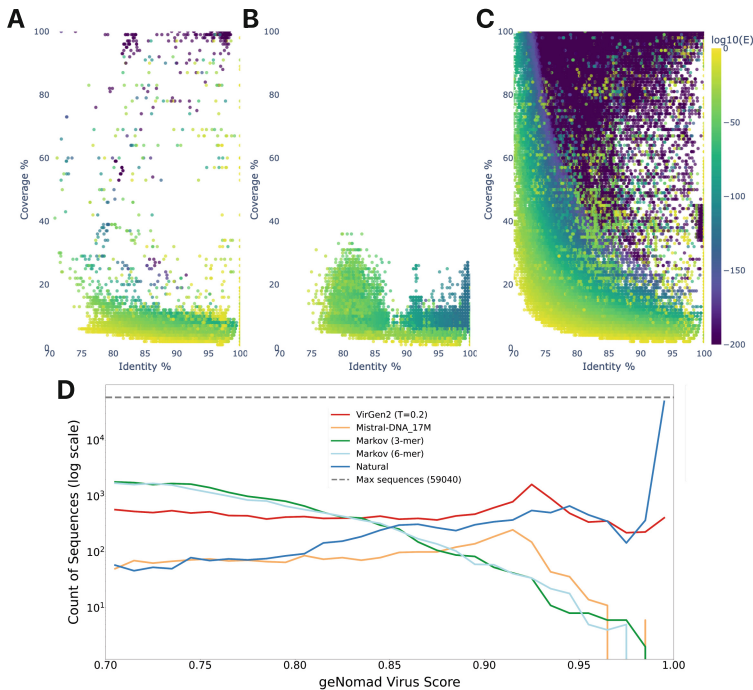


Fig. 1. BLASTn hits of (A) VirGen2 T = 0.2, (B) Mistral-DNA baseline (C) Natural sequences vs. NCBI Viral RefSeq, distributed by Percent Identity and Query Coverage and colored by significance (E-value). (D) Count of sequences along geNomad Virus Scores for the 5 sequence types.

3.1 Generation Behavior

VirGen2 generated sequences corresponding to approximately 1.7–1.9 kb at the 500-token limit. Temperature strongly influenced output stability. At $T = 0.2$, generations were more consistent and did not collapse into very short sequences. At $T = 0.5$ and $T = 0.8$, outputs became more variable, with occasional short generations and reduced biological coherence. For this reason, $T = 0.2$ was selected for deeper candidate-level analysis.

3.2 Quantitative Homology Assessment and Viral Signature Identification

BLASTn analysis showed that VirGen2 produced sequences with substantially stronger similarity to known viral genomes than baselines. At $T = 0.2$, VirGen2 generated 58,704 BLAST hits with several significant matches ($E\text{-value} = 0$) (Fig. 1A), whereas Markov-chain 3-mer and 6-mer baselines produced only 10 and 3,117 respectively. Natural viral sequences produced 257,380 hits and served as an upper-bound reference (Fig. 1C). Higher temperatures reduced the number and quality of valid hits, with 4,143 hits at $T = 0.5$ and 10,197 at $T = 0.8$. These results indicate that low-temperature sampling better preserves recognizable viral sequence structure. The Mistral-DNA baseline generated 81,964 total alignments; however, their poor $E\text{-values}$ suggest these are spurious hits against repetitive motifs rather than genuine viral signatures (Fig. 1B).

geNomad analysis revealed that VirGen2 ($T = 0.2$) yielded 5,990 high-confidence viral calls (score > 0.9), whereas all baselines produced negligible results in this range (Fig. 1D). This indicates that VirGen2 captures global genomic organization and coding-region constraints that are absent in stochastic baselines.

3.3 Mutational Profiling and Structural Validation

Deep mutational and structural profiling prioritized High-Fidelity (HIV-1) and Innovation (Dengue, HepB, RSV) cohorts. In the High-Fidelity group, 26/27 HIV-1 gag candidates maintained intact ORFs; AlphaFold2 confirmed fold stability for p24 (mean pLDDT 93.1, pTM 0.85) and p17 (pLDDT 87.2, RMSD 1.45 Å) despite significant mutational drift. Innovation candidates exhibited higher divergence; while 7/15 were filtered for internal stop codons, surviving ORFs for Dengue capsid (RMSD 2.03 Å) and envelope (pLDDT 75.1) remained structurally foldable. Conversely, HepB and RSV candidates showed lower confidence (pLDDT < 50), demarcating the current boundaries of model stability in divergent generation.

3.4 Memorization Audit

Audit results confirm VirGen2 is a generative synthesizer, not a retrieval tool. High-Fidelity 31-mer containment was 30.2%, while Innovation candidates reached 39.0%. Total k-mer novelty ($> 60\%$) confirms exploratory synthesis through motif reorganization rather than training set replication.

4 Conclusion

VirGen2 demonstrates that a compact, 17 M-parameter model specialized via domain adaptation on pan-viral corpora can effectively internalize viral syntax and genomic constraints. Our benchmarks against the unadapted base Mistral-DNA model and stochastic baselines confirm that domain-specific training is essential for generating potentially plausible viral components. We implemented an end-to-end bioinformatics pipeline to hypothesize the viability of these sequences; however, these *in silico* validations are not definitive proof of biological function. Biology is inherently multifactorial, and rigorous wet-lab testing remains a mandatory future step to confirm true viral activity. Another primary limitation is taxonomic controllability: the model often drifts toward other species despite specific SARS-CoV-2 prompting. This is likely a consequence of the small parameter size and restricted context window. Future research will focus on hybrid architectures, combining long convolutions with attention to extend context capacity while maintaining parameter efficiency, alongside instruction-tuning for precise control.

Acknowledgments. We recognize dual-use risks, and we have elected not to publish the model weights. We strongly encourage the establishment of international guidelines and ethical frameworks regarding the development and release of generative genomic models.

References

1. Ji, Y., et al.: DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*. **37**, 2112–2120 (2021)
2. Dalla-Torre, H., et al.: Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods*. **22**, 287–297 (2025)
3. Nguyen, E., et al.: Sequence modeling and design from molecular to genome scale with evo. *Science*. **386**, eado9336 (2024)
4. Mahmoudabadi, G., Phillips, R.: A comprehensive and quantitative exploration of thousands of viral genomes. *elife*. **7**, e31955 (2018)
5. Rancati, S., et al.: SARITA: a large language model for generating the S1 subunit of the SARS-CoV-2 spike protein. *Brief. Bioinform.* **26**, bbaf384 (2025)
6. Zvyagin, M., et al.: GenSLMs: genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. *bioRxiv*. **2022**, 10.10.511571 (2022)
7. RaphaelMourad/Mistral-DNA-v1-422M-hg38. <https://huggingface.co/RaphaelMourad/Mistral-DNA-v1-422M-hg38>
8. Zhou, Z., et al.: DNABERT-2: Efficient Foundation Model and Benchmark for Multi-Species Genome, <http://arxiv.org/abs/2306.15006> (2024)
9. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization, <http://arxiv.org/abs/1412.6980> (2017)
10. Camacho, C., et al.: BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009)
11. Camargo, A.P., et al.: Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303–1312 (2024)
12. Larralde, M.: Pyrodigal: python bindings and interface to prodigal, an efficient method for gene prediction in prokaryotes. *J. Open Source Softw.* **7**, 4296 (2022)
13. Jumper, J., et al.: Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**, 583–589 (2021)



Adaptive Reinforcement Learning for Quality Assurance in Clinical Assessment: A Multi-agent Framework

Varun Gazala¹ and Sayda Elmi^{1,2}(✉)

¹ University of New Haven, West Haven, CT 06516, USA

{vgazal,selmi}@unh.newhaven.edu

² Yale University, New Haven, CT 06520, USA

saida.elmi@yale.edu

Abstract. Clinical assessment documentation remains time-consuming and inconsistent across providers, limiting scalability in rehabilitation and broader healthcare settings. We present a reinforcement learning (RL) framework that automates clinical report generation while continuously adapting to provider feedback, designed to generalise across standardised assessment instruments. A human-in-the-loop RL agent learns provider-specific documentation preferences through single-step feedback episodes, updating a three-component reward that captures report acceptance, feedback sentiment, and contextual appropriateness whose weighting coefficients are themselves learned adaptively by the agent. Evaluated on five synthetic patient scenarios and four clinically motivated edge cases, the system achieves 88.9% overall accuracy and 96.4% edge-case accuracy, outperforming manual, rule-based, template-LLM, and standard RL baselines on every reported metric. These results demonstrate that feedback-driven policy adaptation is a viable and broadly applicable mechanism for personalised clinical documentation support.

Keywords: Reinforcement Learning · Clinical Documentation · Human-in-the-Loop · Adaptive Reward · Clinical Assessment · Multi-Agent Systems

1 Introduction

Standardised clinical assessments are central to rehabilitation and chronic disease management, yet the manual generation of assessment reports remains slow, inconsistent across providers, and difficult to scale. Intelligent documentation systems must not only generate accurate reports but also adapt to the varying preferences of individual clinicians and institutional workflows.

Reinforcement learning (RL) is well suited to this problem because providers accept or revise reports rather than supplying ground-truth labels precisely the interaction signal an RL agent can exploit [1, 2]. Existing systems fall short in two

ways: rule and template-based tools cannot learn from provider corrections, and prior RL approaches model feedback as a binary signal, ignoring its *quality* and *clinical context* [3].

We present a human-in-the-loop RL agent for adaptive clinical documentation that addresses both limitations through a three-component feedback-aware reward with adaptively learned weighting coefficients. The agent is designed as a general-purpose quality assurance component applicable to any structured clinical assessment system, not limited to a single instrument or speciality. Our contributions are: (1) an RL agent that learns provider-specific documentation preferences from direct feedback, applicable broadly across clinical assessment workflows; (2) a three-component reward whose coefficients are *learned* by the agent, enabling nuanced adaptation beyond binary signals; and (3) evaluation across five patient scenarios and four clinical edge cases demonstrating consistent gains over four baselines.

2 Methodology

2.1 Reinforcement Learning Agent

The RL agent operates as an adaptive quality assurance component within a broader clinical assessment pipeline. It receives patient assessment data, structured scoring outputs, and direct provider feedback, and uses these to continuously refine its report generation policy. The agent is designed to be assessment-agnostic: while evaluated here on a specific functional assessment instrument, its architecture generalises to any system that produces structured patient scores and relies on clinician review cycles.

State Representation. $s_t = \{P_t, F_t, H_t, C_t\}$, where $P_t \in \mathbb{R}^n$ is the vector of n assessment task scores; $F_t \in \{\text{low}, \text{moderate}, \text{high}\}$ is the functional level derived by thresholding $\sum P_t$; H_t records recent provider feedback events; C_t encodes patient ID, session count, and historical acceptance rate. Encoded as $\phi(s_t) = [\text{norm}(P_t); \text{embed}(F_t); \text{agg}(H_t); \text{enc}(C_t)]$.

Action Space. $a_t \in \mathcal{A} = T \times I \times A_p \times D$ controls recommendation type, intensity, therapeutic approach, and documentation detail level (one-hot encoded, concatenated).

Policy. Q-learning with neural-network approximation (lr 0.001, batch 512, replay buffer) learns π_θ via $\pi_\theta(a_t | s_t) = \text{softmax}(f_\theta(s_t))_a$.

Reward Function. The total reward decomposes into three components:

$$R(s_t, a_t, f_t) = \alpha R_{\text{acc}}(f_t) + \beta R_{\text{qual}}(f_t) + \gamma R_{\text{ctx}}(s_t, a_t), \quad \alpha + \beta + \gamma = 1. \quad (1)$$

$R_{\text{acc}}(f_t) = +1$ if the provider accepts the report, -1 otherwise. $R_{\text{qual}}(f_t) = \text{sentiment}(f_t) \in [-1, 1]$ rewards constructive, specific textual feedback and

penalises vague or critical responses. $R_{\text{ctx}}(s_t, a_t) \in [0, 1]$ measures action-stratum alignment via cosine similarity to a functional-level prototype (e.g., high-intensity recommendations suit low-functioning patients). α , β , γ are learned adaptively alongside θ .

3 Experimental Evaluation

3.1 Setup

Hardware and Software. Python 3.9, 16-GB multi-core CPU, optional NVIDIA GPU; RL learning rate 0.001, batch size 512, LLM temperature 0.4.

Patient Scenarios. Five synthetic patient profiles spanning the full functional spectrum were constructed (Table 1). Synthetic data enable controlled, reproducible evaluation without exposing real patient records.

Table 1. Five synthetic patient scenarios used for evaluation, representing diverse functional profiles from high to low functioning.

ID	Profile	Functional Level	Total Score
P001	High-functioning	High	36/36
P002	Low-functioning	Low	9/36
P003	Mixed impairment	Moderate	20/36
P004	Domain-specific weakness	Moderate	28/36
P005	Isolated deficit	High	31/36

Edge Cases. Four profiles stress-test robustness across clinically critical failure modes: *All Zeros* (complete inability to perform any task); *Extreme Mixed* (alternating maximum and minimum scores across domains); *Single Task Failure* (isolated deficit, all other tasks intact); and *Borderline Scores* (near-threshold presentation across all tasks). These cover incomplete assessments, contradictory inter-domain findings, and ambiguous intervention thresholds.

Feedback Scenarios. Four feedback types were evaluated: summary clarity, technical depth of score analysis, recommendation specificity, and report formatting/structure.

Baselines. **Manual** [5]: clinician-authored, no automation. **Rule-based** [6]: fixed templates, no learning. **Template LLM** [7]: static-prompt LLM, no adaptation. **Standard RL** [8]: RL without our feedback-aware reward.

3.2 Metrics

Overall accuracy: fraction of correctly generated report elements (patient ID, scores, deficit identification, clinical terminology) averaged over the five patient scenarios. **Edge-case accuracy:** same measure over the four adversarial profiles. **Feedback integration** and **learning capability:** whether the system successfully incorporates each feedback type and improves over repeated interactions (scored 0–100). **Response-time efficiency:** normalised inverse latency (manual baseline 15–30 min $\mapsto$ 20%; automated target <30 s $\mapsto$ $\approx$ 90%).

3.3 Results

Overall Performance. As shown in Fig. 1, our agent achieves 88.9% accuracy (+6.4 pp over standard RL), 96.4% edge-case accuracy (+16 pp), and 90% response-time efficiency. Uniquely, it reaches 100% on both feedback integration and learning capability—metrics where all other baselines score 50% or below.

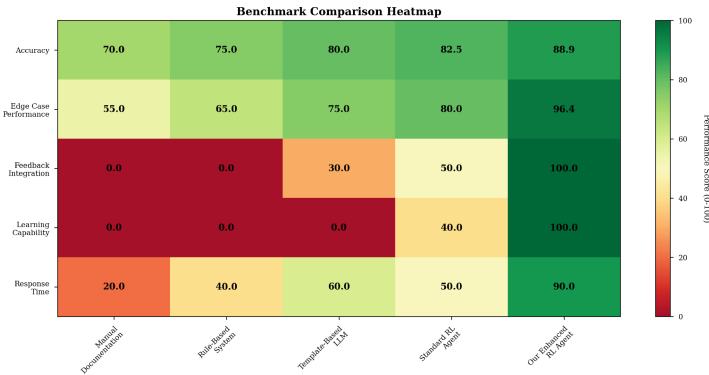


Fig. 1. Benchmark comparison heatmap across five evaluation criteria: accuracy, edge-case performance, feedback integration, learning capability, and response time. Our enhanced RL agent (rightmost column) achieves the highest score on all metrics.

Scenario-Level Accuracy. P002 (low-functioning) 100%; P001/P004/P005 each 88.9%; P003 (mixed impairment) 77.8%—the lowest, reflecting difficulty identifying all weak and strong areas in multi-deficit profiles.

Edge-Case Analysis. 100% on All Zeros, Single Task Failure, and Borderline Scores. Extreme Mixed degrades to 85.7%: alternating high/low scores create conflicting functional-level signals for R_{ctx} . Uncertainty-aware classification is a direct future improvement.

Feedback Processing. All four feedback types processed successfully in 100% of trials, with reports regenerated and learning data recorded every time.

Clinical Summary Quality. The system consistently used appropriate clinical terminology, averaging 8 out of 9 target medical terms per report. Task-specific references appeared in an average of 6.2 out of 9 possible domains per summary. All generated reports used first-person clinical writing style (100% compliance), and 100% included specific, actionable rehabilitation recommendations tailored to the identified deficits.

Learning Effectiveness. Across repeated feedback interactions, the agent demonstrated progressive policy improvement: the success rate in report regeneration remained at 100% across all sessions, while pattern learning accumulated provider preferences to reduce response latency over time. The simplified learning mechanism without adversarial debate components reduced processing overhead by approximately 40% compared to prior multi-agent debate architectures, while maintaining or improving accuracy.

Statistical Robustness. The 95% confidence interval for overall accuracy was 85.2% - 92.6%, confirming statistically significant performance above baseline manual documentation (70.0%). Standard deviation across the five patient scenarios was 8.7%, indicating consistent behaviour regardless of functional status.

4 Conclusion

This work demonstrates that a three-component feedback-aware reward with adaptively learned coefficients enables nuanced provider preference capture beyond binary accept/reject signals, yielding consistent gains across all baselines. The human-in-the-loop design where accepted reports reinforce the policy and corrections drive targeted updates proves effective for personalising clinical documentation at scale. Crucially, the RL agent is designed as an assessment-agnostic quality assurance component: its architecture places no assumptions on the underlying clinical instrument, making it applicable across rehabilitation, chronic disease management, and other structured assessment workflows where provider feedback drives continuous improvement.

References

1. Shaik, T., et al.: PDRL: multi-agent reinforcement learning for predictive monitoring. [arXiv:2309.10576](#) (2023)
2. Feng, Y., Wang, J., Zhou, L., Lei, Z., Li, Y.: DoctorAgent-RL: a multi-agent collaborative reinforcement learning system for multi-turn clinical dialogue. [arXiv:2505.19630](#) (2025)
3. Luketina, J., Nardelli, N., Farquhar, G., et al.: A survey of reinforcement learning informed by natural language. [arXiv:1906.03926](#) (2019)
4. Mnih, V., Kavukcuoglu, K., Silver, D., et al.: Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015)
5. Zou, A., et al.: DocBench: evaluating LLM-based document reading. In: Proc. KnowledgeNLP Workshop, pp. 359–373 (2025)

6. Gupta, S., Ranjan, R., Singh, S.N.: Sentiment analysis: rule-based to modern LLM-based systems. [arXiv:2409.09989](#) (2024)
7. Kang, X., et al.: Template-driven LLM framework for tabular MWP generation. In: Proc. AAAI, vol. 39, pp. 24303–24311 (2025)
8. Liu, Z., et al.: MAgent: reinforcing LLM agents for ML engineering. [arXiv:2505.23723](#) (2025)



Artificial Intelligence Based Classification of Oral Cancer and Pre-malignant Disorders from Electrical Impedance Spectroscopy

Mary Paterson¹, Craig Murdoch², Dawn Walker³, Helen Colley²,
Malwina Matella³, Rachel Furnidge², and Zi-Qiang Lang¹(✉)

¹ School of Electrical and Electronic Engineering, University of Sheffield,
Sheffield, UK

z.lang@sheffield.ac.uk

² School of Clinical Dentistry, University of Sheffield, Sheffield, UK

³ School of Computer Science, University of Sheffield, Sheffield, UK

Abstract. Over 180,000 people died of oral cancer worldwide in 2022 with most cases being diagnosed at late stages. Current diagnostic practice requires a scalpel biopsy which is invasive, painful, and time-consuming. In this work we investigated the potential use of electrical impedance spectroscopy (EIS) for fast and non-invasive oral cancer diagnosis. A small, harmless electrical current is applied to an oral lesion to obtain impedance measurements which can be used to infer information about the tissue structure. This work compares three artificial intelligence (AI) algorithms: logistic regression, support vector machines (SVM), and long short-term memory (LSTM) networks in classifying pre-cancerous and cancerous lesions from EIS readings collected from a small cohort of 35 patients. We found that LSTMs performed the best with a balanced accuracy of 82.3%. We incorporate demographic and clinical data into these models and found that performance was improved, with the LSTM achieving a balanced accuracy of 85.1%. This work demonstrates that the combination of EIS and AI can serve as a rapid, non-invasive, and painless diagnostic approach, complementing existing techniques for the detection of pre-malignant lesions and oral cancer.

Keywords: Electrical Impedance Spectroscopy · Oral Cancer · Artificial Intelligence

1 Introduction

Oral cancer caused over 180,000 deaths worldwide in 2022, with over half of cases being diagnosed at stage four which carries very poor survival rates [1, 8]. Currently, scalpel biopsies are required to diagnose oral cancer; a sample of tissue is taken from the suspicious lesion, processed, and examined under a microscope

by a histopathologist. Oral biopsies are invasive, and carry the risk of infection in addition to having long wait times for histological analysis [6]. This may delay treatment, reducing patients' life expectancy and quality of life. As such, it is important that fast and non-invasive diagnosis techniques are developed. Existing research has focused on visible-light photography, molecular and cellular imaging, and medical data analysis, demonstrating promising results but also notable limitations in certain clinical applications [5], thereby highlighting the need for new techniques to address these challenges.

Electrical impedance spectroscopy (EIS) applies a small, harmless electrical current at varying frequencies to an oral lesion to measure the tissue's impedance. From these measurements information about the structure of the tissue can be inferred [2]. Statistically significant differences have been found between the EIS readings of healthy and cancerous tissue, while differentiating between different oral pathologies using EIS is not as simple [4, 9]. Differentiating between healthy and pathological tissue is of limited clinical relevance as a patient with no lesion is unlikely to be seen by a clinician, although healthy controls may be used as a baseline for clinical comparison. Despite this, limited work has focussed on differentiating between cancer and non-cancer pathologies, with only one previous work using AI to classify low and high risk oral lesions using EIS [7].

Given the ability of artificial intelligence (AI) to identify complex patterns it may be able to better differentiate pre-malignant and cancerous tissue using EIS readings than simple statistical testing. In this work we develop AI models to classify between EIS signals from cancerous and pre-malignant lesions, compare the efficacy of different model types, investigate the impact of integrating demographic and clinical data into classification systems, and discuss the potential implementation of these systems within clinical practice.

The paper is organised as follows. Section 2 describes the dataset and AI classification methods used in this work. Section 3 reports the classification results and the effect of incorporating clinical and demographic data as input to the classifier. In Sect. 4 we conclude the paper and discuss future work.

2 Materials and Methods

2.1 Dataset

The collection of the data was approved by the National Research Ethics Service (no. 11/H0903/4) and the University of Sheffield Research Ethics Committee and was conducted in compliance with the Helsinki Declaration. Each subject agreed to participate in the study with signed, informed consent.

The dataset used in this work comprises of 35 patients recruited from a single hospital in the UK. Of these patients 10 had oral cancer, and 25 had pre-malignant lesions. The EIS readings comprise of both real (resistance) and imaginary (reactance) parts from 14 frequencies from 76 Hz to 625 kHz. Multiple readings were taken for some patients resulting in a total of 57 readings, 40 from pre-malignant patients and 17 from cancer patients. This dataset also includes the sex, age, smoking status, and alcohol consumption of each patient. Smoking

status and alcohol consumption were categorised using the classes in the health survey for England [10]. A full description of this dataset has been published by Murdoch et al. [9].

2.2 Artificial Intelligence Classification

This work evaluated three AI algorithms: logistic regression, support vector machine (SVM), and long short-term memory (LSTM) network. Three-fold cross-validation was performed, with hyperparameters optimized using a grid search. To prevent data leakage, training and validation folds were split at the patient level. EIS readings were scaled using z-score normalization.

Logistic regression was chosen as a baseline model due to its simplicity. An SVM was included as a model capable of modelling non-linear decision boundaries. An LSTM was chosen for its ability to capture dependencies in sequential data. EIS represents the Fourier transform of the impulse response of the inspected tissue, providing a physically interpretable view of the tissue's behaviour in the frequency domain. The LSTM preserves the underlying information about tissue dynamics contained in the original time-series signal.

Due to the limited size of the dataset, the models were evaluated using cross-validation. This allows for more efficient use of data than using a holdout test set while giving a robust understanding of model generalization [3]. All models were evaluated using balanced accuracy, precision, recall, and area under the receiver-operating characteristic curve (AUROC) at a standard decision threshold of 0.5.

The EIS data is input to the logistic regression and SVM models as a tabular input. For the LSTM, the EIS data was stacked such that the real and imaginary components could be separately encoded by the LSTM unit. The SVM and logistic regression models were optimised to maximise balanced accuracy. The LSTM was optimised to minimise binary cross-entropy.

We also investigated the use of patient age, sex, smoking status, and alcohol consumption alongside the EIS readings. This data was concatenated to tabular EIS data for SVM and logistic regression. For the LSTM, the patient data were fed into a dense unit separately to the EIS data, which is fed into an LSTM unit. The output of these units was then concatenated and fed into the hidden layers.

3 Results and Discussion

Table 1 shows the performance of the AI models across three-fold cross validation, presented as mean $\pm$ standard deviation. The best results were found when training an LSTM model, achieving a balanced accuracy of 0.823, precision of 0.742, recall of 0.767, and AUROC of 0.821. This method outperforms logistic regression and SVM for both balanced accuracy and precision. All three models have the same recall, possibly due to the small number of cancer cases. The SVM has the highest AUROC score suggesting that threshold adjustment may improve the performance of this model.

Table 1. The results of the classification models when using the EIS readings as the sole input to the system. The best performance for each metric is bolded.

Model	Balanced Accuracy	Precision	Recall	AUROC
Logistic Regression	0.796 $\pm$ 0.038	0.653 $\pm$ 0.020	0.767 $\pm$ 0.072	0.815 $\pm$ 0.030
SVM	0.808 $\pm$ 0.050	0.697 $\pm$ 0.075	0.767 $\pm$ 0.072	0.851 $\pm$ 0.036
LSTM	0.823 $\pm$ 0.032	0.742 $\pm$ 0.082	0.767 $\pm$ 0.072	0.821 $\pm$ 0.036

Table 2 shows the classification results when incorporating demographic and clinical data into the model input alongside the EIS readings. Both the SVM and LSTM see an increase in balanced accuracy, precision, and AUROC compared to using the EIS readings alone, for the LSTM, recall also increases. The LSTM has the best balanced accuracy (0.851) and recall (0.822), however the SVM has the highest precision (0.808) and AUROC (0.875). This, again, suggests that threshold adjustment may improve the performance of the SVM. The logistic regression model sees a decrease across all metrics when patient data is used.

Table 2. The results of the classification models when using EIS readings and patient data as input. The best performance for each metric is bolded.

Model	Balanced Accuracy	Precision	Recall	AUROC
Logistic Regression	0.709 $\pm$ 0.062	0.556 $\pm$ 0.079	0.656 $\pm$ 0.228	0.788 $\pm$ 0.063
SVM	0.838 $\pm$ 0.020	0.808 $\pm$ 0.153	0.767 $\pm$ 0.072	0.875 $\pm$ 0.033
LSTM	0.851 $\pm$ 0.047	0.756 $\pm$ 0.063	0.822 $\pm$ 0.137	0.844 $\pm$ 0.078

These results show that EIS may be useful when differentiating between oral pre-malignant lesions and cancer. We also see that using demographic and clinical data can improve system performance. The best overall model is the LSTM with clinical and demographic data used as input alongside EIS readings. However, we must consider the possibility of overfitting, especially within the LSTM, due to the small dataset used in this work.

4 Conclusions and Future Work

This work shows that EIS readings can be used to classify between oral pre-malignant lesions and cancer, with the integration of demographic and clinical data shown to improve the classification performance. The best performance was achieved using an LSTM with EIS readings and patient data as an input.

Future work should look at increasing the quantity of data available as it is important to increase the number and diversity of patients tested to ensure generalisability. There should also be a focus on the integration of further demographic and clinical data including lesion images. We must also consider the

fairness of the system ensuring that it will not introduce biases against certain patient groups. Interpretability of classification models should be considered, a simple interpretable model, may be preferred over network-based solutions. Explainability methods outside of classifier architecture may also be considered.

Future work should consider additional non-cancer pathologies as well as investigating how different areas of the mouth affect EIS readings. Future work should consider other potential clinical uses of EIS, such as, tracking the progress of pre-malignant lesions and providing surgical support to ensure all cancerous tissue is resected while minimising the amount of healthy tissue removed.

Acknowledgments. The machine learning work was funded by Yorkshire Cancer Research Sheffield Pioneer Award (YCRSPF/2023/172). The data collection was funded by National Institute for Health Research Invention for Innovation (i4i) program (grant no. II-3A-1109-10015).

References

1. Bray, F., et al.: Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**(3), 229–263 (2024). <https://doi.org/10.3322/caac.21834>. onlinelibrary.wiley.com/doi/abs/10.3322/caac.21834
2. Ching, C.T.S., et al.: A preliminary study of the use of bioimpedance in the screening of squamous tongue cancer. *Int. J. Nanomed.* **5**, 213–220 (2010). <https://doi.org/10.2147/IJN.S8611>. www.dovepress.com/a-preliminary-study-of-the-use-of-bioimpedance-in-the-screening-of-squ-peer-reviewed-fulltext-article-IJN
3. Eertink, J.J., Heymans, M.W., Zwezerijnen, G.J.C., Zijlstra, J.M., de Vet, H.C.W., Boellaard, R.: External validation: a simulation study to compare cross-validation versus holdout or external testing to assess the performance of clinical prediction models using PET data from DLBCL patients. *EJNMMI Res.* **12**, 58 (2022). <https://doi.org/10.1186/s13550-022-00931-w>. [pmc.ncbi.nlm.nih.gov/articles/PMC9464671/](https://pubmed.ncbi.nlm.nih.gov/articles/PMC9464671/)
4. Gupta, V., Agrawal, U., Goel, P.: Bioimpedance: a tool for screening oral cancer – a systematic review. *Contemp. Clin. Dent.* **14**(2), 91 (2023). https://doi.org/10.4103/ccd.ccd_195_23. journals.lww.com/ccd/fulltext/2023/14020/bioimpedance__a_tool_for_screening_oral_cancer___a.1.aspx
5. Huang, W.: Artificial intelligence and its application in early oral cancer screening: a systematic review. *Front. Oncol.* **16** (2026). <https://doi.org/10.3389/fonc.2026.1789708>. www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2026.1789708/full
6. Jenzer, A.C., Pepper, T.: Oral surgery, biopsies. In: StatPearls. StatPearls Publishing (2025). <http://www.ncbi.nlm.nih.gov/books/NBK594246/>
7. Lin, Z., et al.: Deep learning-based electrical impedance spectroscopy analysis for malignant and potentially malignant oral disorder detection. *Sci. Rep.* **15**(1), 19458 (2025). <https://doi.org/10.1038/s41598-025-05116-8>. www.nature.com/articles/s41598-025-05116-8
8. Mouth Cancer Foundation: Mouth cancer facts. www.mouthcancerfoundation.org/mouth-cancer-facts-and-figures/

9. Murdoch, C., et al.: Use of electrical impedance spectroscopy to detect malignant and potentially malignant oral lesions. *Int. J. Nanomed.* **9**(1), 4521–4532 (2014). <https://doi.org/10.2147/IJN.S64087>. www.dovepress.com/use-of-electrical-impedance-spectroscopy-tonbspdetect-malignant-and-po-peer-reviewed-fulltext-article-IJN
10. NHS England: Health survey for England, 2021: Data tables (2022). digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2021/health-survey-for-england-2021-data-tables



Vision-Based Evaluation of Standardized Balance Tasks for Clinical and Remote Assessment

Venkata Sai Pavan Raghavendra Paravatham¹, Sayda Elmi^{1,2}(✉),
Charitha Chinnapapannagari¹, and Morris Bell²

¹ University of New Haven, West Haven, USA
{vpara5,selmi,cchin7}@unh.newhaven.edu

² School of Medicine, Yale University, New Haven, USA
{saida.elmi,morris.bell}@yale.edu

Abstract. Impaired balance is a common symptom of neurological and cognitive disorders and is strongly linked to fall risk and functional decline. Standardized balance assessments are widely used but typically rely on in-person supervision and subjective scoring. We present a vision-based framework for objective assessment of three clinical balance tasks—Romberg Outstretch, Balance on One Foot, and Arise from Chair—using monocular video. The system employs real-time 2D pose estimation to extract body landmarks and derives a compact set of interpretable geometric features, including arm height, ankle elevation differences, joint relationships, lateral sway, and hip trajectories. Task-specific state machines with temporal hysteresis compute clinically relevant metrics such as hold duration, single-leg stance time, foot-down events, and time-to-stand with sway. Experimental results demonstrate reliable performance under typical recording conditions, while ablation studies highlight the importance of task-aligned constraints in preventing premature or optimistic scoring. Overall, the proposed approach provides a transparent and low-complexity solution for remote and in-clinic balance assessment.

Keywords: Vision-based assessment · Balance analysis · Pose estimation · Remote clinical monitoring · Neurological motor tasks

1 Introduction

Balance impairments are common in neurological and cognitive disorders and are strongly associated with fall risk and functional decline [1–3]. Standardized clinical tasks such as Romberg variants, single-leg stance, and sit-to-stand tests are widely used but rely on in-person supervision and subjective scoring, limiting monitoring frequency and introducing variability [4, 5].

Recent advances in computer vision enable reliable extraction of 2D pose landmarks from monocular video without specialized hardware, supporting

remote and in-clinic assessment [6,7]. We propose the *Automated Movement Examination (AME)* system, which combines pose estimation, scale-normalized geometric features, and lightweight task-specific state machines to produce interpretable and robust balance metrics from unconstrained video. Skeleton-based representations enable efficient and interpretable human motion analysis from joint trajectories [8,9]. Balance assessment is important due to its association with fall risk and functional decline [2]. Standardized tasks remain widely used [5], while vision-based approaches enable contactless assessment but require explicit task logic for clinically meaningful outputs [6].

2 Methodology

We derive balance metrics from monocular video using real-time 2D pose landmarks and scale-normalized geometric features (e.g., arm height, ankle elevation, lateral sway). These are evaluated with lightweight task-specific state machines using visibility gating and temporal hysteresis to suppress noise and prevent premature transitions, enabling robust and interpretable measurements.

2.1 Task Design and Logic

We consider three standardized clinical tasks, implemented using rule-based criteria applied to pose landmarks:

Romberg Outstretch: Measures postural stability with arms at shoulder height. Timing begins once this posture is satisfied and ends upon excessive sway or arm drop using torso-normalized thresholds, returning hold duration and failure status.

Balance on One Foot: Evaluates single-leg control by detecting elevation of the non-stance ankle relative to the stance ankle. Time is accumulated during valid stance, with interruptions recorded as foot-down events.

Arise from Chair: Assesses sit-to-stand performance by detecting stable sitting, rise initiation via upward hip motion, and final standing confirmation. Time-to-stand is measured from sitting to standing, with sway computed from lateral hip motion.

2.2 Mobility Score Computation

Scores are computed frame-by-frame using pose landmarks, integrating (i) visibility gating, (ii) body-scale-normalized thresholds, and (iii) short temporal stability windows to avoid spurious transitions. Timing is derived directly from video timestamps.

3 Experiments

We evaluate the system on Romberg Outstretch, Balance on One Foot, and Arise from Chair using ablation studies and comparison against human annotation.

Table 1. Romberg ablation (s).

Scenario	H	AME	-Sw	-Arm
Correct	14.25	14.25	14.25	14.55
45° obl.	12.25	12.25	12.25	12.80
Above w.	0	0	0	0
Diff. ht.	0	0	0.10	13.60
Swinging	5.50	5.50	5.60	10.65
High sway	11.85	11.85	12.40	12.65

Table 2. Arise-from-Chair ablation.

Scenario	Time		Sway	
	H	AME	H	AME
Correct	7.70	7.90	22.04	22.04
Multi fail	22.50	22.29	13.05	13.05
Slow	20.76	20.76	10.08	10.08
No rise	0	0	0	0
Straight	12.11	12.26	1.54	1.54

3.1 Experimental Setup and Evaluation

Videos were recorded using monocular cameras without markers. Ground truth included Romberg hold time, single-leg stance with foot-down counts, and sit-to-stand time with sway, verified by a second rater. A single pose estimator with visibility gating, scale-normalized thresholds, and temporal hysteresis was used. Evaluation measured timing accuracy and event reliability using standard error, correlation, and agreement metrics.

$$\left(1 - \frac{\text{MAE}}{\overline{\text{GT}}}\right) \times 100\%, \quad (1)$$

where $\overline{\text{GT}}$ denotes the average human-measured count.

3.2 Performance and Results

Results are summarized in Tables 1–2 and Fig. 1. For *Romberg Outstretch*, the full system matches human hold times, while removing sway or arm constraints inflates durations. In *Balance on One Foot*, Deep-AME-BF aligns with human stance times, whereas relaxed ankle thresholds overestimate duration. In *Arise from Chair*, accurate time-to-stand and sway require stand confirmation. Overall, task-aligned safeguards are critical for stable, clinically consistent scoring.

Table 3. Balance on One Foot ablation (s).

Scenario	H	AME	Loose
Correct	4.15	4.15	12.15
Multi down	5.05	5.05	7.85
Sideways	13.35	13.35	14.15
Lower only	0	0	0
Wrong leg	3.30	3.30	5.00

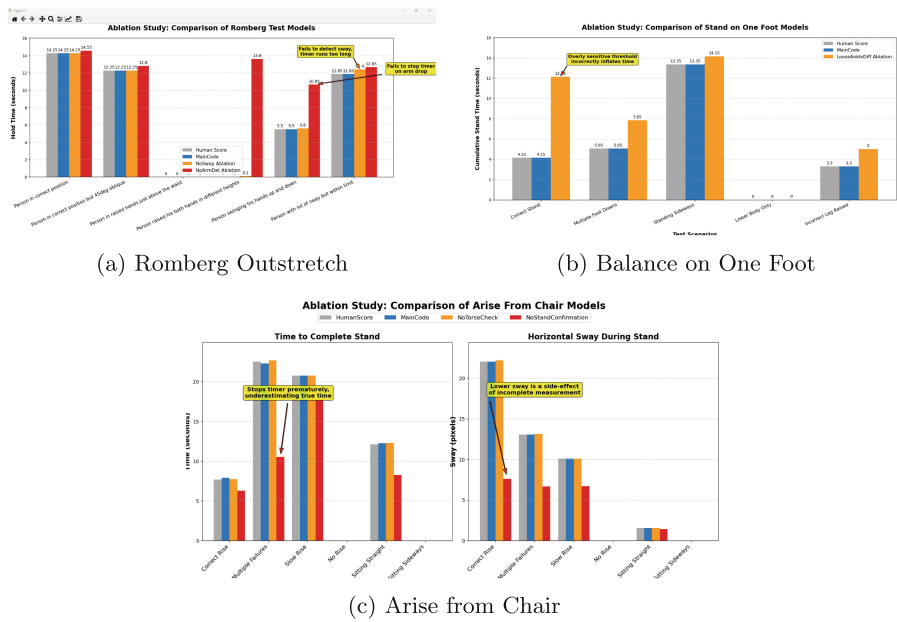


Fig. 1. Ablation plots across the three balance tasks. Removing task-specific safeguards introduces optimistic or premature scoring.

Robustness: Fixed vs Scaled Thresholds Across Distance We evaluate robustness under varying camera distances by comparing fixed, as summarized in Table 4 anatomically normalized thresholds with scale-adaptive thresholds across 0.5 m to 2.0 m.

Table 4. Fixed vs Scaled thresholds across distance for all tasks (time in seconds).

Dist.	Romberg		Balance on One Foot		Arise	
	Fixed	Scaled	Fixed	Scaled	Fixed	Scaled
0.5 m	14.20	7.15	14.60	14.85	13.85	13.50
1.0 m	—	14.55	13.25	13.90	12.75	10.75
1.5 m	5.85	5.80	11.45	11.90	11.86	12.10
2.0 m	7.25	7.15	13.15	14.20	9.56	8.05

Fixed thresholds provide stable and consistent measurements across distances, while scaled thresholds introduce variability and occasional premature or inconsistent outputs, particularly in multi-stage tasks.

3.3 Ablation Study

Ablation evaluates the removal of task-specific safeguards (H: human, AME: full model, -Sw: no sway, -Arm: no arm check, Loose: relaxed ankle threshold), as shown in Tables 1, 2, and 3

Removing constraints leads to overestimation in Romberg and Balance on One Foot, and unstable or underestimated outputs in Arise from Chair, confirming that task-aligned constraints are essential for reliable scoring.

4 Conclusion

We presented a vision-based approach for assessing standardized balance tasks using monocular video. By combining pose landmarks, normalized features, and task-specific state machines, the system produces interpretable clinical metrics.

Results show that task-aligned constraints are essential for reliable assessment, and fixed, anatomically normalized thresholds outperform scale-adaptive methods.

References

1. Leodori, G., et al.: Postural instability and risk of falls in patients with Parkinson's disease treated with deep brain stimulation: a stabilometric platform study. *Brain Sci.* **13**(9), 1243 (2023)
2. Müller, K., Cattin, G., Timm, W., Schmid, V.: Differences in fall-related characteristics across cognitive disorders. *Front. Aging Neurosci.* **15**, 10289027 (2023)
3. Buckley, C., et al.: The role of movement analysis in diagnosing and monitoring neurodegenerative conditions: insights from gait and postural control. *Brain Sci.* **9**(2), 34 (2019)
4. Bergquist, R., Weber, M., Schwenk, M., et al.: Performance-based clinical tests of balance and muscle strength used in young seniors: a systematic literature review. *BMC Geriatr.* **19**, 9 (2019)
5. Mancini, M., Horak, F.B.: The relevance of clinical balance assessment tools to differentiate balance deficits. *Eur. J. Phys. Rehabil. Med.* **46**(2), 239–248 (2010). PMID: PMC3033730
6. Lindroth, H., et al.: Applied artificial intelligence in healthcare: a review of computer vision technology application in hospital settings. *J. Imaging* **10**(4), 81 (2024)
7. Zhou, Y., Rashid, F.A.N., Mat Daud, M., Hasan, M.K., Chen, W.: Machine learning-based computer vision for depth camera-based physiotherapy movement assessment: a systematic review. *Sensors* **25**(5), 1586 (2025)
8. Bazarevsky, V., et al.: Mediapipe holistic: simultaneous face, hand, and pose prediction on device. Google AI Blog (2020). <https://research.google/blog/mediapipe-holistic-simultaneous-face-hand-and-pose-prediction-on-device/>
9. Kim, J.-W., Choi, J.-Y., Ha, E.-J., Choi, J.-H.: Human pose estimation using mediapipe pose and optimization method based on a humanoid model. *Appl. Sci.* **13**(4), 2700 (2023)



Machine Learning for Early Prediction of Gestational Diabetes Mellitus Using Routinely Collected Clinical Data

Syeda Rizvi¹ , Mais Alkhateeb² , Shehel Yoosuf³ , Aliya Tabassum⁴ ,
Arfan Ahmed³ , and Alaa Abd-Alrazaq³

¹ Hamad Bin Khalifa University, Doha, Qatar

² Lusail University, Doha, Qatar

³ Weill Cornell Medicine-Qatar, Doha, Qatar
aaa4027@qatar-med.cornell.edu

⁴ Qatar University, Doha, Qatar

Abstract. This study evaluated eight machine learning models for early GDM prediction using first-visit demographic, anthropometric, clinical, and laboratory data from 475 pregnant women (68 GDM-positive, 407 GDM-negative), analyzed using 60:40 stratified train-test split and class balancing using SMOTENC. Models exhibited trade-offs between sensitivity and specificity. Logistic Regression achieved the highest F1-score (0.34), with moderate recall (0.53) and accuracy (0.71), while MLP had the highest overall accuracy (0.78). BMI, maternal age, prior GDM, and early HbA1c were among the most influential predictors. The findings suggest that interpretable ML models applied to routinely collected first-visit data can support early risk stratification, potentially guiding targeted monitoring and interventions in antenatal care, while highlighting the intrinsic sensitivity-specificity trade-off in early GDM prediction.

Keywords: Machine Learning · Gestational Diabetes · Clinical Data · Prediction

1 Introduction

Gestational Diabetes Mellitus (GDM) refers to glucose intolerance that occurs during pregnancy. It is one of the most prevalent complications of pregnancy, affecting about 14% of pregnancies worldwide [1]. Although it may be resolved after childbirth, GDM poses significant short-term and long-term health risks for both the mother and the offspring. Beyond immediate complications such as preeclampsia, macrosomia, and increased cesarean delivery risk, GDM carries significant long-term consequences; up to 50% of affected women develop Type 2 Diabetes Mellitus (T2DM) within five years of delivery [2]. Therefore, early identification and management of GDM is important not only for improved pregnancy outcomes but also for reducing the long-term burden of chronic diabetes. Current clinical guidelines recommend screening for GDM between the 24th and 28th weeks of gestation using an Oral Glucose Tolerance Test (OGTT), with earlier testing for high-risk individuals. However, 2nd-trimester screening delays opportunities for preventive intervention.

Machine Learning (ML) based risk-prediction methods are increasingly used in healthcare research to provide more timely and effective solutions for early detection of conditions, leveraging their ability to model high-dimensional, non-linear interactions between clinical demographics and novel biomarkers. Despite the growing literature on the application of ML for evaluating GDM, several studies focus on the detection of GDM as opposed to predicting its future incidence [3, 4]. Other studies have explored novel predictor domains such as social network data [5], lipidomic profiles [6], and amino acid biomarkers [7], which substantially increase the complexity and cost of data collection, limiting their scalability in routine clinical settings. Recently, several studies have demonstrated the potential of ML models to predict GDM using early-pregnancy clinical data [5, 8]; however, these efforts have largely been confined to population-specific cohorts. Thus, developing models that use clinical data and generalize across populations is an open research question.

This study aims to evaluate ML models for early GDM risk estimation using routinely collected antenatal data including demographic, clinical, anthropometric, and laboratory data. We compare the performance of ML models under class-imbalance aware training with clinically relevant metrics. We also examine the trade-offs between model interpretability and predictive complexity to assess whether models can provide stable and clinically meaningful results in low-prevalence diagnostic settings.

2 Methodology

The dataset was obtained from a publicly available Kaggle repository and comprises structured clinical and demographic data from pregnant women [9]. The original dataset contained 600 records. Following data quality check measures, 125 records were removed due to anomalous or incomplete GDM related data, including missing GDM diagnosis labels ($n = 42$), absence of OGTT/GCT screening ($n = 57$) which is the clinical test used for GDM diagnosis, or inconsistent or contradictory diagnostic entries when recorded screening measures were incompatible with the assigned diagnosis labels ($n = 26$), suggesting entry errors. These exclusions were applied prior to any model development to prevent label noise and diagnostic ambiguity from confounding classifier performance. The final retained dataset contained 475 records, comprising 68 (14.3%) GDM-positive and 407 (85.7%) GDM-negative.

Prior to model development, missing values were handled via median and mode imputation for continuous and categorical variables respectively, followed by standardization and one-hot encoding. To address class imbalance, SMOTENC was applied exclusively to the training data within a pipeline-based approach to prevent data leakage. All preprocessing steps were performed within each fold on training pipeline to avoid data leakage. Additional information on the methodology employed in this study is available in Supplementary Material 1.

Multiple supervised ML algorithms were evaluated for GDM prediction, representing a diverse range of learning paradigms. The evaluated algorithms included Logistic Regression (Log Reg), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), Naive Bayes (NB) and Extreme Gradient Boosting (XGB), implemented using default hyperparameter settings. Model training and evaluation employed a 60:40 stratified train-test split

(285 training and 190 testing samples), to preserve class distribution. 5-fold stratified cross-validation was conducted exclusively on the training set to assess stability and reduce overfitting, while the held-out test set was reserved for final performance evaluation. Class imbalance was addressed via SMOTENC, incorporated within the training pipeline. All analysis was conducted using Python version 3.12.13.

Model performance was assessed using accuracy, precision, recall (sensitivity), specificity, F1-score, the area under the receiver operating characteristic curve (ROC-AUC), and the area under the precision–recall curve (PR-AUC). Calibration performance was evaluated using the Brier score.

3 Results

Results demonstrate that early GDM prediction using routine data is fundamentally constrained by sensitivity–specificity trade-offs rather than model choice alone. Logistic Regression achieved the highest F1 score (0.34) with accuracy (0.71), while MLP achieved second highest F1 score (0.32) and highest accuracy (0.78). Random Forest and MLP achieved the highest specificity (0.87 and 0.85, respectively), but with lower recall, indicating reduced detection of GDM-positive cases. RF achieved the best calibration according to Brier score (0.16). Performance metrics for all models are presented in Supplementary Material 2. 95% Confidence Interval is reported for all metrics, presented in Supplementary Material 3.

To improve interpretability and conduct cross-model comparison, feature importance was examined for all models. Logistic regression had the best F1 score, therefore we considered it as the best performing model. The top-ranked features were BMI, maternal age, previous history of GDM, and family history of diabetes/GDM emerging as the most influential predictors, although their relative importance varied between models. Feature importance for all models are presented in Supplementary Material 4.

Threshold sensitivity analysis, presented in Supplementary Material 5, demonstrated that lowering the classification threshold improved recall at the expense of precision, highlighting the importance of threshold tuning for clinical applications.

4 Discussion

This study evaluated ML models for early GDM prediction using routine first-visit clinical data. No model performed strongly across all metrics due to sensitivity–specificity trade-offs. Logistic Regression was the most balanced model, with the highest F1 score (0.34) and moderate accuracy (0.71), making it suitable for preliminary risk screening. MLP, Random Forest, and XGB showed higher specificity but missed more GDM cases, while Naïve Bayes had high recall but very poor specificity and calibration. The moderate ROC-AUC of Logistic Regression (0.61), class imbalance, weaker external validation, and distribution shifts indicate limited discrimination and generalizability. Therefore, the model is not ready for clinical deployment.

These findings are consistent with previous GDM prediction studies showing variable performance and limited benefit from complex models when only routine early-pregnancy variables are used [5, 10]. Important predictors included BMI >30,

age > 30 years, family history of GDM, previous GDM, medication intake, and HbA1c, aligning with established risk factors and SHAP-based evidence [5]. However, routine first-visit data capture only part of GDM risk compared with enriched datasets using advanced biomarkers [4].

This study has several limitations. The modest dataset size, low GDM prevalence, and participant heterogeneity may reduce model stability and generalizability. Although SMOTENC was applied to address class imbalance, the low F1-scores indicate limited ability to correctly identify GDM-positive cases, and the high accuracy and specificity likely reflect majority-class dominance rather than strong predictive performance. This trade-off is an inherent challenge in early GDM screening, where positive cases are naturally rare. Synthetic oversampling may have also introduced artifacts, and important predictors such as diet, physical activity, and longitudinal glucose measurements were unavailable. Decision curve analysis, presented in Supplementary Material 6 indicated net clinical benefit only within a narrow low-probability threshold range, and the modest performance ceiling (best F1 = 0.34), partly attributable to default hyperparameters, suggests that clinical relevance claims remain preliminary. While preliminary external validation is presented in Supplementary Material 7, broader multi-site prospective validation is still required. Future work should use larger multi-center cohorts, improve imbalance-handling methods, use longitudinal clinical and laboratory data, and perform prospective external validation before clinical implementation.

5 Conclusion

This study demonstrates the potential of ML in supporting early GDM prediction using routinely collected first-visit clinical data. Among the evaluated models, Logistic Regression offered the most balanced performance, while more complex algorithms favored specificity at the expense of sensitivity. No single model consistently achieved high accuracy, sensitivity, and specificity, reflecting the multifactorial nature of GDM and the limitations of first-visit clinical predictors. Key risk factors, BMI >30, age > 30 years, prior GDM, and family history, align with established clinical evidence and highlight features that could guide early risk stratification. Although promising, the models are not yet suitable for clinical deployment. Future research should focus on multicenter and prospective validation, inclusion of longitudinal and diverse datasets, and exploration of temporal and postnatal outcomes such as progression to type 2 diabetes.

Acknowledgments. Disclosure of Interests The authors declare no competing interests.

Declaration of AI Assistance Technologies During Manuscript Preparation. The authors used ChatGPT for language editing. After using this technology, the authors reviewed the results and take full responsibility for the contents of the manuscript.

Supplementary Materials. Supplementary Materials associated with this work is available at the following Open Science Framework link: <https://osf.io/dehsa>

References

1. Wang, H., et al.: IDF diabetes atlas: estimation of global and regional gestational diabetes mellitus prevalence for 2021 by international association of diabetes in pregnancy study group's criteria. *Diabetes Res. Clin. Pract.* **183**, 109050 (2022). <https://doi.org/10.1016/j.diabres.2021.109050>
2. Bellamy, L., Casas, J.-P., Hingorani, A.D., Williams, D.: Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. *Lancet.* **373**, 1773–1779 (2009). [https://doi.org/10.1016/S0140-6736\(09\)60731-5](https://doi.org/10.1016/S0140-6736(09)60731-5)
3. Ji, Q., et al.: Early prediction of gestational diabetes mellitus using machine learning-integrated metabolomic and clinical features. *Front. Endocrinol. (Lausanne)*. **16**, 1687146 (2025). <https://doi.org/10.3389/fendo.2025.1687146>
4. Zaky, H., et al.: Machine learning based model for the early detection of gestational diabetes mellitus. *BMC Med. Inform. Decis. Mak.* **25**, 130 (2025). <https://doi.org/10.1186/s12911-025-02947-3>
5. Li, Q., et al.: Development and external validation of a non-invasive early gestational diabetes mellitus prediction model integrating social network variables: a machine learning-based prospective cohort study. *BMC Pregnancy Childbirth.* **25**, 1333 (2025). <https://doi.org/10.1186/s12884-025-08500-4>
6. Lappas, M., et al.: The prediction of type 2 diabetes in women with previous gestational diabetes mellitus using lipidomics. *Diabetologia.* **58**, 1436–1442 (2015). <https://doi.org/10.1007/s00125-015-3587-7>
7. Wu, H., Chen, D., Li, X., Zhou, M., Wu, Q.: A novel single-test approach for GDM diagnosis: identification and prediction of high-risk postprandial hyperglycemia. *Meta.* **16**(1), 27 (2026). <https://doi.org/10.3390/metabo16010027>
8. Bigdeli, S.K., Ghazisaedi, M., Ayyoubzadeh, S.M., Hantoushzadeh, S., Ahmadi, M.: Predicting gestational diabetes mellitus in the first trimester using machine learning algorithms: a cross-sectional study at a hospital fertility health center in Iran. *BMC Med. Inform. Decis. Mak.* **25**, 3 (2025). <https://doi.org/10.1186/s12911-024-02799-3>
9. Karkera, A.: Gestational Diabetes, <https://www.kaggle.com/datasets/amithakarkera/gestational/data>. Last accessed 17 Jan 2026
10. Belsti, Y., et al.: Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM machine learning model. *Int. J. Med. Inform.* **179**, 105228 (2023). <https://doi.org/10.1016/j.ijmedinf.2023.105228>



Bridging the Pathology Communication Gap: Early Experience with Osler, a Patient-Facing AI Chatbot

Morgan Black¹(✉), Will Dowon Jeong¹, Sherman Lin¹, Ali Brown²,
Matthew Cecchini¹, and Jason K. Wasserman³

¹ Western University, London N6A 3K7, ON, Canada
Morgan.Black@lhsc.on.ca

² American Society for Clinical Pathology, Chicago 60603, IL, USA

³ The University of Ottawa, Ottawa K1H 8L6, ON, Canada

Abstract. Patients increasingly access pathology reports via electronic portals, yet these clinician-oriented documents are often misunderstood. AI-driven chatbots may support patient understanding, but concerns regarding accuracy, scope, and safety limit clinical use. We developed Osler, a pathology-specific chatbot deployed on *MyPathologyReport.com*, using curated educational content and embedded guardrails to constrain scope, acknowledge uncertainty, and redirect users to their healthcare team. A descriptive analysis of user interactions was conducted over two months post-launch (Dec 10, 2025-Feb 9, 2026). Osler was accessed by 2,408 users across 75 countries, generating 5,860 interactions (median 1.00; mean 2.43 queries per user). Safety-trigger responses occurred in 2.58% of interactions. Median readability was grade 10.37, and no clinically actionable medical advice was identified in a manually reviewed subset of 100 responses. These findings demonstrate the feasibility of a domain-constrained, safety-focused pathology chatbot and support further prospective evaluation of patient-facing AI tools in oncology communication.

Keywords: Patient-centred AI · Pathology Communication · Digital Health

1 Introduction

The expansion of electronic health record portals has increased patient access to diagnostic information, with pathology reports now routinely available in real time following policy changes such as the 21st Century Cures Act. Although this enhances engagement, these clinician-oriented documents contain specialized terminology, probabilistic language, and structured phrasing that are often difficult for patients to interpret accurately [1, 2]. Patients frequently misunderstand key elements of pathology and laboratory reports, including tumour staging, grading, margin status, and biomarker results, contributing to anxiety, confusion, and increased communication burden on clinical teams [1, 3]. While educational resources exist, many patients seek clarification through unsupervised internet searches, where information may be inconsistent or lack context [4].

Artificial intelligence (AI), including conversational agents, offers a potential approach to provide real-time, tailored explanations of complex medical information. Large language models have demonstrated strong performance in medical question answering and patient education [5], but patient-facing applications raise concerns regarding accuracy, hallucination, scope control, and unintended medical advice [6, 7].

Osler was developed as a pathology-specific chatbot embedded within MyPathologyReport.com, an educational platform providing accessible explanations of pathology terminology [8]. Unlike general-purpose AI tools, Osler uses curated, domain-specific content with embedded guardrails to constrain scope and promote appropriate redirection to healthcare providers. This study describes the development of Osler and evaluates early real-world usage, with objectives to assess adoption and engagement, response characteristics (including readability and scope-limiting behaviours), and preliminary safety outcomes during the first two months post-launch.

2 Methods

2.1 Platform and Chatbot Development

Osler was embedded on *MyPathologyReport.com*, a publicly accessible educational website providing plain-language explanations of pathology terminology. The chatbot launched on Dec 10, 2025, and was accessed voluntarily via user-initiated interactions, with no recruitment, prompting, or electronic health record integration.

Osler was developed using curated, clinician-reviewed *MyPathologyReport.com* content within a retrieval-augmented generation (RAG) framework to constrain outputs and reduce hallucination risk. The knowledge base comprised 1,065 web pages converted to text and segmented into paragraph-preserving chunks ($\leq 10,000$ characters, no overlap), embedded using *text-embedding-ada-002* and stored in ChromaDB. At query time, the top 5 nearest chunks were retrieved using L2 distance (equivalent to cosine similarity for normalized embeddings) and filtered with a 0.4 threshold (similarity = $1 - \text{distance}$). A secondary lightweight LLM (GPT-4.1-nano) performed relevance filtering to avoid inappropriate disease-specific associations, retaining up to 3 sources. Retrieved content was incorporated into prompts for generation using GPT-5-nano (v2025-08-07), with default sampling (temperature = 1.0, no fixed seed) and conversation history limited to the 5 most recent turns. Layered guardrails were implemented at both prompt and application levels to ensure safety and scope control. The model was restricted to informational responses, avoiding treatment recommendations, prognosis, medication guidance, and medical record interpretation, while acknowledging uncertainty and redirecting users to their healthcare team when appropriate. Additional safeguards included escalation pathways for self-harm disclosures and predefined safety templates for boundary enforcement. Application-level controls included rate limiting (20 requests/hour/IP), a global token cap, source logging, and post-response PHI detection with redaction of stored queries.

2.2 Study Design

We conducted a retrospective descriptive analysis of chatbot interactions during the first 2 months following deployment (Dec 10, 2025-Feb 9, 2026). The units of analysis

included unique users and individual interactions (user query + chatbot response). No identifiable personal health information was collected. Geographic location was inferred from IP-based analytics at the country level only.

2.3 Data Collection and Metrics

User- and interaction-level variables were extracted. User-level metrics included unique users, country-level location, sessions per user, and queries per session. Users were defined by distinct IP addresses, with all queries from the same IP attributed to a single user. Tracking relied solely on IP logging; cookies, account-based identifiers, and device fingerprinting were not used. Geographic location was estimated using IP-based geolocation (MaxMind GeoLite2 Country database; geoip2 Python library), with country assignments cached locally for reproducibility. Interaction-level metrics included total queries, mean and median queries per user, response length (word count), readability (Flesch–Kincaid grade level), and safety- and scope-related indicators.

2.4 Statistical Analysis

Descriptive statistics were calculated for all variables. Continuous variables are reported as means (SD) or medians (IQR). Categorical variables are reported as proportions. No inferential statistical testing was planned for this preliminary evaluation. Analyses were conducted using Python 3.12.3 (Python Software Foundation) with the pandas library (version 3.0.0) for data processing and statistical analysis.

3 Results

3.1 User Adoption and Geographic Reach

During the two-month study period, Osler was accessed by 2,408 unique users across 4,316 sessions in 75 countries, generating 5,860 total interactions. Engagement per user was low overall (median 1.00; mean 2.43 ± 3.45 queries per user), with most users submitting a single query, although a subset demonstrated higher repeat use (Fig. 1). The top five countries by user volume were Canada, the USA, the Netherlands, Indonesia, and Austria, reflecting broad international uptake.

3.2 Interaction Characteristics

The median chatbot response readability corresponded to a Flesch–Kincaid grade level of 10.37, exceeding commonly recommended health literacy targets of approximately grade 6–8 for patient-facing materials. Mean response length by word count was 268.89 ± 118.61 words. Qualitative categorization of interactions demonstrated that the majority of chatbot responses involved explanatory definitions of pathology terminology, clarification of diagnostic phrases, interpretation of biomarker results at a general educational level (without individualized advice), and broader contextual educational content related to pathology diagnoses.

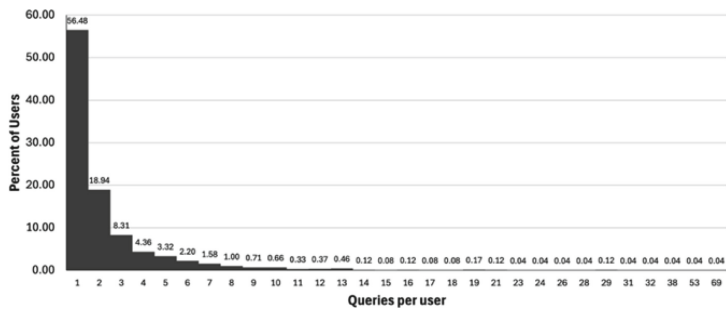


Fig. 1. Distribution of queries per user. User engagement with the Osler chatbot showing a right-skewed distribution, with most users submitting a single query during the study period.

3.3 Safety and Scope Control

Scope-limiting or safety-redirect responses occurred in 2.58% of interactions ($n = 151/5,860$), indicating that guardrail mechanisms were activated in a small but measurable proportion of queries. Redirection language encouraging users to consult their healthcare team was seen in 98.28% of responses (Wilson 95% CI 97.95–98.62), and uncertainty acknowledgment statements were observed in 67.91% (95% CI 66.70–69.12), reflecting consistent reinforcement of clinical boundaries and transparency. A subset of responses ($n = 100$) was randomly selected and manually reviewed to support evaluation of safety-related behaviour.

4 Discussion

This early evaluation demonstrates the feasibility of a pathology-specific chatbot with explicit safety and scope controls. Rapid global uptake suggests unmet demand for accessible pathology-related clarification. System behaviour reflected a safety-first design, with frequent redirection and uncertainty language indicating consistent boundary reinforcement. This likely reflects intentional prompt design and domain-constrained retrieval, though the balance between safety signaling and usability warrants refinement to avoid reducing engagement. Variation in response length likely reflects differences in query complexity, retrieved content, and safety templating. The low rate of scope-trigger events should be interpreted cautiously. It may indicate effective domain constraint but could also reflect the educational context and user selection of in-scope queries. In addition, detection depends on guardrail definitions and may miss more subtle boundary testing, requiring further auditing and stress testing.

Readability improved relative to traditional reports but remained above recommended health literacy levels. The domain-constrained architecture prioritizes factual grounding and scope control, but response quality was not formally evaluated; accuracy and completeness therefore remain uncertain and require targeted review of query-response pairs. Pathology reports are a critical communication interface, and tools such as Osler may support patient understanding and clinical discussions while reducing clarification burden. However, safe implementation depends on transparent limitations, consistent boundary enforcement, and clear escalation pathways.

This study is limited by its descriptive design, lack of comparator, absence of patient-reported outcomes, and reliance on readability as a proxy for comprehension. User counts based on IP addresses may both under- and over-estimate unique users (e.g., shared networks vs. device switching), limiting per-user inferences; language detection was not performed despite international use. Response quality was not formally assessed, and the short observation period limits insight into longitudinal performance.

Future work will include longitudinal monitoring, structured query analysis, and expert evaluation of response quality, beginning with targeted review of query–response pairs to assess accuracy and identify error patterns, followed by prospective studies and comparisons with general-purpose AI systems.

5 Conclusions

Early real-world deployment of Osler on *MyPathologyReport.com* demonstrates that a domain-constrained, guardrail-enabled pathology chatbot can achieve global reach while maintaining low rates of safety-trigger events and no detectable clinically actionable advice during initial review. These findings support further prospective evaluation of pathology-specific AI tools designed to enhance patient understanding.

Acknowledgments. We thank *MyPathologyReport.com* users and the development and content team for their contributions to this evaluation.

Disclosure of Interests Dr. Jason Wasserman is the founder and operator of *MyPathologyReport.com*, which hosts the Osler chatbot. This study was supported by an unrestricted educational grant from AstraZeneca, which had no role in study design or interpretation. No fees were charged to users. All other authors declare no competing interests. Generative AI was used for language refinement and formatting only; it was not used for data generation, analysis, or interpretation. All content was reviewed and approved by the authors, who take full responsibility for the work.

References

1. Giardina, T.D., Modi, V., Parrish, D.E., Singh, H.: The patient portal and abnormal test results: an exploratory study of patient experiences. *Patient Exp. J.* **2**(1), 148–154 (2015)
2. DesRoches, C.M., Bell, S.K., et al.: Patients managing medications and reading their visit notes: a survey of OpenNotes participants. *Ann. Intern. Med.* **171**(1), 69–71 (2019)
3. Zikmund-Fisher, B.J., Exe, N.L., et al.: Numeracy and literacy independently predict patients' ability to identify out-of-range test results. *J. Med. Internet Res.* **16**(8), e187 (2014)
4. Tan, S.S., Goonawardene, N.: Internet health information seeking and the patient–physician relationship: a systematic review. *J. Med. Internet Res.* **19**(1), e9 (2017)
5. Kung, T.H., et al.: Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit. Health.* **2**(2), e0000198 (2023)
6. Singhal, K., Azizi, S., Tu, T., et al.: Large language models encode clinical knowledge. *Nature.* **620**(7972), 172–180 (2023)
7. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 610–623. ACM, New York (2021)
8. Lafreniere, A., Purgina, B., Wasserman, J.K.: Putting the patient at the centre of pathology: an innovative approach to patient education – *MyPathologyReport.ca*. *J. Clin. Pathol.* **73**(8), 454–455 (2020)



Graph Attention Networks for Donor-Recipient Matching in Kidney Transplantation

Sheida Majouni¹(✉), Karthik Tennankore², Samina Abidi³, Amanda Vinson²,
and Syed Sibte Raza Abidi¹

¹ NICHE Research Group, Department of Computer Science, Dalhousie University, Halifax, NS, Canada

Sheida.Majouni@dal.ca

² Division of Nephrology, Department of Medicine, Dalhousie University, Halifax, NS, Canada

³ Department of Community Health & Epidemiology, Dalhousie University, Halifax, NS, Canada

Abstract. Kidney transplantation (KT) requires matching donor organs to the optimal recipients to maximize graft and patient survival. Donor-recipient matching (DRM) is a many-to-many decision problem in which individual transplant decisions are influenced by outcomes observed across the broader transplant population. Machine learning (ML) approaches evaluate D-R pairs independently and do not capture population-level dependencies or leverage clinical knowledge. We present an attention-based DRM framework using Graph Attention Networks (GAT) on a knowledge-driven bipartite graph, with a symmetric edge-pruning strategy that reduces the graph size while improving recipient survival prediction performance, and a temporal-aware graph-splitting strategy for robust evaluation. On adult recipients (≥ 65 years of age), our framework outperformed dense-graph and non-attention baselines (F1-score = 0.71 ± 0.02 and AUC = 0.76 ± 0.02).

Keywords: Kidney Transplantation · Graph Attention Networks · Donor-Recipient Matching · Knowledge-Driven Graph

1 Introduction

Donor-Recipient Matching (DRM) pairs a donor organ with the most suitable recipient by balancing biological constraints (blood group, human leukocyte antigen (HLA)), clinical factors, and ethical considerations. Suboptimal matching can result in transplant rejection, graft loss, and recipient mortality [1]. These challenges are especially significant for older transplant candidates (≥ 65 years) who face higher risks of graft failure and death [2]. In clinical practice, DRM relies on statistical methods (Cox regression, Kaplan-Meier) and clinical scoring systems (e.g., EPTS [3]). ML methods have also been applied to optimize post-transplant outcomes [4, 5]. A common limitation across these approaches is that they evaluate D-R pairs independently, ignoring interactions across domains and the population-level relationships among donors and recipients.

Building on our previous work [6, 7], where we introduced the bipartite DRM-Graph and a GraphSAGE-based DRM framework, this paper extends the approach in three key

directions: (i) attention-based representation learning via GAT, (ii) a scalable symmetric edge pruning strategy, and (iii) a Temporal Graph k-Fold Cross-Validation (TG-k-FCV) strategy that prevents future information leakage.

2 Related Work and Background

DRM is a hybrid feature-set matching problem combining (1) explicit matching of biological constraints and (2) implicit matching of soft clinical compatibility scores derived from historical outcomes. Traditional ML methods evaluate D-R pairs independently and cannot capture population-level dependencies, where the outcomes of similar cases inform the optimal match. Graph-based representations address this limitation by modeling DRM as a relational matching problem on heterogeneous graphs, where donors and recipients are nodes connected by edges encoding clinical knowledge and compatibility rules [8]. Graph Neural Networks (GNNs) such as GraphSAGE and GAT learn node embedding via iterative neighborhood aggregation [9]. GAT improves on GraphSAGE by using attention mechanisms to weight neighbors' importance. Other GNN approaches in transplantation address kidney exchange, a different problem focusing on cycle-finding among incompatible pairs [10]. Survival regression models such as DeepSurv operate on individual feature vectors and cannot exploit population-level relational structure, making them architecturally mismatched to DRM [11].

3 Methodology

This study used data from the Scientific Registry of Transplant Recipients (SRTR). The SRTR data system includes data on all donors, wait-listed candidates, and transplant recipients in the US, submitted by the members of the Organ Procurement and Transplantation Network (OPTN). The Health Resources and Services Administration (HRSA), U.S. Department of Health and Human Services provides oversight to the activities of the OPTN and SRTR contractors. The data reported here have been supplied by the Hennepin Healthcare Research Institute (HHRI) as the contractor for the Scientific Registry of Transplant Recipients (SRTR). The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government.

We applied the following inclusion criteria: (i) recipient age ≥ 65 at transplantation, (ii) deceased donor transplants only, (iii) transplant year between 2000 and 2014, and (iv) complete 5-year post-transplant follow-up, resulting in 13,669 D-R pairs (46.1% positive). Donor and recipient features (23 and 29) were selected in consultation with transplant nephrologists, prioritizing variables with established clinical relevance and excluding post-transplant variables. Minimal missing data (1.1%) were imputed; numerical features were scaled, and categorical features were one-hot encoded.

The DRM-graph is bipartite with donor and recipient nodes, each with feature vectors X_D, X_R , connected by three edge categories: *Direct Compatibility Edges* (E_{DC} , positive if 5-year survival, negative otherwise); *Hard Constraint Edges* (E_{HC} , ABO compatibility), and *Soft Preference/Knowledge Edges* (E_{SP} , clinical score compatibility derived

from deceased donor score and recipient risk score). A fully connected bipartite DRM-graph would have 187 M edges. After applying biological feasibility constraints, the graph contains 49 M edges across three categories: direct compatibility (13,669), blood-type (35 M), and knowledge-driven (13 M). This high density is driven by common blood types (O, A, B exceed 85%) and dense soft-preference connections, leading to slow convergence and unstable neighborhood sampling. Our edge pruning strategy operates symmetrically on source and target nodes. We identify the K most similar source entities using cosine similarity and retain edges to target historically linked via direct compatibility to at least one similar source.

We frame DRM as a link-prediction problem on the DRM-Graph. Compared with a GraphSAGE [6], GAT assigns learned attention weights to D-R neighbors rather than uniform aggregation. The model learns compatibility via message passing over E_{HC} and E_{SP} , a D-R pair is regarded as a successful match when the GNN’s inferred compatibility aligns with the observed historical relationships (Fig. 2A). Standard transductive splitting risks data leakage, since future recipients/donors in the test set become visible during training (Fig. 1B); inductive splitting prevents leakage but produces a sparse test graph (Fig. 1C). Our TG-k-FCV (Fig. 1D) orders cases chronologically into k sequential blocks. Each iteration uses one block for testing, the next for validation, and the remaining for training. Edges within each fold are assigned to four roles: training message edges, training supervision edges, validation edges, and test edges. For validation/testing, the DRM-graph is expanded to include new patients and their connections to historical cases, ensuring sufficient neighborhood structure while preventing future information leakage.

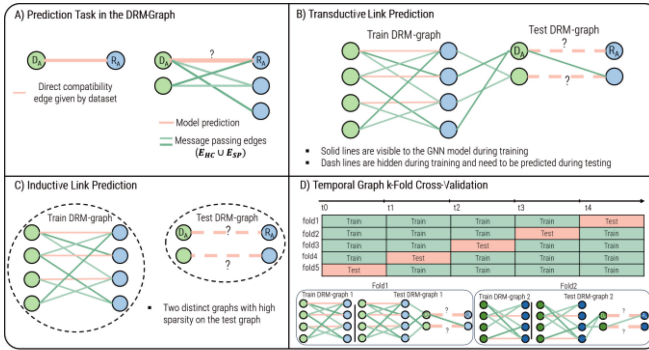


Fig. 1. (A) Link prediction task. (B-D) A comparison of data-splitting strategies, showing limitations of transductive and inductive approaches versus the proposed TG-k-FCV.

Our DRM framework operates in three stages. First, separate MLPs project heterogeneous donor and recipient features into a shared latent space (H^0). Second, a two-layer GAT aggregates information with attention coefficients, where Layer 1 performs cross-type message passing (donors aggregate from recipient neighbors and vice versa) and Layer 2 expands to second-hop, same-type neighbors. Each layer is followed by batch normalization, a non-linear activation, and dropout. Residual skip connections combine embeddings across depths (H^0, H^1, H^2). Finally, an MLP head maps concatenated D-R

embedding pairs to a continuous compatibility score. E_{DC} edges provide inherent positive/negative labels, eliminating negative sampling. The model is trained end-to-end with binary cross-entropy loss, RMSprop, neighborhood sampling [10, 5], 128 hidden channels, batch size 64, and 80 epochs.

4 Results

Symmetric edge pruning reduced the graph from 49 M to 905,354 edges ($\sim 2\%$ of the original) while preserving structural properties: (i) the semantic edge composition shifted only marginally (single-rule to dual-rule ratio 72:28% $\rightarrow$ 69:31%), indicating no disproportionate removal of edge types; (ii) connectivity remained balanced (median ~ 100 neighbors per node and only 0.16% under-connected with <5 neighbors); and (iii) multi-hop reach expanded naturally (334 nodes at Layer 1, 731 at Layer 2), consistent with a well-connected graph (Fig. 2).

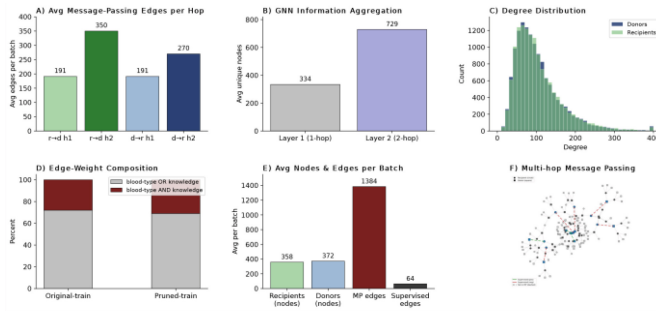


Fig. 2. Graph-level analysis of the pruned DRM-graph: (A) message-passing edges per hop, (B) neighborhood expansion across layers, (C) degree distribution, (D) edge-weight composition, (E) supervised vs. message-passing edges per batch, and (F) example of multi-hop message passing.

We evaluated the framework using 5-fold TG-k-FCV. Table 1 compares GNN models across graph settings, classifier heads, and architecture. The pruned DRM-Graph with GAT, MLP head, and residual connections achieved the highest accuracy (0.72 ± 0.03) and AUC (0.76 ± 0.02), despite being $\sim 2\%$ the size of the original. It outperformed models on the original graph (Accuracy = 0.63 ± 0.06 , AUC = 0.73 ± 0.04). Increasing neighborhood sampling on the original graph (from [10, 5] to [35, 30]) slightly improved sensitivity but not accuracy or AUC, confirming that the pruned graph retains key structural properties without the computational cost of a dense graph. GAT outperformed GraphSAGE on the pruned graph (AUC 0.76 vs. 0.73), with the highest F1_{neg}, indicating better ability to reject incompatible pairs, a critical property for clinical use.

Compared to ML baselines, our GAT achieved comparable AUC to RF and XGB (0.76 vs. 0.75) but significantly outperformed them in accuracy (0.72 vs. 0.68, $p = 0.02$) and F1_{macro} (0.71 vs. 0.67, $p = 0.04$). DeepSurv, as a representative deep survival regression baseline, achieved AUC = 0.55, confirming that individual-level survival regression cannot capture the population-level relational compatibility that graph-based

Table 1. GNN model comparison, 5-fold TG-k-FCV (P = pruned, O = original, R = residual).

Configuration	F1_pos	F1_neg	F1_macro	Accuracy $\pm$ std	AUC $\pm$ std
<i>P, GAT, MLP, no R</i>	0.61	0.71	0.66	0.67 ± 0.07	0.72 ± 0.02
<i>P, GAT, MLP, +R</i>	0.64	0.76	0.71	0.72 ± 0.03	0.76 ± 0.02
<i>P, SAGE, MLP, +R</i>	0.70	0.56	0.63	0.64 ± 0.02	0.73 ± 0.01
<i>O, GAT, MLP, +R</i>	0.69	0.52	0.61	0.63 ± 0.06	0.73 ± 0.04
<i>O - [35, 30], GAT, MLP, +R</i>	0.69	0.55	0.62	0.63 ± 0.02	0.73 ± 0.02

methods model. Counterfactual analysis showed a marginal donor (KDPI = 78), initially paired with a low-survival recipient (35%), was better matched to an alternative candidate (73%); allocations were unchanged when no better options existed.

5 Conclusion

This paper introduces an attention-based GNN framework for DRM in KT that captures explicit and implicit matching via biological constraints and compatibility scores. Our contributions include the graph construction strategy, the symmetric edge-pruning method, and the TG-k-FCV temporal evaluation protocol. We use standard GAT to isolate the effects of these contributions. Edge pruning reduces graph size while improving prediction; residual connections provide stable, balanced predictions; and GAT improves accuracy and interpretability through attention. The framework supports clinical decision-making by ranking candidate recipients in real time, complementing rather than replacing clinical judgment. Future directions include systematic comparison with relation-typed architectures, incorporating fairness constraints, modeling dynamic graphs, and applying the framework to other transplantation domains.

References

1. van de Klundert, J., et al.: The comparative performance of models predicting patient and graft survival after KT: a systematic review. *Transplant. Rev.* **39**(3), 100934 (2025)

2. Raza Abidi, S.S., et al.: Multiview clustering to identify novel kidney donor phenotypes for assessing graft survival in older transplant recipients. *Kidney.* **360**, 951–961 (2023)

3. Sandikci, B., et al.: A new simulation model for KT in the United States. In: *Proceedings - Winter Simulation Conference* (2019)

4. Ramalhetete, L., et al.: Revolutionizing KT: connecting ML and AI with next-generation healthcare, from algorithms to allografts. *BioMedInformatics.* (2024)

5. Briceño, J., et al.: AI and liver transplantation: looking for the best D-R pairing. *Hepatobiliary Pancreat. Dis. Int.* (2022)

6. Majouni, S., Tennankoreet, K., et al.: A graph neural network approach for data-driven DRM in kidney transplantation. In: *Intelligent Health Systems – From Technology to Data and Knowledge*, pp. 262–272 (2025)

7. Majouni, S., et al.: A heterogeneous bipartite graph framework for DRM in kidney transplantation. *Stud. Health Technol. Inform.* **327**, 532–536 (2025)
8. Khemani, B., et al.: A review of GNNs: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *J. Big Data.* **11**, 18 (2024)
9. Zhang, S., et al.: Graph convolutional networks: a comprehensive review. *Comput. Soc. Netw.* **6**(1), 1–23 (2019)
10. Pimenta, P.F., et al.: Solving the kidney exchange problem via GNNs with no supervision. *Neural Comput. & Applic.* **36**, 1–16 (2024)
11. Katzman, J.L., et al.: DeepSurv: personalized treatment recommender system using a cox proportional hazards deep NN. *BMC Med. Res. Methodol.* **18**, 24 (2018)



Multimodal Survival Prediction of Patients with Metastatic Esophagogastric Cancer Based on Clinical Progress Reports and Tabular Data

Irene Cara^{1,2,3} , Dimitris Katsimpokis^{3,4} , Gijs Geleijnse^{3,5} ,
Hanneke W. M. van Laarhoven^{2,3} , and Rob H. A. Verhoeven^{1,2,3} 

¹ Department of Medical Oncology, Amsterdam UMC location University of Amsterdam, Amsterdam, The Netherlands

i.cara@iknl.nl

² Cancer Treatment and Quality of Life, Cancer Center Amsterdam, Amsterdam, The Netherlands

³ Department of Research & Development, Netherlands Comprehensive Cancer Organization (IKNL), Utrecht, The Netherlands

⁴ Department of Clinical Informatics, Amsterdam UMC location University of Amsterdam, Amsterdam, The Netherlands

⁵ Norwegian Institute of Public Health, Cancer Registry of Norway, Oslo, Norway

Abstract. Existing prognostic models of overall survival (OS) in patients with metastatic esophageal or gastric cancer (EGC) rely on structured clinical data. Incorporating additional data modalities, particularly unstructured clinical notes, can further improve predictive performance. The aim of this study was to develop multimodal models, incorporating both tabular and text data, for OS prediction in patients with metastatic EGC and to compare their performance with unimodal models trained on tabular data or clinical notes only.

We developed both unimodal and multimodal survival models using Cox proportional hazards regression. Tabular and text data, in Dutch language, were obtained from the Electronic Health Records of 896 patients. Unstructured clinical text was embedded using multilingual sentence transformers finetuned with few-shot learning. The resulting representations were fused with tabular features prior to model training. Model performance was assessed using cross-validation and evaluated in terms of the time-dependent concordance index, dynamic area under the curve, and time-dependent Brier score.

Overall, the models demonstrated adequate discriminative ability and calibration, especially when structured data was included, with C-indexes within 0.6–0.67, Brier scores within 0.19–0.21 and AUC within 0.64–0.68. Best-performing model was the finetuned multimodal one, with an improvement on the C-index up to 5% and up to 13% compared to the tabular model and text model respectively. Sodium, age and the first text PCA component emerged as the most predictive variables. Future work will explore more advanced multimodal fusion strategies, including cross-attention mechanisms between data modalities, as well as the incorporation of longitudinal trajectories of structured clinical data.

Keywords: Multimodal prediction · Survival prediction · Clinical notes

1 Introduction

Esophageal and gastric cancer (EGC) affected around 1.5 million people and caused 1.1 million deaths in 2022 [1]. As EGC is often diagnosed when the disease is at an advanced stage, a large proportion of patients ultimately receive treatment with palliative intent. Patients and their families rely on physicians to synthesize patient's risk factors into an accurate overall assessment of life expectancy, or overall survival (OS), although physicians' estimates are often imprecise [2]. A more precise and data-driven individual OS prediction is essential, as it can contribute to better anticipation of palliative needs. Existing prognostic models typically rely on structured clinical data [3–5]. Incorporating additional data modalities can further improve predictive performance, for example adding imaging or genomics [6, 7] and text data [8, 9].

Unstructured EHRs include clinical notes, pathology reports or discharge letters in the form of free text. In the past, unstructured data were underutilized because extracting meaningful information required manual processing and conversion into numerical formats compatible with statistical and machine learning models. Thanks to the advent of transformers, manual extraction can be automated through embeddings, which are multidimensional representations of the information contained in text [10].

This study investigates the integration of survival models and natural language processing (NLP) techniques to enhance the accuracy of OS with multimodal prediction models for patients with metastatic EGC and compare their performance with unimodal prediction models.

2 Methods

Data were obtained from the electronic health records (EHRs) of Amsterdam University Medical Center and included both structured information and unstructured information of 896 patients. Patients were selected if diagnosed or treated at our center for synchronous or metachronous metastatic squamous cell carcinomas or adenocarcinomas of the esophagus, of the stomach or of the gastro-esophageal junction within 2016–2021.

The structured EHRs included clinical and demographical variables in a tabular format. The most recent measurement was extracted or, in case of missingness, the value was imputed with K nearest neighbor. The original dataset had 1144 covariates, which is more than the number of patients in our dataset. We selected only covariates which had less than 50% missing values, resulting into age, sex, tumor location, metachronous/synchronous/disease or recurrence, Body Mass Index and seven lab values. Numerical structured data were standardized, and categorical data were one-hot encoded.

The unstructured data included the most recent progress reports, written in Dutch. Some patients (16%) did not have any report by the prediction time. Instead of removing these patients, the text was reported as '[Missing]' to ensure the robust working of the pipeline also in case of a missing modality.

Five models were trained and compared: three unimodal models, one trained on the tabular data, two on the text data, and two models trained on the fused structured and unstructured data, of which one included few-shot finetuning of the embedding model.

The fusion was performed by concatenating the preprocessed structured data and the embeddings of the unstructured data, after reducing the dimensionality using Principal Component Analysis (PCA). For predicting OS of patients with EGC, we trained Cox Regression.

For embedding the text, we used the Dutch version of the Fairly Multilingual ModernBERT Embedding Model (Belgian Edition) [11], a Sentence Transformer with a context window of 8192 tokens to accommodate the potentially long clinical notes. As some notes were still longer than the context window, we implemented a sliding window with final pooling of the obtained window embeddings. The embeddings describe the general meaning of the text, but the model was not trained specifically on clinical data. For this reason, we used a few shot Sentence Transformer Fine-tuning (SetFit) framework [12] to make the embedding model capture the difference between long and short term survival in the notes. From the training set we sampled 16 notes associated with short survival (less than median survival minus 60 days) and other 16 associated with long survival (longer than median survival plus 60 days), which were used by SetFit to finetune the embeddings with contrastive learning. The full pipeline is depicted in Fig. 1.

All models were trained and compared in terms of time-dependent concordance (c-index), dynamic area under the curve (AUC) and time-dependent Brier score, evaluated at 3 months and 12 months after prediction time. We repeated all preprocessing steps in a nested 5-folds cross-validation, testing the performance in the outer loop and optimizing the hyperparameters of Cox (ridge regularization penalty) in the inner loop. Predictive power was measured with permutation feature importance.

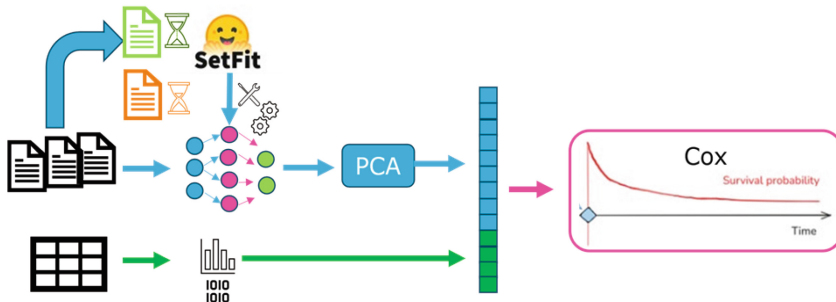


Fig. 1. Schematic view of pipeline used for the model trained on two (fused) modalities.

3 Results

The results of the 5-fold cross-validation of all models are reported in Table 1. The datasets used for the experiment are tabular data only, text data only, and the fusion of tabular and text data. All models including structured data show a discrete performance evaluated at 3 and 12 months, with C-indexes ranging from 0.61 to 0.67, Brier scores in the range 0.19–0.21 and AUC between 0.64 and 0.68. The unimodal text model shows an inferior overall performance (AUC 0.58–0.6 and C-index equal to 0.56–0.6).

The best performing model is the Cox model trained on tabular and text data with SetFit, showing a 0.03 increase in C-index, 0.02 on AUC and a 0.01 decrease in Brier score evaluated 3 months after prediction. In general, the C-index of the multimodal SetFit model had up to 5% increase with respect to the tabular unimodal model and up to 13% with respect to the unimodal text dataset. Based on the permutation of feature importance, the top three relevant features in the multimodal model were: sodium, age (tabular data) and first PCA component of the embeddings (text).

Table 1. Model performance of all five pipelines. C-index, Brier and AUC scores are given with standard deviation in parenthesis. Best model is in bold.

Model	C-index	C-index 3 months	C-index 12 months	Brier score 3 months	Brier score 12 months	AUC 3 months	AUC 12 months
Unimodal tabular	0.6 (0.03)	0.64 (0.05)	0.61 (0.03)	0.19 (0.01)	0.2 (0.01)	0.66 (0.04)	0.65 (0.07)
Unimodal text	0.56 (0.02)	0.6 (0.03)	0.56 (0.02)	0.2 (0.01)	0.21 (0.01)	0.6 (0.03)	0.58 (0.04)
Unimodal text SetFit	0.57 (0.01)	0.6 (0.02)	0.58 (0.01)	0.2 (0.01)	0.21 (0.01)	0.59 (0.02)	0.6 (0.02)
Multimodal	0.6 (0.03)	0.65 (0.03)	0.61 (0.03)	0.19 (0.01)	0.20 (0.01)	0.67 (0.04)	0.64 (0.07)
Multimodal SetFit	0.61 (0.02)	0.67 (0.03)	0.62 (0.02)	0.18 (0.01)	0.20 (0.01)	0.68 (0.03)	0.65 (0.05)

4 Discussion

This study investigates the integration of survival prediction and NLP techniques to enhance the accuracy of OS prediction models for patients with metastatic EGC. To our knowledge, this is the first study on multimodal survival prediction models for metastatic EGC. The model trained on the combination of structured and unstructured data had a better performance with respect to the unimodal models, showing potential for investigating in this direction. The presented methodology can be applied even when the text modality is not present. Thanks to the few-shot finetuning, the model can work with text in a low resource language, for which there is scarcity of good clinical foundation models

In addition to progress reports, other clinical notes are available in the EHRs. Their predictive power has yet to be tested. Although there is potential in adding other modalities, like the progress reports in this study, the gain in performance should be big enough to justify the decrease of applicability in clinical practice due to the more complex input to the prediction model. Current results should be further improved by extending the sample size, testing more advanced attention-based fusion methodologies and exploring the addition of the temporal modality, which can summarize the temporal evolution of

dynamic clinical variables. Future work will focus on the abovementioned aspects and on the explainability of the different modalities.

Disclosure of Interests. DK has received an institutional research sponsorship from AstraZeneca Netherlands. HvL has received research funding and/or medication/material supply from: AMGEN, Auristone, BMS, Incyte, Merck, MyeloidTx, ORCA, Servier, has had a consultant/advisory role: AMGEN, Amphera, Astellas, AstraZeneca, Beigene, BMS, Boehringer, Daiichi, MSD, MyeloidTx and was a speaker for AstraZeneca, BMS, Congress Care, Daiichi, Medtalks, MSD, Uitgeverij JAAP, Travel Congress Management. RV has received research grants from Bristol Myers Squibb and Amgen and performed consultancy for Daiichi Sankyo, all paid to the institute. GG has received research funding from Janssen, AstraZeneca and Roche; has a consultancy role for WeStat. All other authors have no competing interests.

References

1. Bray, F., et al.: Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**(3), 229–263 (2024)
2. Vasista, A., et al.: Accuracy and prognostic significance of oncologists' estimates and scenarios for survival time in advanced gastric cancer. *Oncologist*. **24**(11), e1102–e1107 (2019)
3. van den Boorn, H.G., et al.: SOURCE: prediction models for overall survival in patients with metastatic and potentially curable esophageal and gastric cancer. *J. Natl. Compr. Cancer Netw.* **19**(4), 403–410 (2021)
4. Jung, J.O., et al.: Machine learning for optimized individual survival prediction in resectable upper gastrointestinal cancer. *J. Cancer Res. Clin. Oncol.* **149**(5), 1691–1702 (2023)
5. Su, Y., et al.: Prediction of survival in patients with esophageal cancer after immunotherapy based on small-size follow-up data. *IEEE Open J. Eng. Med. Biol.* **5**, 769–782 (2024)
6. Luo, H., et al.: Multimodal multi-instance evidence fusion neural networks for cancer survival prediction. *Sci. Rep.* **15**(1), 10470 (2025)
7. Gomaa, A., et al.: Comprehensive multimodal deep learning survival prediction enabled by a transformer architecture: a multicenter study in glioblastoma. *Neurooncol. Adv.* **6**(1), vdae122 (2024)
8. Ibing, S., et al.: Electronic health records-based identification of newly diagnosed Crohn's disease cases. *Artif. Intell. Med.* **159**, 103032 (2025)
9. Song, S., et al.: Multimodal cancer modeling in the age of foundation model embeddings. In: *Mach. Learn. Health* (2025)
10. Kauffman, J., et al.: Embedding methods for electronic health record research. *Annu. Rev. Biomed. Data Sci.* **8**(1), 563–590 (2025)
11. Fairly Multilingual ModernBERT Embedding Model (Belgian Edition): Dutch Version [05-02-2026]. Available from: <https://huggingface.co/Parallia/Fairly-Multilingual-ModernBERT-Embed-BE-NL>
12. Tunstall, L., et al.: Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*. (2022)



Clinical Deployability: A Socio-technical Construct for Evaluating Real-World Readiness of Medical AI

Federico Cabitza^{1,2} and Mauro Dragoni²(✉)

¹ Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano, Italy

² Digital Health and Wellbeing Center, Fondazione Bruno Kessler (FBK), Trento, Italy
dragoni@fbk.eu

Abstract. Medical AI evaluation still largely emphasizes model performance under controlled conditions. However, whether an AI-enabled clinical capability can be introduced, used safely and appropriately, and sustained in routine care depends on additional socio-technical conditions. We propose *clinical deployability* as a second-order, multi-level construct capturing this readiness for real-world clinical work. Deployability integrates three coordinated domains: (A) task *usability* and safety-in-use; (B) implementation *workability* (embedding, routinization, sustainment); and, (C) *accountability* (in-use decision performance and risk governability). We outline operational criteria and a pragmatic measurement strategy based on a *deployability dashboard* rather than a single scalar score. This paper presents an ongoing research program; future work will apply the framework in real-world implementation case studies to produce comparative deployability assessments and refine the indicator set.

Keywords: AI Impact Assessment · Digital Health · Deployability · Workability · Usability · Accountability

1 Motivation

Many medical AI systems achieve promising discrimination or calibration in retrospective evaluations yet fail during implementation, are abandoned after initial uptake, or remain fragile once support decreases. These outcomes are increasingly understood as socio-technical rather than purely technical: an AI-enabled capability is an evolving configuration of model, interface, data pipeline, workflow integration, professional roles, governance, training, monitoring, and maintenance.

This creates a recurring methodological gap. “Evaluation” is often equated with clinical effectiveness or cost-effectiveness, while implementation-focused evidence is treated as secondary. In contrast, HCI and implementation science treat

feasibility, appropriate use, sustainment, and governability as legitimate evaluative objects. We propose *clinical deployability* to make this distinction explicit and operational.

2 Clinical Deployability

Clinical deployability is defined as:

the degree to which an AI-enabled clinical capability can be introduced, used appropriately, and sustained in real clinical work, while meeting minimum requirements for task-level usability and safety, and demonstrating adequate in-use decision performance in the settings for which it is deployed.

Deployability is *not* a claim about downstream clinical outcomes. It is an implementation-focused evaluation construct: it concerns readiness, robustness, and governability in context. Deployability is also *second-order* (composed of coordinated subdomains) and *multi-level* (emerging across individuals, teams, units, and organizations). We model deployability through three domains:

Domain A: Task Usability and Safety-in-Use. Usability is framed at the work-task level, consistent with ISO 9241-11 [6], but with safety-in-use treated as a first-class requirement. Relevant failure modes include misinterpretation of uncertainty, alert fatigue, unsafe defaults, and automation surprises under time pressure and interruptions.

Domain B: Implementation Workability. Workability captures whether the capability can be embedded and sustained as routine practice. It is aligned with implementation outcomes such as adoption, appropriateness, feasibility, fidelity, penetration, and sustainability [9], as well as with routinization mechanisms described by Normalization Process Theory [7,8].

Domain C: Accountability (In-Use Decision Performance and Risk Governability). This domain concerns whether decision performance is acceptable *in use* and whether residual risk is governable over time. It includes monitoring for performance drift, calibration and threshold behavior, and explicit indicators of *explainability* (such as *causability* [5]) and *appropriate reliance* [1] and AI influence, such as those defined in [2]. Critically, it also includes and: whether explanations support clinically meaningful understanding, counterfactual reasoning, and the rational evaluation of alternative options.

The domains follow the principle of *non-compensability* since they are conceptually separable but practically interdependent:

- strong performance cannot compensate for unsafe interaction;
- high usability cannot compensate for absent governance;
- high adoption cannot compensate for systematic over-reliance.

Deployability, therefore, cannot be reduced to a single number without explicit normative assumptions. The interconnection of the three domains is shown in Fig. 1.

3 Operationalization: The Deployability Dashboard

To support rigorous and feasible assessment across heterogeneous clinical settings, we propose a *deployability dashboard*: a compact, structured representation in which each domain is populated by a small set of indicators that combine quantitative measures (where available) and qualitative evidence (where necessary).

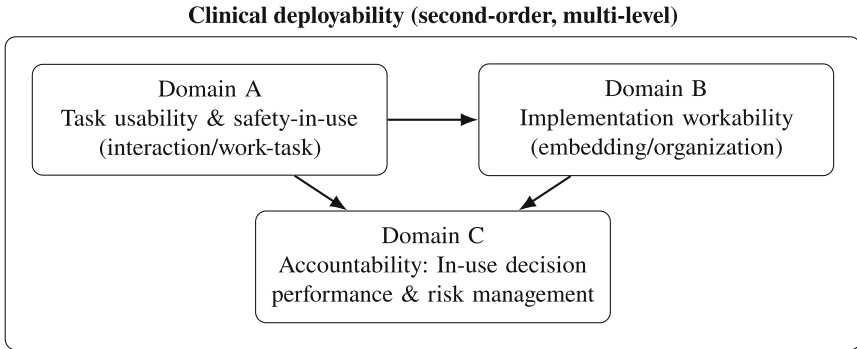


Fig. 1. Clinical deployability as a second-order construct composed of three coordinated domains.

In practical terms, the three domains can be assessed through complementary evidence streams. *Usability* should be evaluated through structured task testing, brief standardized questionnaires, and targeted observation of safety-relevant interaction breakdowns. *Workability* should be assessed as an implementation outcome profile [9], complemented by interviews and observation documenting workflow fit, training and competence-building, workarounds and fidelity to intended use, and organizational arrangements enabling routinization [7, 8]. *Risk governability* should be assessed through evidence that performance and reliance are monitored and actionable: explicit roles, escalation pathways, incident-learning procedures, change management for model/interface/workflow updates, and dedicated indicators for appropriate reliance, explainability, and causability. Post-go-live techno-vigilance [4] is treated as constitutive rather than optional [3].

Table 1 summarizes candidate indicators.

4 Research Program and Future Work

This paper introduces clinical deployability as a unifying evaluation construct and provides an initial operationalization. The work is ongoing. Future work will apply the framework to at least two real-world case studies of AI-enabled clinical capabilities, selected to vary in clinical task, workflow integration, and

Table 1. A deployability dashboard organizes evidence across three coordinated domains.

Domain A	Task usability & safety-in-use
Indicators	A1: Task completion and error types (frequency, severity, recoverability). A2: Time-on-task under realistic constraints. A3: Interruption sensitivity and resumability. A4: Perceived usability, workload, and satisfaction. A5: Comprehension of uncertainty and confidence cues. A6: Interpretability of model outputs (users can identify what is being predicted, scope constraints, and limitations). A7: Alert fatigue and override behavior. A8: Unsafe defaults and automation surprises. A9: Explanation usability (users can identify main drivers, limitations, and applicability conditions). A10: Explanation <i>actionability</i> (users can identify what additional information would plausibly change the recommendation).
Domain B	Implementation workability
Indicators	B1: Uptake patterns (who uses it, when, for which tasks). B2: Acceptability and perceived burden. B3: Feasibility (infrastructure, staffing, workflow fit). B4: Training coverage and competence-building (including explanation literacy and uncertainty literacy). B5: Fidelity to intended use and documented workarounds. B6: Penetration across units and roles. B7: Routinization and integration with existing information systems. B8: Sustainment resourcing. B9: Socio-technical fit (redistribution of work, accountability, coordination demands). B10: Governance readiness (clear ownership, escalation pathways, incident-handling procedures). These align with implementation outcomes [9].
Domain C	Decision Accountability (In-use decision performance & risk)
Indicators	C1: Decision accuracy and error distribution in representative use conditions. C2: Subgroup performance and distribution shift monitoring. C3: Calibration and threshold behavior under workflow constraints. C4: Distribution of reliance modes (appropriate reliance, over-reliance, under-reliance), reliance calibration, and override/acceptance rates stratified by risk and uncertainty. C5: Disagreement patterns between users and model (and resolution outcomes). C6: Explainability & causability, explanation completeness, stability, and consistency. C7: Concordance with domain knowledge. C8: Counterfactual plausibility (clinically feasible “what-if” tests). C9: Causal interpretability (distinguishing correlational cues from plausible causal factors). C10: Traceability and auditability (input provenance, versioning, explanation record). C11: Incident and near-miss learning (including explanation failures). C12: Monitoring signals and triggers for intervention (retraining, UI changes, workflow redesign). C13: Post-go-live change management for model/interface/workflow updates [3,4].

organizational complexity. The empirical goal is to (i) produce deployability dashboards that enable within- and cross-case comparison, (ii) identify recurring socio-technical failure modes (e.g., reliance breakdowns, explanation failures, governance gaps), and (iii) refine the indicator set into a minimal, reusable assessment template for medical AI projects.

5 Conclusion

Clinical deployability reframes evaluation for medical AI as an implementation-focused assessment of readiness, robustness, and governability in real clinical work. The construct integrates task usability and safety-in-use, implementation workability, in-use decision performance and risk management, and explicitly includes appropriate reliance, explainability, and causability. Rather than collapsing evidence into a single score, deployability is assessed through a dashboard approach that remains actionable for design, governance, and post-go-live oversight. The next phase of the project will validate and refine the framework through comparative real-world case studies.

Declaration of AI Assistance Technologies During Manuscript Preparation. The authors used ChatGPT Plus 5.2 for the refinement, correction, editing and formatting of the manuscript to improve clarity of language. After using this technology, the authors reviewed the results and take full responsibility for the contents of the manuscript.

Disclosure of Interests. The authors declare they have no conflict of interest.

References

1. Cabitza, F., et al.: Five degrees of separation: investigating the unexpected potential of displaced human-AI collaboration protocols for apter AI support. *Proc. ACM Hum.-Comput. Interact.* **9**(7), 1–28 (2025)
2. Campagner, A., Fregosi, C., Natali, C., Cabitza, F.: Under what influence: measuring AI influence to fit user profiles in decision-making. *Int. J. Hum.-Comput. Stud.* (2026, in press)
3. Greenhalgh, T., et al.: Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *J. Med. Internet Res.* **19**(11), e367 (2017). <https://doi.org/10.2196/jmir.8775>. <https://doi.org/10.2196/jmir.8775>
4. Harvey, H., Cabitza, F., et al.: Algorithms are the new drugs? Reflections for a culture of impact assessment and vigilance. In: *IADIS International Conference ICT, Society and Human Beings 2018* (2018)
5. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of artificial intelligence in medicine. *WIREs Data Min. Knowl. Discov.* **9**(4), e1312 (2019)
6. International Organization for Standardization: ISO 9241-11:2018 ergonomics of human-system interaction — part 11: Usability: Definitions and concepts. International Standard ISO 9241-11:2018, ISO, Geneva, Switzerland (2018). <https://www.iso.org/standard/63500.html>, reference number 63500; publication date 2018-03

7. May, C., Finch, T.: Implementing, embedding, and integrating practices: an outline of normalization process theory. *Sociology* **43**(3), 535–554 (2009). <https://doi.org/10.1177/0038038509103208>
8. May, C.R., et al.: Development of a theory of implementation and integration: normalization process theory. *Implement. Sci.* **4**(1), 29 (2009). <https://doi.org/10.1186/1748-5908-4-29>
9. Proctor, E., et al.: Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Adm. Policy Ment. Health* **38**(2), 65–76 (2011). <https://doi.org/10.1007/s10488-010-0319-7>



A Knowledge Acquisition Platform and Sharing Library for the TMR Computable Guideline Model

Saif Latifi^(✉) and Vasa Curcin

Digital Health Group, Department of Population Health Sciences, King's College
London, London, UK
{saif.latifi, vasa.curcin}@kcl.ac.uk

Abstract. Clinical Practice Guidelines often fail in multimorbidity contexts, where aggregating single-disease recommendations creates therapeutic conflicts. The Transition-based Medical Recommendations model (TMR) deduces conflicts through physiological logic, but technical complexity hinders adoption. This paper presents an open-source web platform designed to bridge this “knowledge acquisition bottleneck”.

Our system combines a “no-code” Knowledge Acquisition Dashboard and an Interaction Visualiser. By integrating the Z3 theorem prover with a SnowStorm terminology server, the system automates the formalisation of clinical logic using SNOMED CT. This allows clinicians and policy makers to graphically encode guidelines and detect inconsistencies – such as contradictions – without technical expertise.

A collaborative library supports a “Duplicate and Modify” workflow, enabling stakeholders to adapt protocols safely, in line with local guidance. By visualising interactions through a user-friendly frontend, the platform empowers domain experts to audit medical protocols before dissemination, demonstrating a scalable path toward safer guideline management.

1 Introduction and Background

Clinical Practice Guidelines (CPGs) serve as the cornerstone of evidence-based medicine [2]. To operationalise this knowledge, narrative documents are increasingly translated into Computer-Interpretable Guidelines (CIGs) – formal representations that allow systems to process clinical logic and automate decision support [3]. Yet, traditional formats struggle to accommodate the complexities of multimorbidity. Blind aggregation of single-disease guidelines frequently results in contradictory advice. Current approaches, which rely on manually encoded exclusion rules, are inherently unscalable and computationally intractable, as the number of guidelines increases [4].

To address the scalability of conflict detection, TMR (Transition-based Medical Recommendations), originally proposed by Zamborlini et al. [5], introduced a paradigm shift from manual rule curation to effects-based reasoning. Rather

than defining static rules for every possible interaction, TMR models recommendations based on their intended effects – specifically, the transitions they induce in a patient’s physiological state (e.g., shifting blood pressure from ‘high’ to ‘normal’). This formalisation allows a reasoning engine to logically deduce conflicts when two recommendations attempt to drive a physiological parameter in opposing directions, thereby automating the detection process without the need for exhaustive, pre-defined conflict lists.

Despite the theoretical strength of TMR, its practical adoption has been hindered by the ‘knowledge acquisition bottleneck’ [3]. Previous implementations, such as ROAD2H [1], successfully demonstrated the power of the underlying reasoning engine but remained backend-focused, lacking user-facing interfaces for guideline creation and analysis. Consequently, the power of TMR has largely remained inaccessible to the domain experts – clinicians and policy makers – who are best positioned to interpret and utilise these insights. Without intuitive tools to authorise and visualise these complex logical structures, the potential for TMR to aid in guideline development and auditing remains untapped.

In this demo paper, we demonstrate a comprehensive, web-based system that bridges this gap. Unlike previous implementations, our system provides an interactive guideline builder and a visual conflict analyser, making TMR approachable to a wider, non-technical audience. This platform allows users to graphically encode guidelines and visualise detected conflicts, facilitating the static analysis of recommendation sets. This tool is designed not only for clinical utility but as a resource for policy makers, enabling them to verify the internal consistency of medical protocols and identify potential adverse interactions at the guideline level before they are disseminated into practice.

2 Methods

We developed a web-based system comprising a knowledge acquisition interface and a reasoning engine. The system automates the formalisation of clinical logic and the subsequent detection of interactions using an adapted version of the Transition-based Medical Recommendations (TMR) model.

2.1 Adapted TMR Model

Our system utilises a modified TMR model [5] which models recommendations based on intended physiological transitions rather than static rules. We diverge from the original proposal by excluding causation beliefs (probabilities and evidence strength) to minimise the manual encoding burden of narrative guidelines and to avoid hard-coding patient-specific efficacy, thereby enhancing generalisability. Consequently, our model focuses purely on logical structure, enabling the detection of *Divergent Causations*, *Contradiction*, *Repetition*, *Alternative*, *Side-Effect* and *Repairable* interactions, while excluding *Safety* interactions.

2.2 Platform Architecture and Implementation

The platform employs a distributed microservices architecture (built with NestJS) to ensure modularity. It includes Authentication, CIG Management, and an Interaction services, exposed via a standardised API for interoperability. The SvelteKit user interface provides a reactive frontend for visual guideline authoring and real-time interaction analysis (see Fig. 1).

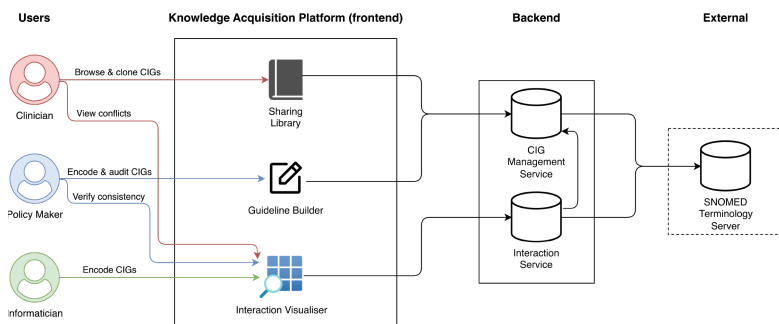


Fig. 1. High-level architecture and user flows of the system, comprising a Knowledge Acquisition Platform, a backend, and SNOMED terminology server. (CIG - Computer-Interpretable Guideline)

2.3 Reasoning Engine and Terminology Integration

To automate interaction detection, the Interaction Service integrates the Z3 theorem prover¹ using FOL. We employ the Strategy design pattern to encapsulate interaction rules dynamically. To handle semantic subsumption (e.g. recognising *Aspirin* as a subtype of *NSAID*), the platform connects to a SnowStorm terminology server². By anchoring definitions to SNOMED CT concepts, the system leverages the ontology’s hierarchy to automate consistency checks across varying levels of granularity. The full source code for the platform and reasoning engine is available as open-source software³.

3 Features of the Platform

The platform provides a user-centric environment designed to bridge the gap between narrative clinical text and executable logic, making formal modelling accessible to non-technical experts.

¹ <https://github.com/Z3Prover/z3>.

² <https://github.com/IHTSDO/snowstorm>.

³ <https://github.com/slatifi/tmr-web>.

3.1 Knowledge Acquisition Dashboard

Built with SvelteKit, this workspace replaces code-heavy editors. Users define guidelines using SNOMED CT semantic auto-completion, ensuring actions and transitions anchor correctly to ontology concepts. This “no-code” approach allows clinicians to formalise knowledge without FOL expertise.

3.2 Interaction Visualiser

To interpret multimorbidity logic, the system renders an interactive graph where recommendations and physiological states act as nodes connected by transitions. When the engine detects inconsistencies like *Contradictions*, it highlights the conflicting nodes directly on the graph. This visual feedback allows users to immediately diagnose the impact of combining multiple guidelines in real-time.

Figure 2 shows an excerpt of the interaction graph between NICE guidelines for Acute Coronary Syndromes (NG185), Osteoarthritis (NG226) and Gastro-oesophageal Reflux Disease (CG184). A full encoding of the guidelines was omitted due to space constraints.

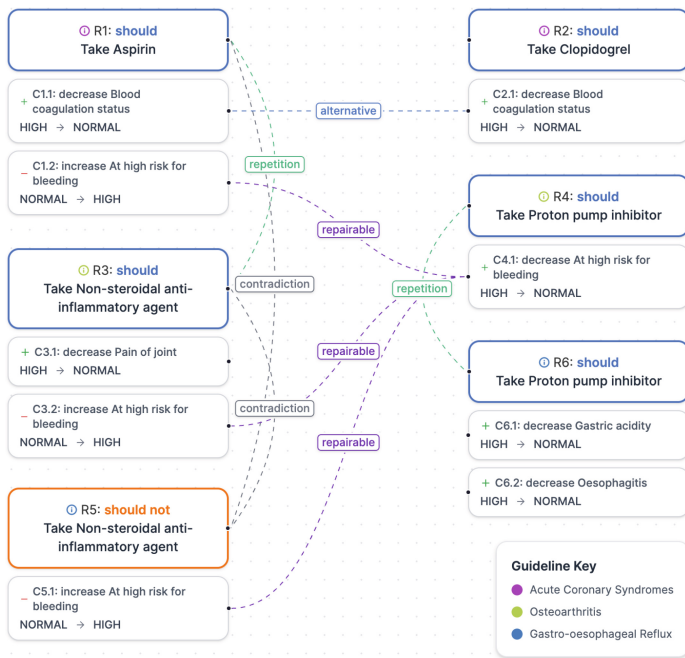


Fig. 2. Interaction graph. R1 & R2 are recommendations from NG185, R3 & R4 are from NG226, and R5 & R6 are from CG184.

3.3 Sharing Library

The platform serves as a collaborative library where users can browse published guidelines and perform interaction detection across them. Furthermore, it supports a “Duplicate and Modify” workflow where users can clone verified guidelines into a private workspace to safely adapt logic, such as substituting a drug based on local availability, and immediately use the visualiser to ensure these adaptations do not introduce new conflicts.

4 Conclusion

Our distributed platform uniquely combines a Knowledge Acquisition Dashboard with an automated reasoning engine, democratising guideline formalisation through a visual, terminology-driven interface. Integrating the Z3 theorem prover with SNOMED CT provides a robust mechanism for visualising logical conflicts, empowering stakeholders to contribute to community-driven guideline engineering and safe multimorbidity management.

Future work involves a two-part evaluation of the platform. First, we will evaluate Z3’s computational scaling with large recommendation sets and ontological subsumptions, followed by a clinical evaluation to assess the engine’s soundness and completeness using real-world guidelines. We also aim to incorporate dynamic, patient-specific decision support and refine the TMR model to represent more complex guideline structures, such as conditional recommendations and hierarchical logic, which is often present in real-world guidelines.

Acknowledgements. This research work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) [Grant reference number EP/Y035216/1] Centre for Doctoral Training in Data-Driven Health (DRIVE-Health) at King’s College London.

References

1. Domínguez, J., et al.: ROAD2H: development and evaluation of an open-source explainable artificial intelligence approach for managing comorbidity and clinical guidelines. *Learn. Health Syst.* **8**(2), e10391 (2024). <https://doi.org/10.1002/lrh2.10391>
2. Institute of Medicine (US). Committee on Standards for Developing Trustworthy Clinical Practice Guidelines: Clinical practice guidelines we can trust (2011)
3. Peleg, M.: Computer-interpretable clinical guidelines: a methodological review. *J. Biomed. Inform.* **46**(4), 744–763 (2013). <https://doi.org/10.1016/j.jbi.2013.06.009>
4. Van Woensel, W., et al.: A community-of-practice-based evaluation methodology for knowledge intensive computational methods and its application to multimorbidity decision support. *J. Biomed. Inform.* **142**, 104395 (2023). <https://doi.org/10.1016/j.jbi.2023.104395>
5. Zamborlini, V., et al.: Analyzing interactions on combining multiple clinical guidelines. *Artif. Intell. Med.* **81**, 78–93 (2017). <https://doi.org/10.1016/j.artmed.2017.03.012>



AI-Informed 30-Day MACE Risk Predictions in ED Arrivals

Ben J. Gross¹(✉) , M.D. Radha Patel² , Kai Zheng³ ,
Shamim Nemati¹ , and M.D. Mattheus Ramsis⁴

¹ UC San Diego, Division of Biomedical Informatics, San Diego, CA, USA
bjgross@health.ucsd.edu

² UC San Diego, Division of General Internal Medicine, San Diego, CA, USA

³ UC Irvine, Department of Informatics, Irvine, CA, USA

⁴ UC San Diego, Division of Cardiovascular Medicine, San Diego, CA, USA

Abstract. Undifferentiated chest pain is a major driver of emergency department (ED) overcrowding and contributes to morbidity and mortality. Early risk stratification for major adverse cardiac events (MACE) allows for early discharge and ED decongestion. We developed a multi-modal machine learning model integrating EHR data, laboratory results, demographic and comorbidity features and chest x-ray impressions to predict 30-day MACE in patients presenting to the ED with chest pain. Using 49,348 ED encounters from two healthcare systems, our model achieved a high area under the curve (AUC) of 82.0% at the development site and 74.6% at the external validation site. Predictions occurred within 6 h of ED arrival and demonstrated consistently high negative predictive values across time, supporting safe early-rule out. This model demonstrates the potential of integrating AI modeling to enhance early MACE risk stratification and improve ED overcrowding and efficiency.

Keywords: MACE · Triage CDSS · Undifferentiated Chest Pain

1 Introduction

Undifferentiated chest pain is one of the most common reasons for emergency department (ED) visits, accounting for approximately eleven million annual presentations in the United States [1]. Although fewer than 10% of these patients are ultimately diagnosed with acute coronary syndrome (ACS), the need to exclude life-threatening conditions contributes substantially to ED overcrowding, delays in care, and increased morbidity and mortality [2, 3]. This creates a critical challenge for clinicians, who must rapidly distinguish low-risk from high-risk patients in a resource-constrained and time-sensitive environment.

Clinical risk scores such as HEART, TIMI, and GRACE are widely used to stratify risk for major adverse cardiac events (MACE). However, these tools are limited by reduced sensitivity and negative predictive value (NPV) at common

decision thresholds, limited generalizability to heterogeneous ED populations, and an inability to incorporate dynamic clinical data such as evolving laboratory results and imaging findings [4–6]. As a result, they may overestimate risk in low-risk patients and contribute to unnecessary admissions and resource utilization.

The increasing availability of electronic health record (EHR) data provides an opportunity to develop more accurate, data-driven approaches to early risk stratification. In this study, we develop and externally validate a multi-modal machine learning model that integrates structured clinical data, laboratory values, and imaging-derived features to predict 30-day MACE in ED patients with chest pain. Our goal is to improve early identification of low-risk patients and support safe discharge decisions to reduce ED overcrowding.

2 Methods

We conducted a multi-center retrospective cohort study of adult ED patients at two academic centers. The development cohort was derived from University of California San Diego Health (UCSD), including encounters from two EDs between January 2016 and December 2023. External validation was performed using data from University of California Irvine (UCI), including ED encounters from January 2023 through August 2024. Institutional review board approval with waiver of informed consent was obtained at both sites.

We included all adult patients (≥ 18 years) presenting during the study periods. Encounters were eligible if a chest radiograph (x-ray) and at least one troponin measurement were obtained within 6 h of ED arrival and prior to any MACE. Predictor variables included the first available values from the comprehensive metabolic panel, complete blood count, lactate, and B-type natriuretic peptide (BNP), measured prior to the prediction time. Up to two troponin measurements were included if both occurred before the prediction time, defined as the time of both troponin measurement and chest x-ray availability.

The primary outcome was 30-day MACE, defined using surgical ICD codes indicating events such as acute myocardial infarction, heart failure, or stroke occurring within 720 h of ED arrival, including subsequent readmissions.

The final cohort included 30,614 encounters at UCSD and 18,734 at UCI. MACE occurred in 240 (0.8%) encounters at UCSD and 479 (2.6%) at UCI. Cohort characteristics stratified by site and outcome are presented in Table 1.

We also track whether patients are given MACE related IV medications during the encounter itself or within 30 days of the encounter start time. These intravenous medications include Heparin, Lasix (Furosemide), and Bumex (Bumetanide). The length of stay in the ED (ED LOS) is measured for all encounters in our cohorts.

In addition to the aforementioned EHR model input features, we also input the gender, race, ethnicity, and age of the patient at the encounter start time. We also note the presence or absence of 15 MACE-related comorbidities from the problem list of the patient at the encounter start time, such as hypertension, diabetes, hyperlipidemia.

Table 1. Cohort demographics and outcomes are shown below. These are broken down by study site and the presence of 30-day MACE, measured from hospital arrival time.

	30-day MACE		No 30-day MACE	
	UCSD (n = 240)	UCI (n = 479)	UCSD (n = 30,374)	UCI (n = 18,255)
Patient Characteristics				
Age, median (years)	65.1	65.3	61.6	62.5
Age, IQR (years)	[56.7–73.5]	[56.0–74.6]	[49.9–72.7]	[47.8–75.6]
Female (%)	36.3	34.4	44.3	49.2
Black Race (%)	9.58	3.1	14.3	3.3
Asian Race (%)	7.5	21.5	7.6	20.6
White Race (%)	44.6	43.0	49.8	45.4
Hispanic Ethnicity (%)	31.3	37.0	23.0	39.8
Patient Outcomes				
ED LOS, median, (hours)	7.3	10.7	7.0	8.7
ED LOS, IQR, (hours)	[4.6–16.8]	[6.5–25.7]	[4.9–10.9]	[5.5–21.8]
Prediction time, median (hours)	3.0	2.5	2.8	2.6
Prediction time, IQR (hours)	[2.0–4.5]	[1.8–3.2]	[2.0–4.0]	[2.0–3.5]
MACE meds during encounter (%)	69.6	78.7	22.7	14.4
MACE meds within 30 days (%)	75.4	82.3	25.0	15.6

Lastly, the model uses the most recent chest x-ray impression available at the time of prediction to capture key MACE risk assessments not present in the EHR. This impression is analyzed using Claude 3.5 Sonnet with instructions to provide a simple MACE risk assessment level (Low Risk, Borderline, High Risk, or Imminent) as an input for our model.

These 40 quantitative and categorical model input features are used to train an XGBoost model [7] to predict MACE within 30 days of the encounter start time. We use an 80/20 training/testing split of our UCSD encounters, thus training the model on 24,491 encounters and testing on the remaining 6,123 encounters. All 18,734 UCI encounters were kept together as an external validation set.

The modeling pipeline was implemented in a HIPAA-compliant cloud environment (AWS EC2), with secure access to Claude via the Amazon Bedrock Converse API.

3 Results

This model obtains an Area Under the Curve (AUC) of 82.0% on the UCSD testing set, and an AUC of 74.6% on the UCI external validation set (Table 2). Model sensitivity is targeted at 70% on the testing set, which determined the threshold risk score for positive model prediction. At this threshold the Positive

Predictive Value (PPV) is low (2.4% for the testing set, 10.6% for the validation set) but our NPV remains high, which aligns with the model’s goal to allow low-risk patients to be safely discharged sooner. In the validation set, the sensitivity falls to 30.1%, suggesting a large distribution shift between these datasets. We note that the majority of model predictions occurred within 3 h of ED arrival, enabling actionable early risk stratification during ED evaluation.

Table 2. Model Performance at Development Site (UCSD) and Validation Site (UCI).

Study Site	AUC	AUC-pr	Sensitivity	Specificity	PPV	NPV
UCSD	82.0%	4.8%	70.8%	77.3%	2.4%	99.7%
UCI	74.6%	9.7%	30.1%	93.3%	10.6%	98.1%

For each study site, we bin the model predictions by hours from hospital arrival to analyze how the model performance changed with the prediction time (Fig. 1). Since the NPV is the most critical metric for whether our model can allow earlier patient discharge, we compute a Spearman rank correlation [8] between hourly NPV and prediction hour of the model. The correlation was not statistically significant ($p = 0.35$ at UCSD, $p = 0.16$ at UCI), indicating no evidence of a significant change in NPV by model prediction time.

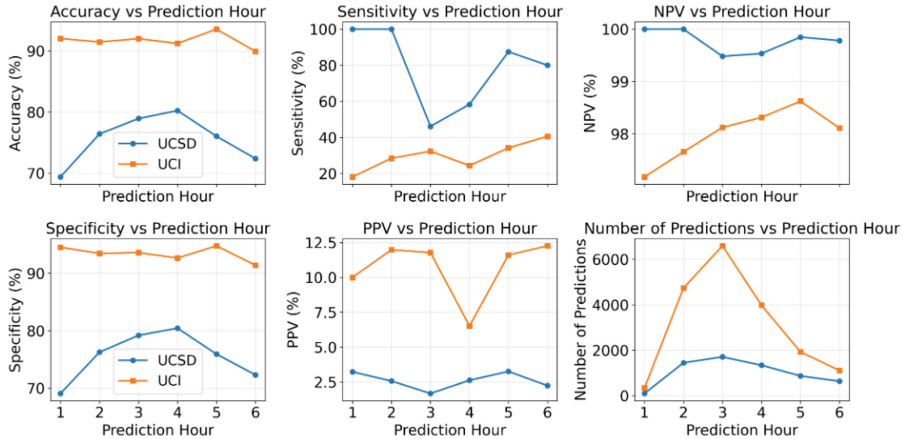


Fig. 1. Model performance by time of prediction (hours from ED arrival).

4 Conclusion and Future Work

ED overcrowding remains a persistent issue that impacts quality of care and increases patient morbidity and mortality. Chest pain, the second most common

presentation to the ED, adds substantial pressure to already strained clinical workflows and excessive patient volume. By accurately distinguishing between low- and high-risk patients, real-time risk stratification tools offer the prospect of reducing ED crowding by discharging low-risk patients sooner.

To improve our model, we would like to add more input features that would be available soon after a patient arrives in the ED. These desired features include ECG waveform data, chest x-ray image data, and clinical notes. Performance may be improved through site-specific retraining or recalibration at the validation site.

Acknowledgments. We thank Dr. Brian Kwan and Dr. Brett Meyers for providing valuable expertise in understanding MACE related workflows at UCSD Health. BJG acknowledges grant funding from the National Library of Medicine (award no. T15LM011271). SN acknowledges grant funding from the National Library of Medicine (award no. R01LM013998) and the National Institute of General Medical Sciences (award no. R35GM143121).

Disclosure of Interests. SN is a cofounder of Clairyon Inc., a predictive analytics startup. The terms of this arrangement have been reviewed and approved by the UC San Diego in accordance with its conflict-of-interest policies. All other authors declare no conflicts of interest.

References

1. Yukselen, Z., Majmundar, V., Dasari, M., Arun Kumar, P., Singh, Y.: Chest pain risk stratification in the emergency department: current perspectives. *Open Access. Emerg. Med.* **16**, 29–43 (2024). <https://doi.org/10.2147/OAEM.S419657>
2. Greenwood-Ericksen, M., Kamdar, N., Swenson, K., et al.: Emergency department boarding, inpatient census, and interhospital transfer acceptances. *JAMA Netw. Open* **8**(5), e2512299 (2025). <https://doi.org/10.1001/jamanetworkopen.2025.12299>
3. Moskop, J.C., Iserson, K.V., Ashton, C., et al.: Another look at the persistent moral problem of emergency department crowding. *Ann. Emerg. Med.* **74**(3), 357–364 (2019)
4. Brown, M.D., Tomaszewski, C.A., Mekwan, J., et al.: Clinical policy: critical issues in the evaluation and management of emergency department patients with suspected non-ST-elevation acute coronary syndromes. *Ann. Emerg. Med.* **72**(5), e65–e106 (2018)
5. Ke, J., Chen, Y., Wang, X., Wu, Z., Chen, F.: Indirect comparison of TIMI, HEART and GRACE for predicting major cardiovascular events. *BMJ Open* **11**(8), e048356 (2021). <https://doi.org/10.1136/bmjopen-2020-048356>
6. Ola, O., Ball, F., Lopardo, N., et al.: Use of the HEAR score for 30-day risk-stratification in emergency department patients. *Am. J. Med.* **136**(9), 918–926.e5 (2023)
7. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: *KDD*, pp. 785–794 (2016)
8. Zar, J.H.: Spearman rank correlation. In: *Encyclopedia of Biostatistics* (2005)



Towards Predicting Sedation Depth in Endoscopy with Large Clinically Annotated EEG Data of Continuous Propofol Sedation

Jakob Garbe^{1,6}, Quang Vu Nguyen², Jan W. Kantelhardt³,
Florian Dünninghaus⁴, Karla Erffmeier⁵, Katja Seeliger²,
and Thomas Schmid^{2,6,7}✉

¹ University Hospital Halle, Department of Internal Medicine I, Halle (Saale), Halle, Germany

² Martin Luther University Halle-Wittenberg, Medical Faculty, Digital Research Methods in Medicine Group, Halle (Saale), Halle, Germany
`thomas.schmid@medizin.uni-halle.de`

³ Martin Luther University Halle-Wittenberg, Faculty of Natural Sciences II, Institute of Physics, Halle (Saale), Halle, Germany

⁴ AMEOS Klinikum Warendorf, Department of Internal Medicine, Warendorf, Germany

⁵ Philipps-University Marburg, Department of Neurology, Marburg, Germany

⁶ University Hospital Halle, Research Center Data Analytics, Halle (Saale), Halle, Germany

⁷ Lancaster University Leipzig, Leipzig, Germany

Abstract. Brain function monitors are a widely used tool in general anesthesia. For monitoring drug effects in procedural sedation, however, no objective device is currently available. While available monitors offer no clinically significant benefit in sedation, the reasons are not well understood and labeled data lacking. To enable a data-driven approach for sedation depth prediction, we collected 2-channel fronto-temporal EEG signals enhanced with additional biosignals during Propofol sedation for gastrointestinal and bronchoscopic endoscopies. A total of 171 recordings were annotated in real-time with clinical events and the *State of Consciousness* (SoC) as well as clinical *Sedation Depth*. Overall, 78.5 h of multimodal biosignal data with 1129 transitions of the SoC and 7328 sedation depth data points measured on the MOAA/S scale have been collected. Applying advanced feature engineering and basic machine learning, we demonstrate that this data is suitable for predicting SoC and provides a sound foundation for research towards a reliable prediction of sedation depth. Notably, we introduce the underlying DOSE-I dataset as the first public large-scale EEG-based multimodal dataset in routine procedural sedation with rich clinical labeling.

Keywords: endoscopy · sedation · Propofol · decision support systems · electroencephalography · EEG · clinical signal processing · clinical annotation · sedation monitoring · time-series classification

1 Use Case and Related Work

Endoscopic procedures are essential for effective diagnostic evaluation and minimally invasive procedures in hollow organs such as the gastrointestinal tract and lungs. Ensuring favorable conditions for patients and endoscopists, various forms of sedation and anesthesia are used. Patients benefit from sedation as it prevents them from experiencing unpleasant sensations, such as pain or fear. Facing an increasing use of minimally invasive endoscopic procedures, endoscopists rely on sedation to maintain stable conditions throughout the procedure.

In gastrointestinal endoscopy, the most prevalent and well-researched method is *nurse administered Propofol sedation* (NAPS). Large scale retrospective and prospective trials, such as the ProSed-2 study with 368,000 endoscopies [1], found severe adverse events, such as the need for intubation, admission to an intensive care unit, or cardiopulmonary resuscitation at a rate of 0.01% and mild adverse events such as hypoxaemia, hypotension or bradycardia at a rate of 0.3%. Other studies have reported similar or higher rates of adverse events [9, 10].

Notably, however, studies in this area predominantly included healthier and fit patients, which may lead to underestimation of the risks of sedation in higher-risk patient populations. In the ProSed-2 study, for example, 79% of the patient were classified with physical status of class I or II according to the American Society of Anesthesiologists (ASA). In their risk factor analysis, the ProSed-2 study identified a higher ASA class (> 2), prolonged procedures (> 60 minutes), and therapeutic or emergency endoscopies as independent risk factors for complications. The combination of such factors – for instance a patient with cholangiosepsis undergoing a therapeutic endoscopic retrograde cholangiography – is associated with expected rates of severe adverse events exceeding 5%.

Although monitoring narcosis depth with brain function monitors (BFMs) in general anesthesia is standard practice, such devices have not been developed for the endoscopy sedation setting. Studies evaluating the utility of BFMs for procedural sedation have yielded mixed and negative results. For example, Chen et al. reported no benefit for *BiSpectral Index* (BIS) monitoring in titrating Propofol in patients undergoing colonoscopy, citing large lag times between changes in the clinical state and corresponding changes in the BIS value [2]. This finding has been corroborated by other studies [4, 5, 8].

In practice, effective monitoring sedation depth requires not only knowing current but also anticipating future sedation depth levels. To this end, a data-driven prediction of sedation states based on EEG signals has been demonstrated to be a valuable tool for monitoring [7]. With this work, we extend our previous data-driven approach and release an open-access, rigorously clinically annotated, multimodal biosignal dataset to foster further research in this domain.

2 The DOSE-I Dataset

The DOSE-I dataset is available online via this link: <https://doi.org/0.5281/zenodo.18483292>.

Patients. Data collection was carried out at the University Hospital Halle (Saale), Germany (trial registration number: DRKS00016605). Patients scheduled for non-emergency endoscopy were included if they were aged 18 years and older, capable of giving informed consent, and subject to a planned sedation duration exceeding 20 minutes. 71 patients were female (42%), ASA physical status ranged from I to III and the mean age was 60.4 (21–83). Types of endoscopies included gastroscopies, colonoscopies, endoscopic ultrasounds, bronchoscopies and endobronchial ultrasounds. A more detailed description of patients and procedures is provided in the supplementary technical documentation [6].

Recordings. Prior to the procedure, monitoring of five modalities (cf. Table 1) was established. For the electroencephalogram (EEG) recording, the patients’ forehead was degreased with alcohol to ensure impedance was below $10k\Omega$. In addition to the raw waveforms, the monitor calculated various derived parameters, including heart rate, blood oxygen saturation SpO_2 , spectral edge frequency, respiratory rate and others (recorded at a frequency of 1 Hz). The final dataset includes 78.5 h in 171 records with durations ranging from 6.7 to 70.8 min (mean: 27.5, SD: 11.6) during 281 endoscopic procedures.

Table 1. Primary biosignals provided by the DOSE-I dataset.

Biosignal	Unit	Frequency	Range	Chs	Lead Configuration
Electrocardiogram (ECG)	mV	500 Hz	$\pm 500^a$	1	RA, LA, LL
Electroencephalogram (EEG)	μV	125 Hz	± 187.5	2	Fp2-Fp1, Fz-F7, Gnd:Fpz ^b
Plethysmography (PLETH)		125 Hz	0–100	1	Finger sensor
Respiration (RESP)	Ω	62.5 Hz	−1 – 2	1	n/a
Blood Pressure (non-invasive)	mmHg	3 min		3 ^c	

^a True biosignal data within approx. $\pm 2.5mV$, higher values are mostly pacemaker artifacts.
^b For left handed patients the setup was mirrored.
^c systolic, mean and diastolic values.

Data Annotation. SoC was assessed by clinical staff using the isolated forearm technique. Transitions between consciousness and unconsciousness were confirmed by an immediately repeated assessment and recorded as either a loss or return of consciousness. Depth of sedation was documented using MOAA/S [3]. Note that both MOAA/S scores 0 (no response to deep stimulus) and 1 (no response to mild prodding or shaking) were recorded as MOAA/S score 1 in order to avoid disturbing sedation and endoscopy. Assessment of both SoC and Sedation Depth were performed every 30 s; when necessary, investigators could conduct them in shorter intervals. Further clinical events (Propofol boluses, start and end of endoscopic procedure) were recorded in real-time.

Post-processing. Event markers were manually reviewed and corrected for input errors. Each record was truncated to a window beginning 1 min before the first administration of Propofol and ending 1 min after regaining consciousness and MOAA/S 5. ECG signals were resampled to a frequency of 250 Hz. Metadata for each recording include record length, cumulative time of unconsciousness, as well as availability and completeness of biosignal channels.

3 Data-Driven Prediction of State of Consciousness

Building upon our previous work [7], we follow an approach of expert-driven feature engineering combined with established machine learning techniques for classification to predict the *next* SoC from the DOSE-I dataset. Preprocessing was implemented in C, machine learning with Python and Scikit-learn. Code is available online via <https://safe-ai-research.github.io/DOSE-I>.

Preprocessing. Raw data were transformed into 34 pEEG features [7] and supplemented additional data for each of the 171 recordings (e.g., 4 parameters from vital signs, and 6 static context variables from case report forms), yielding a comprehensive set of 44 features. With respect to the time series character of this task, we further processed this data into so-called lagged features, aggregating the respective last four time windows into an averaged value.

Prediction Models. Using Scikit-learn’s standard implementations, we applied a decision tree (dtree), a random forest with 250 trees (forest), and a simple feedforward neural network (NN, multilayer perceptron with 1 hidden layer) to predict the next SoC value via binary classification. A stratified dummy classifier, which randomly selects class labels according to the empirical class distribution in the training data, was used as a no-information baseline for comparison.

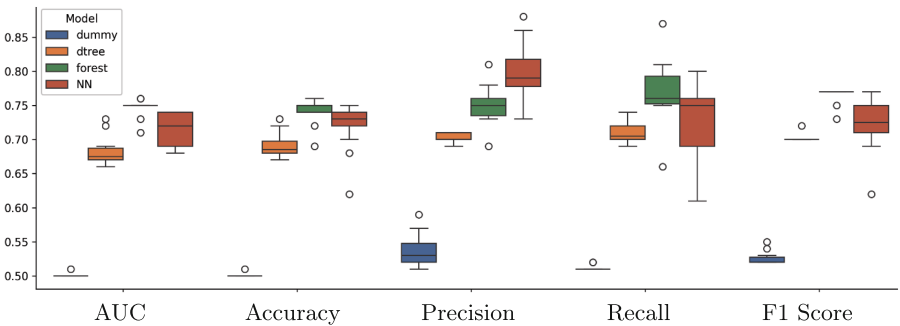


Fig. 1. Results per classifier (dummy, dtree, forest, NN) for predicting State of Consciousness with 10-fold cross validation, grouped by metric.

Training and Evaluation. For classifier training, we applied 10-fold cross-validation over the 171 recordings, i.e., in each fold 90% of the recordings were used for training and 10% for testing. With respect to the 10-fold CV evaluation of the classifiers, we report mean values ($\pm$ standard deviation) of AUC, accuracy, precision/recall, and F1 score for all three classifiers as well as the dummy classifier and display metrics across all ten folds via boxplots (Fig. 1).

Results. Mean accuracy over ten folds was higher for the random forest ($74.1\% \pm 2.1$) than for the neural network ($72.1\% \pm 3.2$) and the decision tree ($69.2\% \pm 1.9$). The same holds for F1 scores (decision tree: $70.2\% \pm 0.6$; random forest: $76.4\% \pm 1.4$; neural network: $72.3\% \pm 3.6$). Notably, mean values of all metrics are consistently higher for random forests than for decision trees, and standard deviations consistently lower for random forests than for neural networks.

4 Conclusions and Outlook

Using the DOSE-I dataset, this preliminary work demonstrates the potential of large clinically annotated EEG datasets for the prediction of State of Consciousness and Sedation Depth in endoscopy. The dataset used here is published in the context, of the *Monitoring of Sedation for Endoscopy with Artificial Intelligence* (SAFE-AI) Consortium, which aims to investigate the unique requirements of sedation monitoring as compared to general anesthesia, generate high-quality data to facilitate research on this topic and ultimately develop a device to effectively monitor sedation in the endoscopy setting. In future work, we will extend this approach from predicting binary States of Consciousness to predicting sedation depth based on the ordinal multi-level MOAA/S scale.

Acknowledgments. This work was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Project number 547230187. We thank the neurologists Dr. Josephin Damm (Halle, Germany) and Dr. Leona Möller (Marburg, Germany) for annotation of EEGs and for their insights into signal quality and bioinformatician Dr Jan Grau for sharing and discussing his early prototype SoC classifier.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Behrens, A., Kreuzmayr, A., Manner, H., et al.: Acute sedation-associated complications in GI endoscopy (ProSed 2 Study): results from the prospective multicentre electronic registry of sedation-associated complications. *Gut* **68**(3), 445–452 (2019)
2. Chen, S.C., Rex, D.K.: An initial investigation of bispectral monitoring as an adjunct to nurse-administered Propofol sedation for colonoscopy. *Am. J. Gastroenterol.* **99**(6), 1081–1086 (2004)

3. Chernik, D.A., Gillings, D., Laine, H., et al.: Validity and reliability of the observer's assessment of alertness/sedation scale: study with intravenous Midazolam. *J. Clin. Psychopharmacol.* **10**(4), 244–251 (1990)
4. von Delius, S., Salletmaier, H., Meining, A., et al.: Bispectral index monitoring of Midazolam and Propofol sedation during endoscopic retrograde cholangiopancreatography: a randomized clinical trial. *Endoscopy* **44**(3), 258–264 (2012)
5. Drake, L.M., Chen, S.C., Rex, D.K.: Efficacy of bispectral monitoring as an adjunct to nurse-administered Propofol sedation for colonoscopy: a randomized controlled trial. *Am. J. Gastroenterol.* **101**(9), 2003–2007 (2006)
6. Garbe, J.: DOSE-I: A multimodal biosignal dataset of procedural sedation for endoscopy. Technical report, arXiv (2026)
7. Garbe, J., Eisenmann, S., Kantelhardt, J.W., et al.: Capability of processed EEG parameters to monitor conscious sedation in endoscopy is similar to general anaesthesia. *UEG Journal* **9**(3), 354–361 (2021)
8. Heo, J., Jung, M.K., Lee, H.S., et al.: Effects of bispectral index monitoring as an adjunct to nurse-administered propofol combined sedation during colonoscopy: a randomized clinical trial. *Korean J. Intern. Med.* **31**(2), 260–266 (2016)
9. Qadeer, M.A., Vargo, J.J., Khandwala, F., et al.: Propofol versus traditional sedative agents for gastrointestinal endoscopy: a meta-analysis. *Clin. Gastroenterol. Hepatol.* **3**(11), 1049–1056 (2005)
10. Rex, D.K., Deenadayalu, V.P., Eid, E., et al.: Endoscopist-directed administration of Propofol: a worldwide safety experience. *Gastroenterology* **137**(4), 1229–1237 (2009)



Establishing a Data-Driven Upper Bound for CAR-T Prognosis: Quantifying the Impact of Heuristic Discretization

Francesco Olivato¹✉, Mirko Farina², Michele Malagola²,
Domenico Russo², and Roberto Gatta¹

¹ Department of Clinical and Experimental Sciences, Università degli Studi di Brescia, Brescia, Italy

`francesco.olivato@unibs.it`

² Unit of Blood Diseases and Bone Marrow Transplantation, Cell Therapies and Hematology Research Program, Department of Clinical and Experimental Science, ASST- Spedali Civili di Brescia, University of Brescia, 25123 Brescia, Italy

Abstract. Prognostic indices for survival following CAR-T therapy failure typically rely on expert-defined binary thresholds to discretize continuous features. While this point-sum approach offers interpretability, it inherently compresses feature variance. This study aims to quantify the information loss caused by heuristic discretization and establish a data-driven upper bound for prognostic performance of well known clinical indices. Using a multi-center dataset of 130 post-CAR-T patients, we compared the standard clinical scoring system with respect to a continuous, gradient-optimized model. We utilized the same set of clinical features but maintained their continuous distributions, optimizing a Cox Proportional Hazards risk function via gradient descent. To ensure robustness and generalizability, the data-driven upper bound was estimated using repeated k-fold cross-validation. Furthermore, we applied spectral clustering to the optimized risk space to identify a density-based patient stratification strategy avoiding heuristic thresholds. Results indicate that heuristic discretization incurs a significant performance penalty. The data-driven model achieved a C-index of 0.74, outperforming the clinical baseline (C-index 0.69) and reducing the log-rank p-value by three orders of magnitude (from $5 \cdot 10^{-4}$ to $1 \cdot 10^{-7}$), particularly between high and intermediate risk groups). Kaplan-Meier analysis revealed that the optimized model yielded distinct risk stratifications with almost completely non-overlapping 95% CIs, solving the issue of overlapping confidence intervals (CIs) which affects the clinical index. We conclude that current clinical indices are not capturing all the significant prognostic information, and that preserving feature continuity and leveraging a data-driven optimization approach offers a more precise, statistically rigorous benchmark for cellular therapy outcomes.

Keywords: Gradient optimization · Spectral clustering · Information loss · CAR-T therapy · Prognostic indices · Cox Proportional Hazards · Patient stratification · C-index · Hematologic malignancies

1 Introduction

In the clinical setting, the gold standard for patient evaluation often relies on prognostic indices that are simple, and easy to use at the bedside. In emathology, in particular, in predicting post-CAR-T therapy relapses, indices such as the Endothelial Activation and Stress Index (EASIX) [7] and the Post-CAR Prognostic Index (PC-PI) [3] are widely utilized. These tools typically rely on point-sum scoring systems, which discretize continuous biological markers by the application of thresholds.

However, a fundamental question remains: *How optimized are those thresholds?* The lack of known benchmarks, arises the problem of determining whether an index is truly optimal or merely the best among heuristic approximations. Discretizing continuous variables inherently compresses feature variance, leading to an invisible yet significant loss of prognostic information.

To address this gap, this work proposes a data-driven artificial intelligence methodology to identify an optimal index. This optimized model serves as a benchmark, of prognostic accuracy given a specific feature space. By establishing the absolute ceiling of predictive capacity, we can quantify the exact amount of information currently captured by existing clinical indices versus the amount lost due to heuristic thresholding.

While demonstrated on CAR-T relapses, this work proposes a methodological framework rather than a novel state-of-the-art model. By intentionally optimizing a standard clinical Cox model, we isolate the performance gained purely by eliminating heuristic discretization.

2 Material and Methods

2.1 Dataset and Clinical Features

This study utilized a multi-center dataset comprising 130 patients who underwent CAR-T cell therapy. To ensure a fair comparison, the feature space was strictly limited to the standard clinical biomarkers currently used to compute the EASIX and PC-PI scores.

2.2 Baseline Clinical Indices

Baseline evaluations utilized the core EASIX and PC-PI formulations, strictly retaining standard multivariate features. As adapting established prognostic indices is a recognized methodological practice to enforce equitable comparisons, baselines were structurally harmonized. To ensure a uniform three-tier evaluation, EASIX was tertiled according to clinical literature norms, while the intermediate PC-PI cohorts were simply merged. Retaining these fixed, literature-based thresholds ensures the baselines reflect actual clinical practice, focusing the comparison strictly on standard discretization limitations.

2.3 Data-Driven Optimization

To construct the benchmark index, we maintained the standard clinical feature space to ensure a stable events-per-variable ratio. Survival probabilities were modeled using a Cox Proportional Hazards (CPH) [1] risk function optimized via gradient descent. Robustness was ensured through repeated k-fold cross-validation. Crucially, to strictly prevent information leakage, the model’s single hyperparameter was tuned entirely internally within the training folds, mitigating overfitting risks on the 130-patient cohort.

2.4 Spectral Clustering for Stratification

To translate the continuous optimized risk space back into clinically actionable groups, we applied spectral clustering [6] [5] to the model’s output. Unlike heuristic binning, spectral clustering identifies natural, density-based stratifications within the optimized hazard ratios, allowing the data structure itself to define the boundaries between low, intermediate, and high-risk patient groups.

3 Results and Discussion

3.1 Kaplan-Meier Survival Analysis

The performance of the traditional indices was visually and statistically compared against the proposed data-driven benchmark using Kaplan-Meier(KM) [4] survival curves.

As observed in the top row of Fig. 1, the standard clinical scores struggle with distinct patient stratification. A recurrent issue across both baseline models is the presence of heavily overlapping 95% Confidence Intervals (CIs), particularly between the intermediate and high-risk patient cohorts. This overlap empirically demonstrates the information loss caused by heuristic discretization.

Conversely, our data-driven solution (Fig. 1, bottom-left) demonstrates a highly distinct risk stratification. By utilizing spectral clustering on a continuous gradient-optimized space, the resulting patient groups exhibit almost completely non-overlapping CIs. The pairwise log-rank tests (Fig. 1, bottom-right) further validate this, showing significantly stronger distinction between the intermediate and high-risk groups compared to standard baselines.

3.2 Performance Benchmarking

To quantify this overall improvement, C-indices [2] were calculated for all models (Fig. 1, bottom-right).

The data-driven model achieved a C-index of 0.74, substantially outperforming the highest-scoring clinical baseline (C-index 0.69). Furthermore, the optimized model reduced the overall log-rank p-value by three orders of magnitude (from $5 \cdot 10^{-4}$ to $1 \cdot 10^{-7}$).

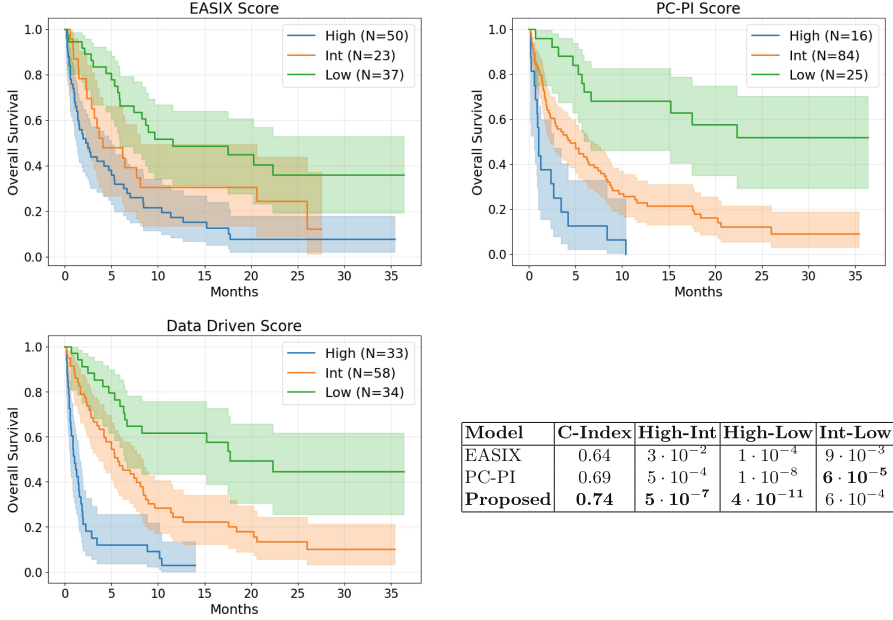


Fig. 1. Kaplan-Meier survival analysis and performance metrics. **Top-Left:** EASIX. **Top-Right:** PC-PI. **Bottom-Left:** Proposed gradient-optimized continuous feature model. **Bottom-Right:** Concordance index (C-index) and pairwise log-rank test p-values between risk stratifications.

Crucially, the benchmark C-index of 0.74 represents the theoretical upper bound of this specific dataset. It confirms that while the standard indices are highly practical, they are currently leaving significant predictive power on the table. The fact that the benchmark is not 1.0 validates our hypothesis: it accurately reflects the actual limit of prognostic information contained within standard biomarkers, defining exactly how much a clinical index can realistically be improved before requiring entirely new types of medical data.

4 Conclusion and Future Work

Current clinical best practices heavily rely on easy-to-use prognostic indices always validated but often without a parameter optimization: as shown in our analysis of post-CAR-T relapses, the heuristic discretization of continuous variables incurs a severe performance penalty. By maintaining feature continuity and optimizing a Cox Proportional Hazards model via gradient descent, we established a data-driven upper bound for prognostic accuracy.

Our optimized model outperformed established scores (EASIX, PC-PI), achieving a C-index of 0.74 and generating distinct, density-based patient stratifications with much more robust confidence intervals. This work provides a

fundamental methodological shift: rather than just proposing a new score, we provide a mathematical benchmark to measure the true information capacity of a threshold-based scoring index. Future work will focus on extending this benchmarking methodology to other complex prognostic settings and the further development of a software library for indeces benchmarking.

Acknowledgments. We sincerely thank the healthcare professionals across all participating centers. While not involved in this computational study, their rigorous clinical work and data collection provided the essential foundation for this research.

Roberto Gatta's collaboration is funded by the Ministerial Project -M.U.R.-Department of Excellence 2023–2027 assigned to the Department of Clinical and Experimental Sciences of the University of Brescia.

References

1. Cox, D.R.: Regression models and life-tables. *J. Roy. Stat. Soc.: Ser. B (Methodol.)* **34**(2), 187–220 (1972)
2. Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. *JAMA* **247**(18), 2543–6 (1982)
3. Iacoboni, G., et al.: Development and validation of the post-CAR prognostic index for large b-cell lymphoma patients after car-t progression in third or later line treatment. *J. Hematol. Oncol.* **17** (2024)
4. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**(282), 457–481 (1958)
5. von Luxburg, U.: A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416 (2007). <https://api.semanticscholar.org/CorpusID:3264198>
6. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: *Proceedings of the 15th International Conference on Neural Information Processing Systems: Natural and Synthetic*, pp. 849–856. NIPS'01, MIT Press, Cambridge, MA, USA (2001)
7. Penack, O., et al.: Endothelial activation and stress index (EASIX) to predict mortality after allogeneic stem cell transplantation: a prospective study. *J. Immunother. Cancer* **12**, e007635 (2024)



PreClaim-GCT: Self-supervised Graph-Transformer Pretraining for Incident Outcome Prediction from Healthcare Claims

Rui Wang^{1(✉)}, Ming-Ching Chang¹, and Emily Leckman-Westin²

¹ University of Albany, Albany, NY 12222, USA
rui.wang@omh.ny.gov, mchang2@albany.edu

² NYS Office of Mental Health, Albany, NY 12229, USA
emily.leckman-westin@omh.ny.gov

Abstract. Administrative claims data provide large-scale coverage of healthcare utilization but pose challenges for predictive modeling due to sparse, high-dimensional coding and limited clinical structure, particularly in psychiatric populations. Existing graph and transformer approaches often rely on fully supervised training or longitudinal sequence modeling, which may not fully exploit within-encounter structure. We propose **PreClaim-GCT**, a graph-transformer framework for encounter-level representation learning using self-supervised masked code modeling. Each inpatient encounter is represented as a graph with heterogeneous diagnosis and medication codes anchored at a central node. The model is pretrained using masked code reconstruction to capture within-encounter co-occurrence patterns without outcome labels, and then fine-tuned for incident post-discharge risk prediction. Across outcomes and horizons, self-supervised pretraining improves performance over strong baselines and the same architecture without pretraining. These results demonstrate that encounter-level self-supervised learning enhances post-discharge risk prediction in administrative claims data.

Keywords: heterogeneous graph · graph transformer · self-supervised learning · administrative claims data · healthcare prediction

1 Introduction

Administrative healthcare claims data provide large-scale coverage of healthcare utilization but pose challenges for machine learning, including sparse high-dimensional coding, limited clinical context, and weak supervision—particularly for predicting conditions following inpatient discharge. These challenges are amplified in psychiatric and comorbid populations, where complex diagnosis and medication patterns complicate representation learning [1,2]. Graph and graph-transformer models improve modeling of structured medical data [3,4],

but many approaches rely on fully supervised training or longitudinal sequence modeling [5, 6], which may be less effective for claims data with sparse labels. We introduce **PreClaim-GCT** (Fig. 1), a self-supervised graph-transformer model for encounter-level representation learning from administrative claims data. Each encounter is represented as a heterogeneous graph of diagnosis and medication nodes anchored at a central node. Unlike prior work, PreClaim-GCT excludes inter-encounter edges to isolate within-encounter structural learning and is pre-trained using masked code reconstruction without outcome labels, followed by supervised fine-tuning for incident post-discharge prediction. Empirically, self-supervised pretraining improves performance over strong baselines and generalizes across outcomes with different prevalence levels.

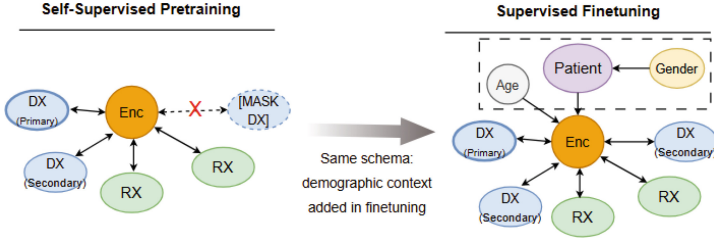


Fig. 1. Encounter-level graph of PreClaim-GCT with masked code pretraining.

2 Methods

Encounter Graph Construction. Each encounter is represented as a heterogeneous graph $G = (V, E)$ (Fig. 1). The graph contains three node types: *Enc* (encounter), *DX* (diagnosis), and *RX* (medication), embedded in a shared space. Primary and secondary diagnoses use separate edge types. During supervised fine-tuning, patient, age, and gender nodes are added for demographic conditioning. A reserved [MASK] token is used during self-supervised pretraining. A subset of diagnosis and medication nodes is masked using [MASK]. Edges connected to masked nodes are removed to prevent identity leakage. Pretraining uses encounter-code relations, while fine-tuning adds patient and demographic edges; no edges connect distinct encounters. **Heterogeneous Graph-Transformer Encoder.** All nodes are embedded in a shared d -dimensional space with relation-specific transformations. Diagnosis and medication embeddings are initialized using code hierarchy. At layer ℓ , node representations $\mathbf{h}_v^{(\ell)}$ are updated via relation-aware attention over neighbors $\mathcal{N}_r(v)$ with $H = 2$ heads. Query and key projections are $\mathbf{q}_v^{(k)} = Q_k \mathbf{h}_v^{(\ell)}$ and $\mathbf{k}_u^{(k)} = K_k \mathbf{h}_u^{(\ell)}$, with attention score $e_{uv}^{(k,r)} = \frac{(\mathbf{q}_v^{(k)})^\top \mathbf{k}_u^{(k)}}{\sqrt{d_k}}$. Attention is normalized within each relation:

$$\alpha_{uv}^{(k,r)} = \frac{\exp(e_{uv}^{(k,r)})}{\sum_{u' \in \mathcal{N}_r(v)} \exp(e_{u'v}^{(k,r)})}. \quad (1)$$

Messages are transformed as $\mathbf{m}_{u \rightarrow v}^{(r)} = W_r \mathbf{h}_u^{(\ell)}$, and node updates are

$$\mathbf{h}_v^{(\ell+1)} = \text{LayerNorm} \left(\mathbf{h}_v^{(\ell)} + \sum_r \sum_{k=1}^H \sum_{u \in \mathcal{N}_r(v)} \alpha_{uv}^{(k,r)} V_k \mathbf{m}_{u \rightarrow v}^{(r)} \right). \quad (2)$$

The final encounter-node representation is used for downstream prediction.

Masked Code Modeling with In-Batch Contrastive Loss. PreClaim-GCT is pretrained using masked code modeling on heterogeneous encounter graphs. Let $\mathcal{C}$ denote diagnosis and medication nodes and let $\mathcal{M} \subset \mathcal{C}$ be the masked subset sampled with masking rate ρ . For each masked encounter-code pair (v_k, c_k) , a positive pair is formed between encounter embedding $\mathbf{h}_{v_k}$ and code embedding $\mathbf{e}_{c_k}$. All other masked codes in the minibatch serve as implicit negatives. We optimize an in-batch InfoNCE objective [7]:

$$\mathcal{L}_{\text{SSL}} = - \sum_{k=1}^M \log \frac{\exp(\text{sim}(\mathbf{h}_{v_k}, \mathbf{e}_{c_k})/\tau)}{\sum_{j=1}^M \exp(\text{sim}(\mathbf{h}_{v_k}, \mathbf{e}_{c_j})/\tau)}, \quad (3)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is the temperature parameter (set to 0.07). The pretrained encoder is later fine-tuned end-to-end for supervised incident outcome prediction.

3 Downstream Tasks and Experimental Setup

Study Cohort and Prediction Setting. We used New York State (NYS) Medicaid claims data (2016–2023), including individuals with at least one psychiatric hospitalization during 2016–2021. The cohort comprised approximately 159,000 patients and 759,000 inpatient encounters (16.9 codes per encounter). Data were split at the patient level into training (70%), validation (15%), and test (15%) sets. All training data were used for self-supervised pretraining, while 20% of training patients were sampled for supervised fine-tuning and resampled five times. Continuous enrollment was required for downstream prediction, with 365 days prior to admission and 365 or 730 days post-discharge. Each prediction instance corresponds to an index encounter (defined here as an inpatient hospitalization) represented as the heterogeneous graph in §2.

Incident Post-discharge Outcomes. We evaluate two tasks using 365- and 730-day windows: **new-onset diabetes** and **new-onset depression**. Incident outcomes are defined as the first diagnosis within the window among individuals without history; task-specific eligibility criteria determine the final evaluation cohort (i.e., 315k for diabetes-365).

Baseline Models and Evaluation Setup. We compare against logistic regression (LR), XGBoost, graph neural networks (GCN [8], GAT [9]), a BEHRT-style transformer [5], and two variants of PreClaim-GCT. All models use identical patient-level splits and input features to ensure fair comparison and are evaluated with AUC-ROC and AUC-PR. LR

and XGBoost use multi-hot features; GCN and GAT use the same graph; and BEHRT uses diagnosis sequences without explicit graph edges. PreClaim-GCT (Random) and (SSL) share the same architecture, isolating the effect of self-supervised pretraining. **Training Details.** Models were trained using Adam with hidden dimension 128, three layers and two attention heads. During pre-training, 15% of nodes are masked and reconstructed using an in-batch InfoNCE objective with temperature $\tau = 0.07$, consistent with contrastive learning [10]. The pretrained encoder is then fine-tuned end-to-end with early stopping based on validation AUC-PR.

Table 1. Incident outcome prediction (mean $\pm$ std across five resampled splits). Best performance in each column is shown in bold.

Model	Diabetes		Depression	
	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
365 days				
LR	0.632 $\pm$ 0.017	0.110 $\pm$ 0.006	0.589 $\pm$ 0.008	0.285 $\pm$ 0.007
XGBoost	0.673 $\pm$ 0.010	0.130 $\pm$ 0.003	0.618 $\pm$ 0.009	0.313 $\pm$ 0.007
GCN [8]	0.675 $\pm$ 0.011	0.121 $\pm$ 0.007	0.597 $\pm$ 0.005	0.284 $\pm$ 0.009
GAT [9]	0.662 $\pm$ 0.021	0.116 $\pm$ 0.010	0.583 $\pm$ 0.006	0.275 $\pm$ 0.006
BEHRT [5]	0.678 $\pm$ 0.009	0.121 $\pm$ 0.006	0.606 $\pm$ 0.013	0.297 $\pm$ 0.011
PreClaim-GCT (Random)	0.677 $\pm$ 0.004	0.118 $\pm$ 0.005	0.623 $\pm$ 0.008	0.309 $\pm$ 0.005
PreClaim-GCT (SSL)	0.690 $\pm$ 0.002	0.134 $\pm$ 0.003	0.633 $\pm$ 0.004	0.322 $\pm$ 0.005
730 days				
LR	0.641 $\pm$ 0.019	0.186 $\pm$ 0.010	0.596 $\pm$ 0.006	0.399 $\pm$ 0.005
XGBoost	0.675 $\pm$ 0.011	0.208 $\pm$ 0.004	0.627 $\pm$ 0.009	0.432 $\pm$ 0.012
GCN [8]	0.670 $\pm$ 0.007	0.197 $\pm$ 0.005	0.596 $\pm$ 0.022	0.389 $\pm$ 0.022
GAT [9]	0.650 $\pm$ 0.014	0.178 $\pm$ 0.011	0.586 $\pm$ 0.027	0.380 $\pm$ 0.024
BEHRT [5]	0.681 $\pm$ 0.003	0.206 $\pm$ 0.008	0.614 $\pm$ 0.014	0.415 $\pm$ 0.015
PreClaim-GCT (Random)	0.671 $\pm$ 0.009	0.197 $\pm$ 0.009	0.619 $\pm$ 0.001	0.423 $\pm$ 0.006
PreClaim-GCT (SSL)	0.686 $\pm$ 0.003	0.215 $\pm$ 0.004	0.634 $\pm$ 0.014	0.444 $\pm$ 0.007

4 Results

Table 1 reports incident prediction performance. PreClaim-GCT achieves the best performance. For diabetes, pretraining improves ROC by $\sim 2\%$ and AUC-PR by 10–15% over the randomly initialized model, with similar gains observed for depression. These trends persist at the 730-day horizon, where PreClaim-GCT (SSL) outperforms all baselines, including XGBoost, BEHRT, and the same architecture without pretraining. **Statistical Significance Analysis.**

We assessed statistical significance following [11] using a Friedman test across all seven models based on AUC-PR. The analysis includes five repeated sub-training splits and four prediction settings, yielding 20 evaluation blocks. The Friedman test indicated significant differences across models ($\chi^2 = 74.16$, $p < 0.001$). PreClaim-GCT (SSL) achieved the best average rank (1.13), followed by XGBoost (2.09) and BEHRT (3.75). Post-hoc Wilcoxon signed-rank tests with Holm correction confirmed that PreClaim-GCT (SSL) significantly outperformed PreClaim-GCT (Random), XGBoost, and BEHRT ($p < 0.001$).

5 Discussion and Conclusion

Interpretation of Representation Gains. Larger improvements for depression suggest that masked code modeling captures dense co-occurrence patterns, whereas diabetes prediction may depend more on longer-term processes. Consistent gains indicate improved representation quality. Although gains over strong baselines are modest, comparison between PreClaim-GCT (Random) and PreClaim-GCT (SSL) isolates the effect of self-supervised pretraining. Because both models share the same encoder, improvements arise from representation learning. In claims data with sparse labels, even modest AUC-PR gains remain meaningful. **Implications for Claims-Based Modeling.** The framework leverages diagnosis–medication co-occurrence to learn relational structure without curated ontologies, making it suitable for large-scale claims data. Operating at the encounter level, it generalizes across outcomes without requiring longitudinal modeling. **Conclusion and Limitations.** We presented PreClaim-GCT, a graph–transformer model combining self-supervised masked code pretraining with supervised fine-tuning for incident outcome prediction from claims data. The approach consistently outperforms strong baselines, demonstrating the effectiveness of encounter-level self-supervised representation learning. This study focuses on a single state claims system. Hyperparameters were selected based on validation performance and common practice; systematic sensitivity analysis is left for future work. Future work will evaluate generalizability and explore integration with longitudinal modeling.

References

1. Iezzoni, L.I. (ed.): Risk Adjustment for Measuring Health Care Outcomes, 4th edn. Health Administration Press, Chicago (2013)
2. Walkup, J.T., et al.: Trends in psychiatric hospitalization among children and adolescents. *J. Am. Acad. Child Adolesc. Psychiatry* **55**(6), 514–522 (2016)
3. Choi, E., et al.: Learning the graphical structure of electronic health records with graph convolutional transformer. In: *Proceedings of AAAI*, pp. 606–613 (2020)
4. Hu, Z., Dong, Y., Wang, K., Sun, Y.: Heterogeneous graph transformer. In: *Proceedings of WWW*, pp. 2704–2710 (2020)
5. Li, Y., et al.: BEHRT: transformer for electronic health records. *Sci. Rep.* **10**(1), 1–12 (2020)

6. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit. Med.* **4**(1), 86 (2021)
7. Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2018). [arXiv:1807.03748](https://arxiv.org/abs/1807.03748) arXiv preprint
8. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations (ICLR 2017). OpenReview.net (2017)
9. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: Proceedings of ICLR (2018)
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of International Conference on Machine Learning (ICML), pp. 1597–1607 (2020)
11. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**(1), 1–30 (2006)



Artificial Intelligence–Driven Wearable Technologies for Early Cardiovascular Disease Prediction: A Narrative Review

Lariza Noletto¹ , Amanda Guimarães² , Pedro Coelho¹ , Diandra Noletto² ,
Matheus Moraes³ , and Marcos Nascimento²

¹ Federal University of Maranhão, São Luís, MA, Brazil

² State University of Maranhão, São João dos Patos, MA, Brazil
marcosdonascimento@professor.uema.br

³ Institute of Science and Technology of the Federal University of São Paulo,
São José dos Campos, SP, Brazil

Abstract. Recent advances in artificial intelligence (AI), machine learning, and data analytics have transformed the acquisition and interpretation of large-scale physiological data, enabling real-time cardiovascular monitoring through wearable devices. This study aimed to analyze the contributions of AI-integrated wearable systems to early cardiovascular disease monitoring and risk prediction. A PRISMA-informed narrative review was conducted using PubMed, SciELO, Web of Science, and Scopus, including studies published between 2020 and 2025. Seven studies met the inclusion criteria and were critically analyzed. The findings indicate that AI-driven wearable systems effectively process multimodal data—such as blood pressure, physical activity, and behavioral patterns—supporting continuous monitoring, early detection, and automated risk assessment. These technologies were associated with clinically relevant reductions in blood pressure, improved patient engagement, and enhanced self-management. Despite these advances, such systems should be considered complementary clinical decision-support tools. In conclusion, AI-integrated wearable technologies represent a scalable and promising approach for cardiovascular risk prediction, highlighting the need for further validation and real-world application.

Keywords: Computer Science · Machine Learning · Health

1 Introduction

Wearable technologies are compact, body-integrated devices that can be classified as primary (e.g., smartwatches) or secondary (e.g., chest straps), enabling continuous physiological monitoring through integrated or transmitted data. Their accessibility and capacity for long-term self-monitoring contribute to improved patient adherence and reduced healthcare costs [1].

Advances in digital health have increased the demand for innovative disease detection strategies. A key milestone was the approval of the Apple Watch by the FDA for atrial

fibrillation monitoring, demonstrating high diagnostic accuracy [2]. Currently, wearable technologies extend beyond arrhythmia detection to include hypertension management, heart failure monitoring, and broader cardiovascular risk control. Studies have shown that devices such as Fitbit Flex 2 and ActiGraph GT3X can reduce systolic blood pressure (SBP) when integrated into exercise programs, while digital therapeutic systems such as HERB-DH1 improve hypertension control through combined monitoring and lifestyle interventions [3, 4].

The integration of artificial intelligence (AI) has further enhanced these systems by enabling automated analysis of large-scale multimodal data. AI-driven models, including convolutional neural networks (CNNs) and long short-term memory (LSTM), have demonstrated high accuracy in detecting cardiovascular abnormalities and predicting risk, supporting early and personalized interventions [5, 6].

In this context, the present study aims to analyze the contributions of artificial intelligence in healthcare when integrated with wearable technologies for early monitoring and prediction of cardiovascular diseases.

2 Methodological Framework

This study was conducted as a narrative review informed by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure methodological transparency and rigor.

A systematic search was performed in the PubMed, Scopus, Web of Science, and SciELO databases, including studies published between 2020 and 2025. The search strategy combined descriptors related to artificial intelligence, wearable technologies, and cardiovascular diseases, using Boolean operators (AND/OR).

Studies were included if they involved the application of artificial intelligence to wearable technologies with a focus on cardiovascular monitoring or prediction, including randomized clinical trials, observational studies, and technological developments. Editorials, letters, and duplicate records were excluded.

The study selection process followed PRISMA stages, including identification, screening, eligibility, and inclusion (Fig. 1). Two reviewers independently conducted the selection, and disagreements were resolved by consensus.

Data extraction included study design, sample characteristics, AI techniques, wearable devices, and main outcomes. A qualitative synthesis was performed, focusing on predictive performance, clinical relevance, and applicability of AI-driven wearable systems.

The guiding research question was defined as follows: What are the contributions of artificial intelligence in healthcare when integrated with wearable devices for the early monitoring of cardiovascular diseases?

3 Results and Discussion

Seven studies were included in this review: Rodrigues [7], Persell et al. [8], Leitner et al. [9], Branch et al. [10], Grazit et al. [11], Kario et al. [4], and Paz et al. [12], selected due to their methodological alignment with the study objective and relevance to AI-integrated wearable technologies for cardiovascular monitoring.

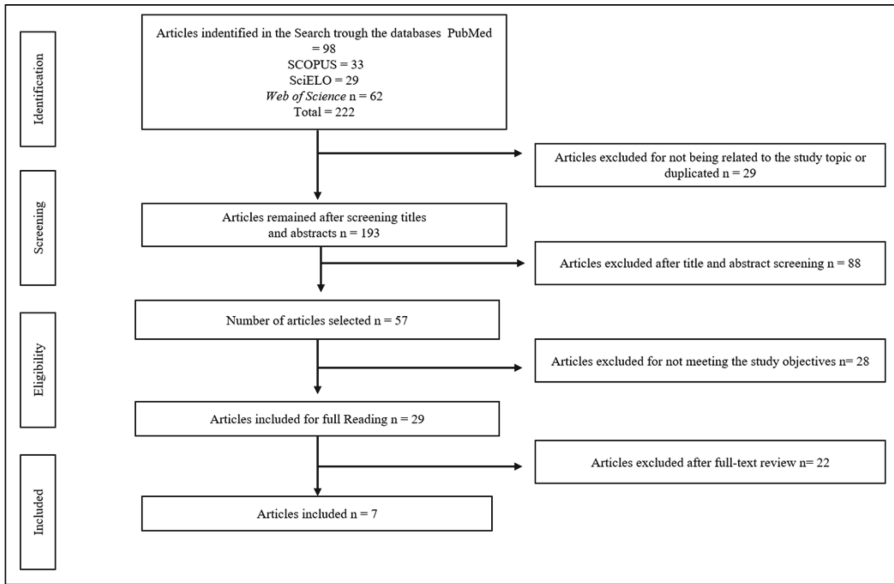


Fig. 1. PRISMA 2020 flow diagram of the study selection process. The diagram illustrates the identification, screening, eligibility assessment, and final inclusion.

The findings indicate a consistent trend in which the integration of AI with wearable devices improves blood pressure (BP) monitoring, particularly through clinically meaningful reductions in systolic blood pressure (SBP) [4, 7–12]. These results reinforce the growing evidence supporting the effectiveness of AI-driven digital health interventions.

Rodrigues [7] highlighted the efficiency of wearable devices in remote BP monitoring, enabling continuous data acquisition and real-time clinical decision support. Similarly, Persell et al. [8] demonstrated that AI-supported applications enhance self-monitoring behaviors and patient engagement, despite modest BP reductions.

Leitner et al. [9] reported significant reductions in SBP and diastolic BP through AI-driven personalized interventions, illustrating the role of machine learning in adaptive digital therapeutics. In line with these findings, Branch et al. [10] observed greater BP reductions in high-risk individuals, emphasizing the importance of AI-based risk stratification.

Additionally, AI-integrated systems enable continuous monitoring of physiological signals and early detection of abnormal cardiovascular patterns, supporting preventive healthcare strategies. Grazit et al. [11] and Paz et al. [12] further demonstrated that mobile health programs contribute to long-term BP control and improvements in cardiometabolic risk factors.

Despite these advances, AI-based wearable technologies should be considered complementary tools rather than replacements for clinical decision-making. Their effective implementation requires integration into healthcare systems, ethical data management, and robust clinical validation [4].

4 Conclusions

The findings of this review indicate that the integration of artificial intelligence with wearable technologies is a promising strategy for continuous monitoring and early prediction of cardiovascular diseases, particularly hypertension. The analyzed studies consistently reported clinically relevant reductions in systolic blood pressure, reinforcing the effectiveness of AI-enabled digital health interventions.

From a health promotion perspective, these technologies support personalized, data-driven interventions based on real-time physiological and behavioral metrics. However, they should be considered complementary tools rather than replacements for clinical expertise.

Their effective implementation requires integration into healthcare systems, sustained patient engagement, and further validation through robust clinical research. Future studies should focus on model transparency, generalizability, and long-term applicability in diverse populations.

Acknowledgments. We would like to express our sincere gratitude to the Research Group in Physiology, Nutrition, and Exercise, as well as to the Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA), for their invaluable support throughout this project.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Tulcanaza Ochoa, G.F., et al.: Efficacy of wearable cardiac monitoring devices versus traditional methods in detecting atrial fibrillation: a systematic review and meta-analysis. *Salud, Ciencia y Tecnología*. **4**, 962 (2024)
2. Peppinkhuizen, S., et al.: Accuracy and clinical relevance of the single-lead apple watch electrocardiogram to identify atrial fibrillation. *Cardiovasc. Digit. Health J.* **3**, 17–22 (2022)
3. Roberts, M.C., et al.: Precision public health in the era of genomics and big data. *Nat. Med.* **7**, 1865–1873 (2024)
4. Kario, K., et al.: Efficacy of a digital therapeutics system in the management of essential hypertension: the HERB-DH1 pivotal trial. *Eur. Heart J.* **40**, 4111–4122 (2021)
5. Wang, R., Veera, S.C.M., Asan, O., Liao, T.: A systematic review on the use of consumer-based ECG wearables on cardiac health monitoring. *IEEE J. Biomed. Health Inform.* **11**, 6525–6537 (2024)
6. Kanagamalliga, S.: Advancements in remote heart monitoring: wearable technology and AI-based approaches for cardiovascular disease detection. In: *AI in the Social and Business World: A Comprehensive Approach*, vol. 1, pp. 102–117. Bentham Science Publishers (2024)
7. Rodrigues, R.J.S.: SBIdC-MPH: sistema baseado em internet das coisas para o monitoramento de pacientes com hipertensão. In: *Dissertação (Mestrado em Engenharia)*. Universidade Federal de São Carlos, Repositório Institucional UFSCar (2022)
8. Persell, S.D., et al.: Effect of home blood pressure monitoring via a smartphone hypertension coaching application or tracking application on adults with uncontrolled hypertension: a randomized clinical trial. *JAMA Netw. Open.* **3**(3), e200255 (2020)

9. Leitner, J., Chiang, P.H., Agnihotri, P., Dey, S.: The effect of an AI-based, autonomous, digital health intervention using precise lifestyle guidance on blood pressure in adults with hypertension: single-arm nonrandomized trial. *JMIR Cardio*. **8**, e51916 (2024)
10. Branch, O.H., et al.: Relationships between blood pressure reduction, weight loss, and engagement in a digital app-based hypertension care program: observational study. *JMIR Form. Res.* **6**(10), e38215 (2022)
11. Gazit, T., Gutman, M., Beatty, A.L.: Assessment of hypertension control among adults participating in a mobile technology blood pressure self-management program. *JAMA Netw. Open*. **4**(10), e2127008 (2021)
12. Paz, E., et al.: Comprehensive cardiovascular risk factor control with a mobile health cardiovascular risk self-management program. *J. Am. Heart Assoc.* **13**(10), e033328 (2024)



Robustness of Transformer-Based Fluence Map Prediction Under Clinically Realistic Perturbations

Ujunwa Mgboh¹, Rafi Ibn Sultan¹, Joshua Kim², Kundan Thind²,
and Dongxiao Zhu¹(✉)

¹ Wayne State University, Detroit, MI, USA

{[ujunwa.mgboh](mailto:ujunwa.mgboh@wayne.edu),[rafis](mailto:rafis@wayne.edu),[dzhu](mailto:dzhu@wayne.edu)}@wayne.edu

² Henry Ford Health, Detroit, MI, USA

{[JKIM8](mailto:JKIM8@hfhs.org),[kthind1](mailto:kthind1@hfhs.org)}@hfhs.org

Abstract. Learning-based fluence map prediction offers a fast alternative to iterative inverse planning in intensity-modulated radiation therapy (IMRT), but its robustness under realistic distribution shifts remains unclear. We study a two-stage transformer pipeline that maps anatomy (CT and contours) to dose and then to beamlet fluence maps. We compare fluence-stage transformer backbones with hierarchical, global, and hybrid attention, trained with a physics-informed loss enforcing energy consistency. Robustness is evaluated under geometric perturbations, radiometric noise, reduced training data, and domain shifts using a prostate IMRT dataset, with additional evaluation of the dose stage on public datasets. Results show smooth degradation under moderate perturbations but sharp failures under severe rotations and noise. Hierarchical transformers (e.g., SwinUNETR) exhibit slower growth in upper-quartile energy error, indicating improved robustness. We further show that SSIM alone fails to capture clinically relevant errors, highlighting the need for physics-informed evaluation.

1 Introduction

Radiotherapy planning operates under unavoidable uncertainty in imaging, patient setup, and anatomy, exposing learning-based models to distribution shifts at deployment. While deep learning has shown strong performance in dose and fluence prediction [1–3], most evaluations assume stable data distributions and do not assess robustness under realistic perturbations encountered in clinical workflows.

In medical AI, models are known to degrade under dataset shift, noise, and geometric perturbations [4, 5]. In radiotherapy, prior robustness studies have focused primarily on segmentation and dose prediction [6], often evaluating sensitivity at the level of contours or dose distributions. Direct fluence map prediction poses additional challenges: fluence maps are constrained by beam geometry and energy conservation, and small perturbations can induce non-linear, clinically significant deviations [7].

Unlike prior work that emphasizes architectural improvements, we treat fluence prediction as a robustness evaluation problem. Building on the FluenceFormer framework [8], we perform a systematic stress test under deployment-relevant perturbations, including geometric misalignment, radiometric noise, domain shifts, and limited training data. We analyze how architectural choices affect degradation and demonstrate that physics-based metrics reveal failure modes not captured by image similarity alone. This reframes fluence prediction as a robustness auditing task and highlights the importance of reliability assessment for safe clinical deployment.

2 Methods

2.1 Problem Formulation and Evaluation Setup

We model IMRT planning as a two-stage pipeline. Let $I \in \mathbb{R}^{H \times W \times Z}$ denote the planning CT, where H , W , and Z are spatial dimensions, and $M \in \{0, 1\}^{K \times H \times W \times Z}$ the anatomical masks.

The first stage predicts dose:

$$g_\phi(I, M) \rightarrow \hat{D}, \quad (1)$$

and the second stage predicts beamlet fluence maps:

$$f_\theta(\hat{D}, B) \rightarrow \hat{F}, \quad (2)$$

where $\hat{D}$ and $\hat{F}$ denote predicted dose and fluence, and B encodes beam geometry.

2.2 Model and Training

We adopt the FluenceFormer framework [8] and compare transformer backbones including SwinUNETR, UNETR, nnFormer, and MedFormer under a shared pipeline.

Training uses a Fluence-Aware Regression (FAR) loss:

$$\mathcal{L}_{FAR} = \alpha \mathcal{L}_{MSE} + \beta \mathcal{L}_{Grad} + \gamma \mathcal{L}_{Corr} + \delta \mathcal{L}_{Energy}, \quad (3)$$

where $\mathcal{L}_{MSE}$, $\mathcal{L}_{Grad}$, $\mathcal{L}_{Corr}$, and $\mathcal{L}_{Energy}$ denote voxel, gradient, correlation, and energy terms, and $\alpha, \beta, \gamma, \delta$ are weights.

2.3 Datasets and Evaluation

We use a 99-case prostate IMRT dataset with CT, contours, dose, and fluence maps, and evaluate cross-dataset generalization on OpenKBP [9] and CORT [10] (CT, contours, dose only). Models are trained on 128×128 grids with normalized CT.

Performance is measured using Structural Similarity Index Measure (SSIM) and upper-quartile energy error ($E_{err}^{Q_{75}}$) to capture structural fidelity and clinically relevant worst-case deviations. We also assess DVHs on public datasets for clinical relevance.

2.4 Robustness Evaluation

We evaluate robustness under four deployment-relevant perturbations: geometric shifts (3–5 mm) and rotations (2° – 5°), Gaussian noise ($\sigma = 0.05$ – 0.20), intensity bias fields, and reduced training data. These perturbations are motivated by robustness evaluation in radiotherapy, where setup uncertainties and worst-case analysis (e.g., ~ 3 mm deviations) assess plan sensitivity [11, 12].

All perturbations are applied at inference to stress-test both stages of the pipeline, enabling analysis of failure modes under structured distribution shifts.

3 Results

3.1 Geometric Robustness

Transformer models show stable performance under shifts but degrade under rotations (Fig. 1), with increased energy error. FluenceFormer backbones are substantially more robust than baselines: at 2° , SwinUNETR achieves 7.32% error vs 17.70% for CNN [13] ($>2\times$ lower), with similar SSIM. This gap widens at 5° (9.33 vs 20.86). Across models, FluenceFormer variants remain below $\sim 12\%$ error, while baselines exceed ~ 18 – 20% , indicating better physical consistency despite comparable SSIM.

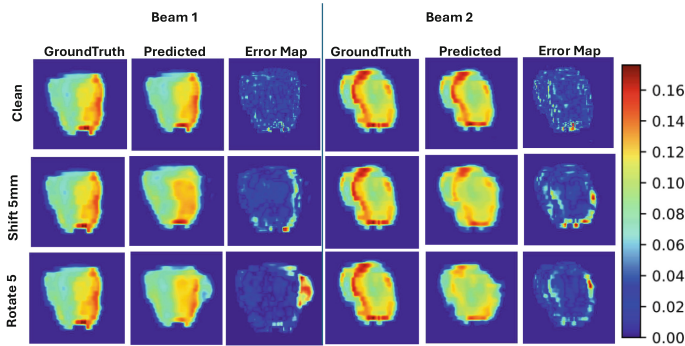


Fig. 1. Qualitative geometric robustness. Ground-truth fluence maps (repeated for visual reference) remain invariant, as geometric perturbations are applied only at inference.

3.2 Radiometric Robustness

All models degrade smoothly as noise increases. SwinUNETR drops from 0.74 to 0.65 SSIM ($\sim 12\%$), UNETR from 0.64 to 0.55 ($\sim 14\%$), and MedFormer from 0.62 to 0.53 ($\sim 15\%$), while nnFormer shows the largest degradation, from 0.55 to 0.44 ($\sim 20\%$). Overall, hierarchical models (SwinUNETR, MedFormer) retain higher performance under noise, whereas nnFormer is more sensitive, indicating reduced robustness to radiometric perturbations.

3.3 Domain and Data Robustness

Under domain shifts, SSIM changes minimally ($\sim 2\text{--}4\%$), while energy error increases by up to $\sim 30\%$, indicating physically inconsistent predictions. SwinUNETR shows smaller error growth ($7.33\rightarrow 9.65$) than CNN baselines ($>14\%$).

With reduced data, SwinUNETR drops $\sim 14\%$ SSIM ($0.76\rightarrow 0.65$), while nnFormer degrades most ($\sim 31\%$) (Table 1). Overall, hierarchical models retain better performance, indicating improved robustness and data efficiency.

Table 1. Data efficiency and scaling. SSIM and upper-quartile energy error $E_{\text{err}}^{Q_{75}}$ as a function of training data fraction.

Backbone	Metric	25%	50%	75%	100%
SwinUNETR	SSIM $\uparrow$	0.65 ± 0.09	0.71 ± 0.08	0.74 ± 0.07	0.76 ± 0.08
	$E_{\text{err}}^{Q_{75}} \downarrow$	10.56	7.32	5.43	4.67
UNETR	SSIM $\uparrow$	0.58 ± 0.09	0.62 ± 0.08	0.65 ± 0.07	0.67 ± 0.07
	$E_{\text{err}}^{Q_{75}} \downarrow$	12.56	10.89	10.56	8.67
nnFormer	SSIM $\uparrow$	0.40 ± 0.09	0.52 ± 0.08	0.55 ± 0.07	0.58 ± 0.06
	$E_{\text{err}}^{Q_{75}} \downarrow$	12.78	11.89	11.56	10.34
MedFormer	SSIM $\uparrow$	0.52 ± 0.08	0.58 ± 0.08	0.62 ± 0.07	0.65 ± 0.06
	$E_{\text{err}}^{Q_{75}} \downarrow$	12.26	11.92	11.45	10.67

3.4 External Robustness of Dose Prediction on Public Datasets

We evaluate cross-dataset robustness by applying the prostate-trained dose model to OpenKBP and CORT. As shown in Fig. 2, DVHs degrade smoothly

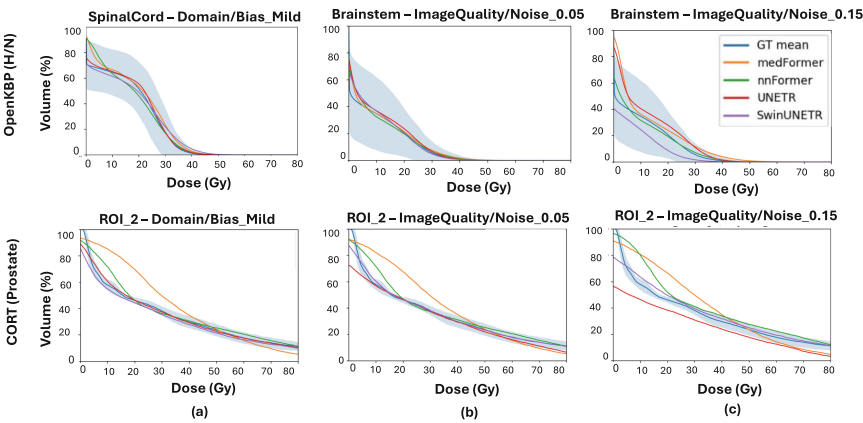


Fig. 2. DVH robustness on public datasets. DVH curves on OpenKBP (top) and CORT (bottom) under bias and noise. Shaded: ground truth mean $\pm$ std; lines: model predictions.

under bias and noise. SwinUNETR and UNETR remain closest to ground truth, while nnFormer and MedFormer deviate more, indicating generalization under moderate domain shift. End-to-end fluence robustness cannot be assessed without fluence labels.

4 Discussion and Conclusion

We evaluate the robustness of transformer-based fluence prediction under clinically motivated perturbations reflecting deployment-time uncertainty. Performance degrades under moderate stress and deteriorates under severe rotations and noise (e.g., SwinUNETR energy error increases from $\sim 7\%$ to $\sim 9\%$, while CNN baselines exceed $\sim 20\%$).

Hierarchical transformers show improved robustness, while SSIM alone fails to capture clinically relevant errors. Energy-based metrics reveal failure modes that are visually plausible but physically inconsistent.

These findings emphasize that robustness evaluation, rather than architectural novelty alone, is critical for assessing clinical reliability. Overall, robustness auditing and physics-informed evaluation are essential for safe and reliable deployment.

References

1. Hu, C., et al.: Trdosepred: a deep learning dose prediction algorithm based on transformers for head and neck cancer radiotherapy. *J. Appl. Clin. Med. Phys.* (2023)
2. Mgbob, U., Sultan, R., Zhu, D., Kim, J.: Fluence map prediction with deep learning: a transformer-based approach. *arXiv preprint [arXiv:2511.08645](https://arxiv.org/abs/2511.08645)* (2025)
3. Lee, H., Sheng, K.: Deep learning of fluence map patterns for deliverable IMRT planning. *Med. Phys.* (2019)
4. Glocker, B., et al.: Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects. *arXiv preprint [arXiv:1910.04597](https://arxiv.org/abs/1910.04597)* (2019)
5. Zech, J.R., et al.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* (2018)
6. Rogowski, V.: Impact of deep learning model uncertainty on manual corrections to auto-segmentation in prostate cancer radiotherapy. *arXiv preprint [arXiv:2502.18973](https://arxiv.org/abs/2502.18973)* (2025)
7. Elith, C., et al.: An introduction to the intensity-modulated radiation therapy (IMRT) techniques, tomotherapy, and VMAT. *J. Med. Imaging Radiation Sci.* (2011)
8. Mgbob, U., et al.: Fluenceformer: transformer-driven multi-beam fluence map regression for radiotherapy planning. *arXiv preprint [arXiv:2512.22425](https://arxiv.org/abs/2512.22425)* (2025)
9. Babier, A., et al.: OpenKBP: the open-access knowledge-based planning grand challenge and dataset. *Med. Phys.* (2021)
10. Craft, D., et al.: Shared data for intensity modulated radiation therapy (IMRT) optimization research: the cort dataset. *GigaScience* (2014)
11. Hansen, C.R.: Plan quality in radiotherapy treatment planning-review of the factors and challenges. *J. Med. Imaging Radiat. Oncol.* **66**(2), 267–278 (2022)

12. Kim, S.: Evaluation of radiotherapy plan robustness: a systematic literature review. *Clin. Oncol.* **48** (2025)
13. Wang, W., et al.: Deep learning-based fluence map prediction for pancreas stereotactic body radiation therapy with simultaneous integrated boost. *Adv. Radiation Oncol.* (2021)



AI-Driven Literature Term Extraction for Optimizing Laboratory Data Capture and Prioritization

Aída Muñoz Monjas^{1,2} , David Rubio Ruiz^{1,2} , David Perez-Rey² ,
Kinwei Arnold Chan¹ , and Matvey B. Palchuk¹ 

¹ TriNetX, LLC, Cambridge, MA, USA

david.rubio.ruiz@alumnos.upm.es,

{karnold.chan,matvey.palchuk}@trinetx.com

² Biomedical Informatics Group, Universidad Politécnica de Madrid, Madrid, Spain

aida.munozm@alumnos.upm.es, david.perez.rey@upm.es

Abstract. Harmonization of clinical laboratory data is essential to scalable secondary use. The focus of this standardization is typically set on the most commonly ordered labs, which account for the largest volume of clinical facts, but this approach underestimates the clinical relevance of less frequent lab tests, crucial in state-of-the-art research. This work explores the quantification of term clinical relevance via representativity in peer-reviewed publications that could be used as a harmonization prioritization indicator. We tested different AI-ML techniques, evaluating the extraction of laboratory components from published abstracts and titles: exact matching, fuzzy matching, Named Entity Recognition (NER), logistic regression and BioMistral. Using UMLS synonyms as data augmentation, combined approaches were trained and tested against an expert-annotated document corpus. NER with logistic regression proved to obtain the most comprehensive and accurate results. This technique combination was then used to extract lab terms from a targeted PubMed dataset, consisting of 2118 abstracts with expected laboratory tests. The obtained terms depict trends observed in literature, whose relevance could complement existing standardization approaches to support groundbreaking research.

Keywords: Standardization · Laboratory Data · Artificial Intelligence · Secondary Use · Knowledge Extraction

1 Introduction

Laboratory results captured within electronic health records are a highly valuable resource for a wide range of secondary applications of clinical data [6], yet this data domain is characterized by significant complexity and heterogeneity, which poses a challenge when standardizing across institutions [3, 4].

The standardization of laboratory tests is a non-trivial task, with precise results requiring of expert input in state-of-the-art automated systems [4].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027

P. Andreev et al. (Eds.): AIME 2026, LNAI 16749, pp. 341–345, 2027.

https://doi.org/10.1007/978-3-032-30813-9_63

Among the different approaches to this task, the historically preferred approach is term prioritization via volume, where laboratory terms are prioritized for standardization according to the number of clinical observations they represent [3]. Despite this approach covering the standardization of the most frequently performed lab tests, it fails to acknowledge the publication trends employing less common tests such as NT-proBNP or Lp(a). These less frequent terms are key in both retrospective peer-reviewed studies and clinical trial acceptance criteria, and evaluating their usage could complement historical standardization efforts impacting data capture [6, 8].

This publication explores laboratory test clinical significance prioritization via term representativity extraction from peer-reviewed publications, testing different Artificial Intelligence (AI) techniques for its automated retrieval.

2 Methodology

With the aim of identifying laboratory tests from peer-reviewed publications, different AI model approaches were tested and evaluated in an expert-annotated corpus. The best performing approach was then applied to a larger dataset.

The target laboratory terms for this task were extracted from LOINC component axis. Components representing single codes were removed from the analysis, excluding surveys and questionnaires, as well as terms related to documents or narrative reports. To expand the target label set, synonyms were extracted from UMLS MRCONSO file for the LOINC components, enriching the target classes with acronyms and alternative labels. Additionally, training examples were generated to describe the context in which laboratory terms appear in publications.

Multiple NLP/AI techniques were explored, each aimed at a different section of the described task.

- Embedding generation. To obtain a vector space representation, embeddings were generated using *BioMistral-7B* pretrained tokenizer and model. This suite of Mistral-based models are pre-trained using text data from PubMed [2], fitting the scope of the defined task.
- Entity recognition. *Spacy* large model was used for the task of Named Entity Recognition (NER), extended to detect the custom entity type ‘‘CLINICAL_MEASUREMENT’’.
- Classification. Weighted logistic regression was employed due to the expected heavy imbalance between target classes.
- Semantic matching. Both exact match and fuzzy matching were evaluated, with expected different resulting performances.

The described techniques were combined according to literature reported results [4]. To evaluate performance, precision and recall were studied over an expert-annotated document corpus, aiming to achieve a balance between them.

1. As a baseline, exact match was applied over the texts.

2. The second approach consisted of a NER system with a fuzzy matcher, evaluating term similarity via normalized Levenshtein distance [9]. Fuzzy matcher target terms were enriched with the obtained UMLS synonyms.
3. The third approach transformed input documents into vectors by generating BioMistral embeddings. Weighted logistic regression was trained using a UMLS-extended synonym list to classify vectors into LOINC components.
4. The last approach combined the four techniques described above, applying BioMistral embeddings over NER entities, then categorized by a weighted logistic regression and filtered using fuzzy matching scores.

The best performing approach was applied over a filtered subset of PubMed titles and abstracts to obtain a representative list of clinically significant lab terms from literature.

3 Results and Discussion

The few-shot NER for entity recognition in documents was trained using 200 laboratory related sentences [1], while the described classifier used 73 876 UMLS synonyms of the target LOINC codes for training. No additional information was necessary for the BioMistral embeddings.

Each of the described approaches were evaluated against a corpus of 450 annotated entities from PubMed publications for the identification of 2730 distinct LOINC components. The performance of the four combined approaches described in the methodology are shown in Table 1. As baseline, a simple exact match approach was used, searching for the exact label of LOINC components in the text. The precision obtained was relatively high, diminished by an overabundance of false positives due to a lack of context processing. The overall high precision highlights the strong adherence to standard labels in clinical literature.

Table 1. Performance comparison between the studied approaches.

Approach	Precision	Recall	Top 5 recall	F1 score
Exact match	48.18	22.75	-	30.91
NER + fuzzy match	21.62	31.90	39.62	25.77
BioMistral embeddings + logistic regression	10.18	4.47	-	6.21
BioMistral embeddings over NER entities + weighted logistic regression + fuzzy match	35.43	37.28	44.60	36.33

The second approach obtained a high entity recall during the NER steps (83.4%), significantly reducing the search space. When transforming the highlighted entities into the target component list, the fuzzy match used led to incorrect classifications, as acronyms and slight differences between components went undetected.

The third approach, logistic regression with BioMistral embeddings, especially trained over PubMed data for clinical text-directed tasks, displayed sub-optimal performance in terms of precision and recall. This behaviour is explained by the nature of generalized LLM models, that despite being trained for a specific data domain underperform in specific, precision-requiring tasks, as reported in literature [7].

The last approach aims to combine the strengths of the selected methods, generating BioMistral embeddings from NER-detected entities, which are later classified by a weighted logistic regression. A fuzzy matching post-processing step was applied to reduce the ratio of false positives. The classification into LOINC components captured the nuances missing in previous approaches, correctly handling small class differences, such as *urobilin* and *urobilinogen*, as well as acronyms and context-aware tagging.

The best overall performance for the defined task was obtained from the fourth approach, which aligns with literature observed findings when applying NLP techniques over clinical real-world texts [4]. By combining the strong points of the different techniques as a sequence of distinct tasks - the reduced search space of NER, clear nuance distinctions from the logistic regression using context-aware embeddings from specialist models and the cutoff specificity of the fuzzy match - a hybrid model was obtained with a more balanced result between precision and recall in both top-1 and top-5 metrics.

This best performing approach was applied over a random subset of 2118 PubMed publications containing common lab specimen information, to ensure the existence of clinical laboratory tests in the sample. The exact query to extract said documents was (*serum*[Title/Abstract]) OR (*plasma*[Title/Abstract]) OR (*blood*[Title/Abstract]) OR (*urine*[Title/Abstract]).

The top two, from the 11 172 identified laboratory components, correspond to the terms **blood** and **age**, in 24% and 9% of documents respectively. The appearance of these terms matches expectations, as age is a frequent cohort descriptor and blood was a term directly included in the extraction query. These terms are followed by components that belong to the metabolic panel and complete blood count, corresponding to routinely tested assays.

Additionally, data elements with less testing frequency were also present in the top 100 terms. For instance, *lipoprotein (little a)* appears in 1% of documents, reflecting its increasing relevance in the last years as a cardiovascular risk indicator. Similarly, *fibrinogen* was detected in 0.7% of documents.

These obtained terms should be prioritized during laboratory term standardization, to improve data capture and availability in the dataset.

Known limitations of this work are not only confined to the inherent limitations of the applied techniques but also the variability of the studied data and the nature of the defined task. The combination of a high number of target categories and a relatively small amount of training data may lead to suboptimal classification performance, and result interpretation should be approached with caution, taking into account both the complexity of the data and the limitations of the applied technique.

4 Conclusion

Evaluating trends in peer-reviewed publications is a necessary step to contextualize and guide semantic standardization in collaborative research networks such as TriNetX [5]. The capture of these valuable clinical terms will result in an increased number of data points available, ensuring network representativity and bias reduction.

This publication explores the extraction of these relevant laboratory terms from peer-reviewed publications, leveraging LOINC component axis and synonyms. Four different AI techniques were tested, combined and evaluated, obtaining the sequential aggregation of these techniques the best performance.

Building on the findings of this study, other AI approaches will be explored to solve the task and improve the reported performance. Additional expert review will be introduced to further validate the granularity of the LOINC component selection.

References

1. Goel, M., et al.: Deep learning based named entity recognition models for recipes. *ACL Anthol.*, 4542–4554 (2024)
2. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M., Dufour, R.: Biomistral: a collection of open-source pretrained large language models for medical domains (2024)
3. McDonald, C.J., et al.: Mis-mappings between a producer’s quantitative test codes and LOINC codes and an algorithm for correcting them. *J. Am. Med. Inform. Assoc.* **30**(2), 301–307 (2023). <https://doi.org/10.1093/jamia/ocac215>
4. Naliyattthaliyazhayil, P., Sangam, V.R., Amlung, J., Kanter, A.S., Purkayastha, S., Payne, J.: Automated logical observation identifiers names and codes mapping with biomedical natural language processing models: enabling scalable health information exchange via the Open Concept Lab. *J. Am. Med. Inform. Assoc.*, ocag010 (2026). <https://doi.org/10.1093/jamia/ocag010>
5. Palchuk, M.B., et al.: A global federated real-world data and analytics platform for research. *JAMIA Open* **6**(2), ooad035 (2023). <https://doi.org/10.1093/jamiaopen/oad035>
6. Rubinstein, M., et al.: Effectiveness of practices to support appropriate laboratory test utilization: a laboratory medicine best practices systematic review and meta-analysis. *Am. J. Clin. Pathol.* **149**(3), 197 (2018). <https://doi.org/10.1093/ajcp/aqx147>
7. Soroush, A., Glicksberg, B.S., Zimlichman, E., Barash, Y., Freeman, R., Charney, A.W., Nadkarni, G.N., Klang, E.: Large language models are poor medical coders — benchmarking of medical code querying. *NEJM AI* (2024). <https://doi.org/10.1056/AIdbp2300040>
8. Xu, K., Bai, X., Sun, Y., Zhao, H., Zhou, W., Zhang, C.: The invisible bridge: how harmonization closes the gap between impossible standardization and clinical necessity. *Clin. Chim. Acta* **578**, 120516 (2026). <https://doi.org/10.1016/j.cca.2025.120516>
9. Yujian, L., Bo, L.: A normalized Levenshtein distance metric. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 1091–1095 (2007). <https://doi.org/10.1109/TPAMI.2007.1078>



A Heuristic for Spatial Rearrangement of Large Cytology Images for Efficient Classification

Rafael Silva Godoy^{1,2(✉)}, Carlos Henrique Costa Ribeiro¹ ,
and Israel Tojal da Silva³

¹ Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, Brazil
{godoyrsg, carlos}@ita.br

² Universidade Federal de São Paulo (UNIFESP), São Paulo, Brazil

³ A. C. Camargo Cancer Center, São Paulo, Brazil
itojal@accamargo.org.br

Abstract. Ultra-high-resolution cytology images provide rich morphological information but pose significant challenges for scalable machine learning due to their size and computational cost. This work presents a preliminary study on an image preprocessing strategy based on spatial rearrangement. The proposed approach reorganizes detected cellular elements into a compact representation under the assumption that global spatial arrangement is weakly informative in cytology.

Importantly, the method does not apply geometric transformations, such as scaling or rotation, to individual elements, preserving their original spatial scale and morphology. Information loss may occur due to segmentation and filtering steps, but not due to resolution reduction. Preliminary experiments using the Clustering-constrained Attention Multiple Instance Learning (CLAM) framework indicate that the method can substantially reduce memory and computational requirements while maintaining comparable classification performance. The results suggest that spatial rearrangement represents a promising direction for efficient learning in large-scale cytology images.

Keywords: Cytology image · Computer Vision · Spatial Rearrangement

1 Introduction

Recent advances in medical imaging have enabled the acquisition of ultra-high-resolution cytology images, providing detailed morphological information at the cellular level. However, these large-scale images, often with tens of thousands of pixels per dimension, impose significant memory and computational challenges for machine learning pipelines [1].

Cytological diagnosis relies primarily on individual cellular morphology rather than global spatial organization [2]. Due to scraping-based collection procedures, cells are detached from their original tissue context and distributed

across the slide in a largely stochastic manner, reducing the diagnostic relevance of spatial arrangement. This contrasts with histopathology, where tissue architecture plays a central role [3].

Motivated by this characteristic, this work investigates a spatial rearrangement preprocessing heuristic that compacts large cytology images by reorganizing detected cellular elements into smaller representations. The objective is to reduce computational cost while maintaining classification performance. Experiments are conducted using the CLAM framework [4].

Unlike downsampling or patch-based sampling, the proposed approach does not reduce image resolution or apply geometric transformations such as scaling or rotation. Instead, elements are repositioned while preserving their original scale. Although some information may be removed during segmentation and filtering, the morphology of retained elements remains unchanged. To the best of our knowledge, this strategy has not been previously explored as a preprocessing approach in cytology.

2 Method Overview

The proposed preprocessing approach compacts large cytology images by spatially rearranging detected cellular elements, leveraging the limited importance of global spatial organization in cytology [2, 3]. The pipeline consists of four stages: background removal, element segmentation, bounding box demarcation [5], and spatial rearrangement.

Background pixels are first separated from foreground elements, followed by connected-component labeling to identify individual elements. For each element, an axis-aligned bounding box is computed and subsequently repositioned using a heuristic two-dimensional packing strategy that maximizes spatial utilization while avoiding overlap.

The compacted image is then reconstructed by placing each element at its new location using the original pixel values. Only translational transformations are applied, preserving the internal structure and spatial scale of retained elements. However, segmentation and filtering steps may exclude portions of the image, potentially introducing information loss depending on parameter choices.

3 Method Details and Results

3.1 The Dataset

The proposed spatial rearrangement preprocessing strategy was evaluated on a dataset of 873 high-resolution cytology images acquired using the pap smear method. Images were annotated by expert pathologists looking for tumors, with 39.3% labeled as positive (HSIL, LSIL and adenocarcinoma) and 60.7% as negative. All experiments were conducted using the CLAM framework.

3.2 Heuristic Workflow for Cytology Image Compaction

The experimental evaluation employed a four-stage spatial rearrangement workflow consistent with the proposed heuristic. For background removal, a global thresholding strategy was adopted due to the strong contrast between the white background and cellular structures in the images. The threshold was automatically determined using Otsu’s method [7], which selects an optimal threshold by maximizing the between-class variance of the image intensity histogram. Element segmentation was performed using connected component labeling with a 4-connected adjacency system. This step assigns a unique label to each connected foreground component, enabling independent manipulation of individual elements [8]. Bounding box demarcation was performed by computing axis-aligned rectangular bounding boxes for each labeled element. Two filtering criteria were applied at this stage: a minimum element size threshold, used to discard small components and reduce noise, and a maximum element size threshold, used to subdivide oversized elements. These filters significantly impact both computational efficiency and packing quality, as the number and size distribution of elements directly affect the packing process. Spatial rearrangement was implemented using a greedy two-dimensional packing strategy based on the First Fit Decreasing (FFD) heuristic [9]. Elements were sorted according to bounding box area and sequentially placed into the target image grid while preventing overlap. Since two-dimensional packing is NP-hard [6], the adopted heuristic prioritizes computational efficiency over optimality. Finally, the compacted image was reconstructed by replacing each optimized bounding box position with its corresponding original pixel values. As only translational transformations were applied, the internal structure of each element was preserved, allowing elements to be traced back to their original locations if required.

3.3 Results

Figure 1 illustrates an example of the proposed spatial rearrangement applied to a $10k \times 10k$ downsampled slide, resulting in a compacted $3k \times 3k$ representation. This example is intended for visualization purposes, highlighting how cellular elements are reorganized while preserving their morphology.

The heuristic was evaluated using a minimum element size of 250 pixels and a maximum threshold of 450 pixels before element splitting, with a packing strategy prioritizing larger elements. On average, preprocessing required approximately 90 s per slide, and the method is fully parallelizable. In our experiments, storage requirements were reduced from approximately 15 GB to 3 GB per slide (5× reduction). The proportion of retained content varied across slides, with an interquartile range between $Q1 = 19\%$ and $Q3 = 79\%$ under the selected hyperparameters.

Model training time was also significantly reduced. Under identical conditions using the CLAM framework with 400 training epochs, the average training time decreased from 18 h and 46 min to 6 h and 43 min per fold, corresponding to a speed-up exceeding 2.5×.

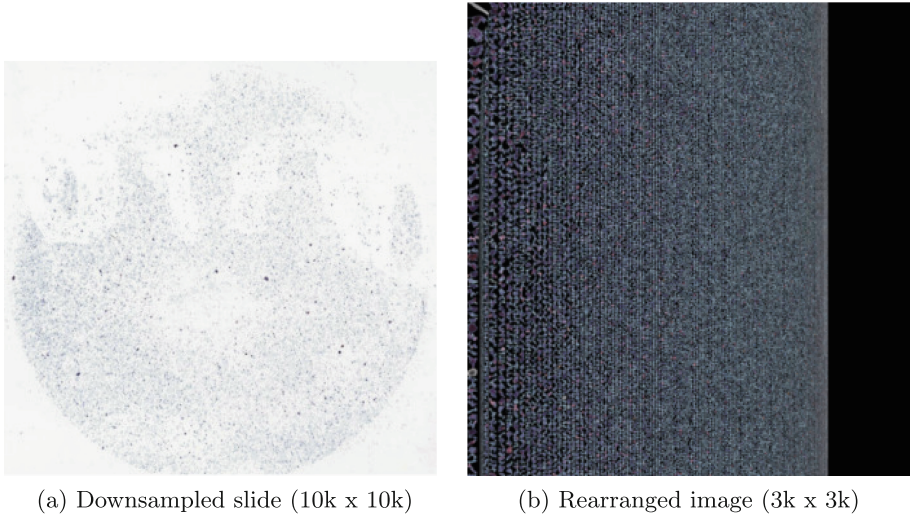


Fig. 1. Downsampled and rearranged cytology images.

To evaluate the impact of spatial rearrangement, three configurations were considered: (i) rearranged images obtained from 10k downsampled slides and compacted to 3k, (ii) a downsampled baseline (100k $\rightarrow$ 10k), and (iii) full-resolution images (100k) without preprocessing. As shown in Table 1, the rearranged configuration achieves performance comparable to the downsampled baseline, while requiring substantially lower computational cost.

These results highlight the trade-off between computational efficiency and predictive performance, suggesting that spatial rearrangement provides a favorable balance for large-scale cytology analysis.

Finally, the proposed heuristic is still under active development. Ongoing efforts focus on improving its scalability and robustness, with the goal of enabling direct processing of full-resolution 100k images without prior downsampling.

Table 1. Comparative CLAM performance (mean $\pm$ std).

Configuration	Initial Size	Final Size	AUROC	Acc.	F1	Recall
Rearrangement	10k	3k	0.699 ± 0.035	0.721 ± 0.032	0.621 ± 0.058	0.600 ± 0.133
Downsample	100k	10k	0.711 ± 0.045	0.732 ± 0.042	0.639 ± 0.070	0.620 ± 0.122
Original	100k	-	0.845 ± 0.047	0.765 ± 0.046	0.772 ± 0.051	0.765 ± 0.046

4 Discussion

This study investigates a spatial rearrangement strategy tailored to cytology, where diagnosis relies primarily on individual cellular morphology rather than

global spatial organization. The method preserves the spatial scale and morphology of retained elements, while potential information loss may arise from segmentation and filtering steps. Our results indicate that spatial rearrangement can substantially reduce memory requirements and training time while maintaining comparable predictive performance. These findings suggest that the approach can improve computational efficiency without directly degrading the morphological information used for classification.

However, altering the original spatial context may impact interpretability for human observers and explainable AI methods, particularly in clinical settings. As this work represents an initial exploratory study, these effects have not yet been fully characterized. In particular, the extent to which discarded regions influence different cytological patterns remains an open question and will be investigated in future work.

Overall, spatial rearrangement should be viewed as a complementary preprocessing strategy that enables more efficient use of high-resolution data. In this context, it may facilitate the training or fine-tuning of large vision and multimodal models on ultra-high-resolution cytology images.

Acknowledgments. I.T.S was supported by FAPESP (#22/12991-0) and CNPq (#33/2024).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Declaration of AI Assistance. The authors used ChatGPT-5.2 exclusively for language translation. No scientific content or results were generated by the AI. The authors take full responsibility for the manuscript.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks, pp. 1097–1105. NeurIPS (2012)
2. Koss, L.G., Melamed, M.R.: Koss' Diagnostic Cytology and its Histopathologic Bases. 5th edn. Lippincott Williams & Wilkins (2006)
3. Bibbo, M., Wilbur, D.C.: Comprehensive Cytopathology, 4th edn. Elsevier (2014)
4. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**(6), 555–570 (2021)
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation, pp. 580–587. CVPR (2014)
6. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman and Company, New York (1979)
7. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979)
8. Weaver, J.B.: Centrosymmetric packing of equal circles. *J. Math. Phys.* **26**(2), 165–170 (1985)
9. Baker, B.S., Coffman, E.G., Jr., Rivest, R.L.: Orthogonal packings in two dimensions. *SIAM J. Comput.* **9**(4), 846–855 (1980)



OpenCDT: An Open Neuro-Symbolic Infrastructure for Executable Clinical Decision Tools

Antonina Spalińska¹(✉) , Krzysztof Spaliński² , Michał Chojnicki³ ,
Magdalena Jasionowska-Skop⁴ , Artur Przelaskowski² ,
and Wojciech Glinkowski⁴

¹ Cardinal Stefan Wyszyński University, Warsaw, Poland
a.slubowska@uksw.edu.pl

² Warsaw University of Technology, Warsaw, Poland

³ Poznań University of Medical Sciences, Poznań, Poland

⁴ Medical University of Warsaw, Warsaw, Poland

Abstract. Clinical scoring systems and decision algorithms guide diagnostic and therapeutic workflows but are typically distributed as narrative publications or embedded in proprietary software without standardized, machine-readable representations. This fragmentation limits interoperability, reproducibility, and safe integration into large language model (LLM)-based clinical systems. We present OpenCDT, an open-source neuro-symbolic infrastructure for structured representation and deterministic execution of clinical decision tools, implemented as a Model Context Protocol (MCP)-compatible server.

Keywords: Large language models (LLMs) · Neuro-symbolic AI · Clinical decision support systems (CDSS) · Model Context Protocol (MCP) · Auditable and uncertainty-aware reasoning

1 Introduction

Clinical decision rules and scoring systems (e.g., CHA₂DS₂-VASc, HEART, NEWS2) translate evidence into structured risk frameworks that standardize interpretation and guide point-of-care decisions [1]. Despite their clinical importance, most are distributed as narrative publications, static tables, or proprietary calculators without machine-readable representations. As large language models (LLMs) enter clinical workflows, a key safety challenge arises: probabilistic models are poorly suited to deterministic rule execution and may hallucinate variables, misapply criteria, or produce non-reproducible outputs [2].

Recent methodological and regulatory efforts emphasize transparent reporting, validation, and post-deployment monitoring of clinical AI. STARD-AI extends diagnostic accuracy standards to AI-based tools, underscoring reproducibility and structured evaluation [3]; policy and governance frameworks

further highlight auditability, safety, and data sovereignty [4, 5]. Computer-interpretable guidelines (CIGs) have a long history of formalizing clinical knowledge for computational use [10, 11], producing modeling languages, execution engines, and EHR-integrated decision support systems [11]. Operational tools target specific consumers: CDS Hooks delivers recommendations to EHR workflows [6], openEHR GDL2 provides a repository of formalized rules [7], and proprietary calculators such as MDCalc serve clinicians directly [8]. None, however, exposes a standardized interface through which LLMs can discover, retrieve, and deterministically execute validated decision tools.

The Model Context Protocol (MCP) [9] provides a standardized interface for exposing structured tools to LLMs. Models can retrieve authoritative definitions and perform deterministic execution, separating probabilistic language understanding from symbolic rule evaluation.

We present **OpenCDT** (Open Clinical Decision Tools), an open-source neuro-symbolic infrastructure for structured representation and deterministic execution of clinical decision tools. Tools are encoded as validated JSON specifications, while LLMs extract clinical information and map it to structured variables, with scoring executed server-side under deterministic semantics. The system computes exact score ranges under incomplete information by enumerating constraint-consistent assignments, enabling explicit uncertainty representation and ranking of missing variables by decision relevance. Trustworthiness is ensured through deterministic evaluation, schema validation, auditability, and full offline deployability. This separation of probabilistic extraction from symbolic execution supports reproducible, interoperable, and safety-aware clinical decision support.

We demonstrate the system using a fully local MCP-enabled LLM setup and evaluate it on a 50-case stress-test set. It was deployed on a Ubuntu workstation using a `llama.cpp`-served GPT-OSS 120B model with native tool calling, interfacing with the OpenCDT MCP server via an OpenAI-compatible API, ensuring fully offline, deterministic tool execution without external dependencies. Code is available on GitHub at: <https://github.com/aspalinka/OpenCDT>

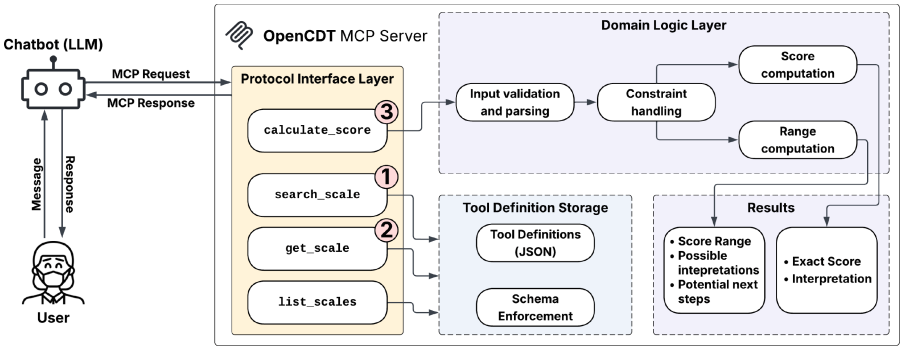


Fig. 1. Architecture of the OpenCDT MCP server. Numbers represent typical order of functions called by LLM.

2 Tool Description

OpenCDT is implemented as a MCP server, enabling any MCP-compatible LLM client to discover and execute clinical decision tools without custom integration code. The architecture separates concerns into three layers: a *tool definition schema*, a *domain logic layer*, and a *protocol interface layer*. The architecture of the tool is illustrated in Fig. 1. The interaction workflow between the LLM and the MCP server is illustrated in Fig. 2.

Tool Definition Schema. Each clinical decision tool is encoded as a self-contained JSON file comprising: (i) metadata (name, purpose, applicability criteria, clinical category, and literature references); (ii) a list of *items*, a clinical description, and an explicit mapping from categorical choices to numeric point values; (iii) a mathematical formula referencing item identifiers, evaluated deterministically; (iv) interpretation rules mapping score ranges to clinical recommendations; and (v) optional *mutual-exclusivity constraints* declaring groups of items where at most one may be non-zero (e.g., age-bracket variables).

Adding a new tool to the repository requires only placing a conforming JSON file in the designated directory; the server discovers and validates it at startup. Tool creation is partially assisted by LLMs, which can generate initial JSON specifications from source publications; these are subsequently manually reviewed and validated.

The current implementation focuses on clinical decision tools with discrete, additive scoring logic, in which individual items contribute fixed point values. All tools are encoded from primary literature sources or authoritative clinical references describing the original scoring logic. The full list of implemented tools and their corresponding specifications is available in the project repository.

Domain Logic Layer. The core contribution of the domain layer is its handling of incomplete information. The LLM-facing interface accepts a structured input for every item in a tool, where each entry carries a value (or `null` if unknown) and a natural-language reasoning field that is required whenever a value is assigned, which documents the clinical evidence behind the assignment. When all values are provided, the system returns an exact score with its interpretation. When some values are missing, the system computes deterministic score bounds by:

1. **Constraint-aware auto-fill:** if a non-zero value is supplied for one member of a mutually exclusive group, all other members are automatically set to zero.
2. **Full enumeration:** the system enumerates all combinations of allowed point values across every remaining missing item, discards any assignment that violates a mutual exclusivity constraint, and records the minimum and maximum score across all feasible assignments. This is correct for any formula, including non-linear ones, as no assumption about per-item contribution direction is made.

The range result lists each missing item with its description, and allowed values, alongside all interpretation categories overlapping the current bounds. This enables the LLM to prioritize follow-up questions targeting the items that most reduce diagnostic uncertainty.

Protocol Interface Layer. The MCP server exposes four tools that define a guided workflow: `search_scales` finds relevant tools by keyword and tag matching; `list_scales` enumerates all available tools; `get_scale` retrieves the full definition including exact item identifiers and allowed values; `calculate_score` accepts structured inputs and returns either an exact or range result. The sequential workflow (`search` → `get` → `calculate`) ensures that the LLM operates on authoritative tool definitions and reduces the risk of hallucinations. All tool responses are returned as JSON, including structured error messages, facilitating reliable parsing and downstream use.

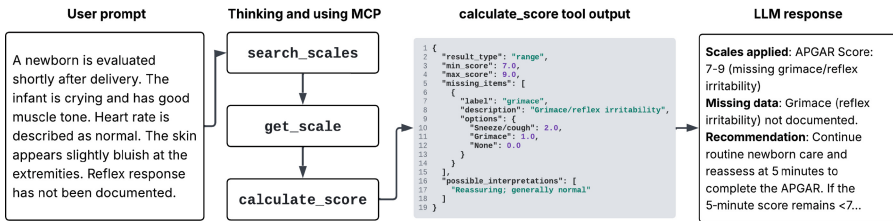


Fig. 2. Typical interaction workflow between an LLM and the OpenCDT MCP server.

Design Principles. Several design choices support trustworthiness in clinical settings: (i) *deterministic execution* – scoring logic uses `numexpr` Python package with no stochastic components, guaranteeing reproducibility; (ii) *auditability* – every item assignment requires documented reasoning, and auto-filled values are explicitly flagged in the output; (iii) *deployment flexibility* – as a local MCP server, the system can operate fully offline with locally hosted LLMs, satisfying data-sovereignty requirements in hospital environments.

3 Evaluation Framework and Results

Evaluation Setup. We conducted an evaluation of OpenCDT using a set of 50 synthetic clinical cases. The cases were grouped into six categories: (i) complete-information cases (all variables required for at least one scale available), (ii) missing-information cases (partial data requiring uncertainty handling), (iii) out-of-scope cases (no supported scale applicable), (iv) contradictory-information cases (conflicting clinical evidence), (v) distractor-noise cases (irrelevant comorbidities or findings), and (vi) short sparse notes requiring targeted

clarification. We compared two configurations: (1) an MCP-connected system, in which the LLM accessed OpenCDT tools for deterministic execution, and (2) the same underlying model operating without MCP access. Responses were evaluated using a structured rubric covering scale selection, mapping of clinical variables, score handling, recommendation correctness, next information suggested, and uncertainty handling. Each dimension was scored on a three-point scale (0 = incorrect, 1 = partially correct, 2 = correct).

Results. Across the evaluated cases, the MCP-connected system showed strong performance when a supported clinical scale was applicable. In complete-information scenarios, it achieved near-perfect correctness, reflecting reliable deterministic execution. In missing-information cases, it avoided hallucinating unavailable variables and represented uncertainty using score ranges derived from partial inputs, although these were occasionally overly wide due to conservative handling.

Compared to the same model without MCP access, the MCP-enabled configuration improved consistency in variable mapping and score calculation. The baseline model often produced clinically plausible responses but more frequently hallucinated scale definitions or scoring logic, leading to incorrect results. Differences were most pronounced in distractor-noise cases, where the MCP system maintained focus on relevant variables, while the baseline model was more susceptible to irrelevant information.

Limitations were most evident in short sparse notes, where the MCP system tended to apply scoring despite minimal available information, returning wide score ranges instead of deferring computation and requesting targeted inputs. Out-of-scope cases exposed challenges in boundary handling: although the system consistently identified when no appropriate scale was available, it sometimes additionally computed partially applicable general severity scores (e.g., NEWS2 or qSOFA), which were not clinically relevant to the task. In contradictory-information cases, uncertainty was consistently represented, but conflicting inputs were not always fully propagated into the computed score bounds.

Quantitative results are summarized in Table 1. A complete set of cases, model outputs, and evaluation annotations is available in the project repository.

Table 1. Quantitative summary of evaluation results by case category. Scores represent proportion of maximum achievable points.

Case category	OpenCDT (MCP)	Baseline (no MCP)
Full information	120/120 (100%)	112/120 (93.3%)
Missing information	114/120 (95.0%)	110/120 (91.7%)
Out of scope	12/20 (60.0%)	—
Contradictory	47/60 (78.3%)	46/60 (76.7%)
Distractor	57/60 (95.0%)	48/60 (80.0%)
Short notes	94/120 (78.3%)	102/120 (85.0%)

4 Discussion and Future Directions

This stress-test evaluation shows that OpenCDT improves correctness and reproducibility of clinical scoring by enforcing deterministic execution and reducing hallucinated variables. Observed limitations were largely driven by incomplete tool coverage and system prompt-dependent interaction behavior. A key next step is the development of a dedicated conversational mode, in which prioritization of the most decision-relevant variables will be essential for guiding efficient follow-up questions. Future work will also expand the tool set and extend the framework beyond discrete scoring systems to support continuous risk models and multi-step decision algorithms.

Declaration of AI Assistance Technologies During Manuscript Preparation. The authors used ChatGPT 5.2 for refinement, correction, editing and formatting the manuscript to improve clarity of language. After using this technology, the authors reviewed the results and take full responsibility for the contents of the manuscript.

Disclosure of interests. The authors declare no competing interests.

References

1. Steyerberg, E.W.: Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 2nd edn. Springer, Cham (2019). <https://doi.org/10.1007/978-3-030-16399-0>
2. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., et al.: Large language models encode clinical knowledge. *Nature* **620**(7972), 172–180 (2023). <https://doi.org/10.1038/s41586-023-06291-2>
3. Sounderajah, V., Guni, A., Liu, X., et al.: The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat. Med.* **31**, 3283–3289 (2025). <https://doi.org/10.1038/s41591-025-03953-8>
4. Ismail, M.A.: The strategic trajectory of artificial intelligence in Qatar’s healthcare sector: a model for UN sustainable development goal 9. *Front. Artif. Intell.* **9**, 1702242 (2026). <https://doi.org/10.3389/frai.2026.1702242>
5. U.S.: Department of health and human services. National AI strategy for health innovation, Washington, DC (2025)
6. HL7 international: CDS Hooks v2.0.1. Ann Arbor: HL7 international (2023). <https://cds-hooks.hl7.org/>
7. Medical informatics group, Cambio healthcare systems. The GDL project. gdl-lang.org (2026). <https://gdl-lang.org/the-project/>
8. MDCalc: MDCalc – medical calculators, equations, scores, and guidelines. New York: MDCalc (2026). <https://www.mdcalc.com>
9. Anthropic: Introducing the model context protocol. San Francisco: Anthropic (2026). <https://www.anthropic.com/news/model-context-protocol>
10. Teije, A., Lucas, P., Miksch, S. (eds.): Computer-Based Clinical Guidelines and Protocols: A Primer and Current Trends. IOS Press, Amsterdam (2008)
11. Peleg, M.: Computer-interpretable clinical guidelines: a methodological review. *J. Biomed. Inform.* **46**(4), 744–63 (2013). <https://doi.org/10.1016/j.jbi.2013.06.009>



Clinician-Supervised Large Language Models Generated Exercise Protocols for Degenerative Knee Disease: A Four-Arm Pilot Randomized Trial

Ömer Alperen Gürses¹(✉) , Nazife Kapan Tunçer² , and İsmail Özsoy³

¹ School of Physiotherapy and Rehabilitation, Kırşehir Ahi Evran University, Kırşehir, Türkiye
omer.alperengurses@gmail.com

² Faculty of Medicine, Kırşehir Ahi Evran University, Kırşehir, Türkiye

³ Faculty of Health Science, Selçuk University, Konya, Türkiye

Abstract. Large language models (LLMs) can generate standardized, evidence-informed rehabilitation protocols, yet their effect on patient outcomes remains uncertain. We conducted a single-centre, four-arm randomized pilot trial in which 52 adults with degenerative knee disease were allocated equally to conventional physiotherapy alone (Control; $n = 13$) or to conventional physiotherapy plus an LLM-generated exercise protocol (ChatGPT-5, Gemini 2.5 Pro, or DeepSeek V3.1; $n = 13$ each). Before recruitment, a single standardized prompt was issued to each LLM to generate an 8-week supervised progressive exercise protocol for a representative patient scenario using standard clinic equipment. After safety review, each protocol was applied unchanged to all participants in its corresponding group alongside identical co-interventions, delivered twice weekly for 8 weeks. The primary outcome was change in Knee Injury and Osteoarthritis Outcome Score (KOOS) total score (0–100) from baseline to week 8. KOOS improved in all groups, with mean changes of $+6.2 \pm 14.5$ in Control, $+14.0 \pm 8.9$ in ChatGPT-5, $+16.6 \pm 10.0$ in Gemini, and $+16.2 \pm 14.6$ in DeepSeek; between-group differences were not statistically significant ($p \approx 0.12$). No serious adverse events occurred. In this pilot trial, LLMs functioned as protocol generators under clinician supervision, producing feasible and safe exercise programs with numerically larger KOOS gains than clinician-designed usual care. Larger trials are needed to confirm comparative effectiveness.

Keywords: Large Language Models · Protocol Generation · Degenerative Knee Disease · Rehabilitation

1 Introduction

Degenerative knee disease (DKD), primarily knee osteoarthritis (OA), is a leading cause of global disability characterized by pain and functional limitation [1]. Major guidelines recommend 8-week supervised exercise therapy as a core first-line intervention [2–4]. However, wide variations in exercise selection, dosage, and progression rules

often lead to heterogeneous clinical outcomes and implementation challenges [2, 3, 5]. Large language models (LLMs) can generate structured exercise schemes, yet their application in orthopaedics reveals variable performance, clinical reasoning gaps, and a lack of patient-outcome trials, necessitating expert “human-in-the-loop” oversight [6–10]. Empirical evidence from randomized trials implementing LLM-generated protocols remains exceptionally scarce [6, 7]. This pilot study evaluates whether fixed, clinician-supervised 8-week exercise protocols generated by three different LLMs (ChatGPT-5, Gemini 2.5 Pro, DeepSeek V3.1) improve Knee Injury and Osteoarthritis Outcome Score (KOOS) outcomes compared to a conventional clinician-designed exercise programme in adults with DKD.

2 Methods

This single-centre, four-arm pilot trial (NCT07267962) enrolled 52 adults aged 40–65 years with degenerative knee disease and ≥ 3 months of pain. Participants were randomized 1:1:1:1 ($n = 13/\text{group}$) by concealed envelopes; assessors were blinded. No formal power calculation was performed [11]. All groups received identical physical agents twice weekly for 8 weeks (hot pack, TENS, ultrasound). The Control group followed a clinician-designed standard exercise programme. In the three experimental arms, ChatGPT-5, Gemini 2.5 Pro, and DeepSeek V3.1 were each used once as protocol generators (accessed Nov 12–14, 2025) with the following exact prompt: “Create an 8-week, supervised, progressive exercise protocol for a typical patient aged 40–65 years with degenerative knee disease, such as Kellgren–Lawrence grade 2–3 knee osteoarthritis. The protocol should be delivered in an outpatient physiotherapy clinic twice per week for 8 weeks. It should be feasible using standard clinic equipment, including treadmills, cycle ergometers, free weights, resistance bands, balance boards, treatment tables and step platforms. The program should focus on knee and hip strengthening, range of motion, proprioception, balance and functional exercises. Please provide week-by-week progression, including exercise names, sets, repetitions, intensity/progression criteria, rest intervals and safety precautions. Avoid high-impact, ballistic or unsafe exercises, excessive painful loading, and any recommendation requiring unsupervised home implementation, medication, injection or surgical decision-making. The protocol should be suitable for supervised implementation by physiotherapists.” Outputs were safety-screened for OARSI/ACR consistency and applied unchanged; temporary load reductions during symptom flares were allowed [2, 3, 12]. The primary outcome was change in KOOS total score from baseline to week 8 [13]. KOOS change was compared using one-way ANOVA ($\alpha = 0.05$, η^2); other outcomes were exploratory [11].

3 Results (Table 1)

All groups showed within-group improvements in KOOS total score from baseline to week 8 (Table 2). Mean KOOS change was +6.2 (SD 14.5) in the control group, +14.0 (SD 8.9) in the ChatGPT-5 protocol arm, +16.6 (SD 10.0) in the Gemini 2.5 Pro arm, and +16.2 (SD 14.6) in the DeepSeek V3.1 arm; group-wise KOOS changes are shown

Table 1. Baseline characteristics of the participants

Variable	Control (n = 13)	ChatGPT-5 (n = 13)	Gemini 2.5 Pro (n = 13)	DeepSeek V3.1 (n = 13)
Age (years)	50.3 ± 8.9	48.9 ± 11.5	51.8 ± 11.4	48.8 ± 15.2
Female (n(%))	9 (69.2)	6 (46.2)	10 (76.9)	11 (84.6)
Height (cm)	167.7 ± 11.2	169.1 ± 10.9	160.6 ± 9.5	163.5 ± 10.9
Weight (kg)	92.0 ± 17.5	92.2 ± 13.3	76.1 ± 8.0	84.8 ± 25.2
BMI (kg/m ²)	33.1 ± 8.0	32.6 ± 6.5	29.8 ± 4.8	31.2 ± 7.0
KOOS (total score)	34.2 ± 13.9	45.7 ± 18.4	44.0 ± 17.0	36.3 ± 16.3

Values are mean ± standard deviation unless otherwise indicated. BMI, body mass index; KOOS, Knee Injury and Osteoarthritis Outcome Score.

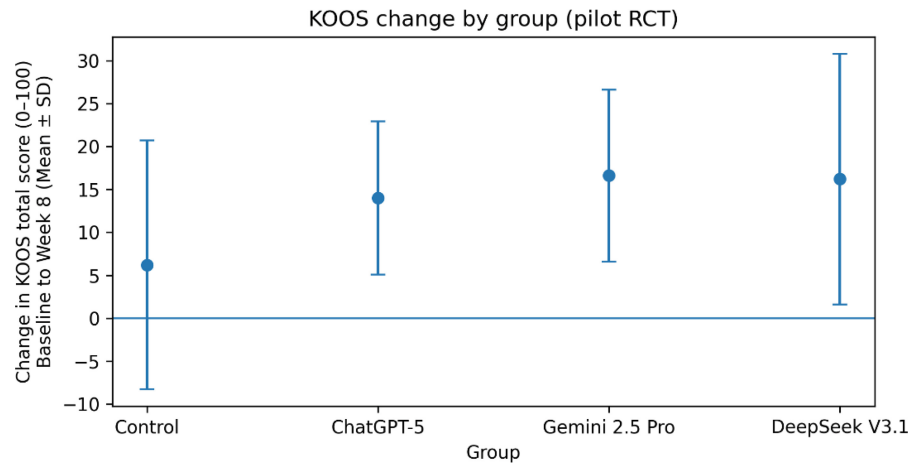


Fig. 1. Change in KOOS total score (0–100) from baseline to week 8 across groups (mean ± SD).

Table 2. Change in KOOS total score from baseline to week 8 (mean ± SD; n = 13 per arm).

Group	ΔKOOS (mean ± SD)
Control	6.2 ± 14.5
ChatGPT-5	14.0 ± 8.9
Gemini 2.5 Pro	16.6 ± 10.0
DeepSeek V3.1	16.2 ± 14.6

in Fig. 1. One-way ANOVA for KOOS change did not reach statistical significance ($p \approx 0.12$; $\eta^2 \approx 0.11$), consistent with the underpowered pilot design.

Secondary outcomes showed within-group improvements across the 8-week intervention in most arms, without a consistent pattern of between-group separation. No serious adverse events occurred; only transient knee or muscle soreness consistent with exercise exposure was reported.

4 Discussion

This pilot RCT suggests that clinician-supervised, LLM-generated exercise protocols are feasible, safe, and yield numerically larger (though non-significant) KOOS improvements than standard care in DKD [11]. Guideline-concordant supervised exercise likely drove the meaningful within-group gains across all arms [2–4]. The greater LLM benefits may stem from these models enforcing more explicit progression rules and broader exercise variety than routine clinical templates, addressing known variabilities in program delivery [2, 3, 5].

While orthopaedic LLM applications face valid concerns regarding hallucinations and limited patient-outcome data [6, 7], our “human-in-the-loop” design mitigates these risks. By using models strictly as protocol generators under expert oversight with identical co-interventions, we align with AI clinical trial governance [12]. Limitations include the single-centre setting, small sample size, 8-week follow-up, and absence of formal fidelity metrics [11]. Future adequately powered trials should incorporate protocol content audits and long-term outcomes.

5 Conclusion

Clinician-supervised LLM-generated rehabilitation protocols for DKD are safe and produced numerically larger KOOS improvements than a clinician-designed standard protocol. Larger trials are required to confirm comparative effectiveness and optimize “LLM-as-protocol-generator” workflows.

Acknowledgments. This study received no external funding or grants.

Disclosure of Interests The authors declare no conflicts of interest.

Declaration of AI Assistance During the preparation of this manuscript, the authors used large language models (including Gemini and ChatGPT) for language refinement and data summarization. The authors reviewed and edited the content as needed and take full responsibility for the final publication.

References

1. Safiri, S., et al.: Global, regional and national burden of osteoarthritis 1990-2017: a systematic analysis of the global burden of disease study 2017. *Ann. Rheum. Dis.* **79**(6), 819–828 (2020)
2. Bannuru, R.R., et al.: OARSI guidelines for the non-surgical management of knee, hip, and polyarticular osteoarthritis. *Osteoarthr. Cartil.* **27**(11), 1578–1589 (2019)

3. Kolasinski, S.L., et al.: 2019 American College of Rheumatology/Arthritis Foundation guideline for the management of osteoarthritis of the hand, hip, and knee. *Arthritis Rheumatol.* **72**(2), 220–233 (2020)
4. Goh, S.-L., et al.: Efficacy and potential determinants of exercise therapy in knee and hip osteoarthritis: a systematic review and meta-analysis. *Ann. Phys. Rehabil. Med.* **62**(5), 356–365 (2019)
5. Whitfield, M., Tomlinson, O.W.: Optimal exercise modalities and doses for therapeutic management of osteoarthritis of the knee. *Front. Aging.* **6**, 1458983 (2025)
6. Ma, Z., et al.: Clinical applications of large language models in knee osteoarthritis: a systematic review. *Front. Med.* **12**, 1670824 (2025)
7. Zhang, C., et al.: Examining the role of large language models in orthopedics: systematic review. *J. Med. Internet Res.* **26**, e59607 (2024)
8. Beheshti, M., et al.: Evaluating the reliability of ChatGPT for health-related questions: a systematic review. *Informatics.* **12**(1), 9 (2025)
9. Gürses, Ö.A., et al.: The role of artificial intelligence large language models in personalized rehabilitation programs for knee osteoarthritis: an observational study. *J. Med. Syst.* **49**(1), 73 (2025)
10. Sawamura, S., et al.: Evaluation of the accuracy of ChatGPT's responses to and references for clinical questions in physical therapy. *J. Phys. Ther. Sci.* **36**(5), 234–239 (2024)
11. Eldridge, S.M., et al.: CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *BMJ.* **355**, i5239 (2016)
12. Liu, X., et al.: Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit. Health.* **2**(10), e537–e548 (2020)
13. Roos, E.M., et al.: Knee injury and osteoarthritis outcome score (KOOS)—development of a self-administered outcome measure. *J. Orthop. Sports Phys. Ther.* **28**(2), 88–96 (1998)



Scaling Healthcare Engagement with AI Voice Agents: A Real-World Evaluation Study

Robert Ellis¹(✉) , Michael Goodman², Sai Moturu¹, Sherene Philip²,
Paula Buzzard², Anmol Madan¹ , and Sameer Berry²

¹ RadiantGraph, Inc., San Francisco, CA 94107, USA
rj.ellis@radiantgraph.com

² Oshi Health, Inc., New York, NY 10003, USA

Abstract. Healthcare organizations often lack the resources to perform member outreach at scale. While conversational AI shows promise across healthcare domains, rigorous evaluations of voice-based outbound engagement remain limited. We evaluated a seven-week controlled deployment of AI voice agents for outreach to individuals across three journey stages within a virtual specialty care program. Recipients of AI calls had significantly higher appointment completion rates, with effects strongest for members who had already created program accounts but had yet to schedule an appointment. Call answer rate and opt-out rate both outperformed industry benchmarks, with zero safety escalations identified. The deployment yielded a nine-fold increase in outbound call volume and significant efficiency gains over human agents alone. Together, these results suggest that AI voice agents represent a viable approach for scaling personalized healthcare outreach.

Keywords: AI voice agents · Member engagement · Controlled evaluation

1 Introduction

Healthcare organizations know who needs outreach: patients who missed appointments, are due for screenings, or are primed for the next step on their care journey. What's lacking is a capacity to act. Call centers are stretched thin, staff overburdened, and personalized outreach at scale remains elusive due to cost and ROI considerations. The result is a reactive system, where intervention arrives too late, if at all.

To date, investigations of conversational agents for healthcare have largely focused on text-based, user-initiated systems for health counseling, behavioral interventions, or chronic disease support [1, 2]. Recent commentary highlights the transformative potential of AI voice agents [3, 4], while noting a need for rigorous validation.

We present the design, deployment, and evaluation of a purpose-built conversational AI platform to deliver natural, HIPAA-compliant voice interactions at scale. The evaluation was a collaboration between RadiantGraph and Oshi Health, a virtual-first specialty clinic for gastrointestinal conditions. To our knowledge, this represents one of the first controlled evaluations of AI-initiated voice outreach for healthcare.

2 Methods

2.1 Platform and Technical Implementation

We designed a platform (RadiantGraph AI Voice Studio) to facilitate the deployment of conversational AI agents capable of naturalistic dialogue while maintaining strict adherence to scripted flows and compliance guardrails. Each agent was defined by its identity, knowledge scope, task sequence (e.g., detect voicemail, verify identity, confirm eligibility, execute call objectives, answer questions, facilitate transfer requests), error handling protocols, and operational guidelines. For a demonstration, please visit <https://radiantgraph.com/platform/ai-voice-studio>.

Multiple AI subsystems were integrated to support live conversation: automatic speech recognition (ASR) for transcription, a large language model (LLM) for conversational reasoning, text-to-speech (TTS) synthesis for natural voice output, and real-time safety monitoring. Safety guardrails were designed to detect medical advice requests, self-harm indicators, emergency situations, and other scenarios requiring handoff to a human agent (or direction to contact emergency services).

The underlying architecture separated a stateless telephony layer from a conversation orchestrator managing the ASR → LLM → TTS cycle. Conversations were built from reusable modules for each task sequence, with explicit goals and transition logic, addressing the limitations of traditional rule-based dialogue systems [1].

2.2 Deployment Design and Analysis

The target program behavior was a member completing their first telehealth appointment with a GI specialist. Three stages were defined among individuals yet to do so. “Stage A” were prospective members without accounts (mean age: 44, SD age: 14; 76% female). “Stage B” members had created an account, but never scheduled their first appointment (mean age: 43, SD age: 13; 67% female). “Stage C” members had previously scheduled, but not completed, their first appointment (mean age: 42, SD age: 13; 69% female). A unique AI voice agent was optimized for each stage. All members in Stages B and C received a weekly cadence of SMS outreach (independent of whether they also received AI voice outreach).

The deployment utilized a randomized allocation approach, with rolling enrollment over seven weeks (September–October, 2025). Each week, newly eligible individuals were assigned to their journey stage; a random selection from each stage received AI voice outreach (Treatment group). Once assigned, re-allocation was prohibited. Treatment allocations were constrained to enable 100% manual QA auditing and accommodate limited live-agent transfer capacity.

Post-hoc analysis of de-identified, aggregate data assessed call impact. The primary metric for all stages was appointment *completion*. The secondary metric was stage-specific *progression*: account creation (Stage A), first appointment scheduling (Stage B), or appointment completion (Stage C). Thus, for Stage C members, primary and secondary metrics were equivalent. Metrics were measured over 30 days from the call (Treatment) or initial eligibility date (Control). A one-tailed Fisher’s exact test was used, given the directional hypothesis that outreach would improve metrics.

3 Results

3.1 Impact on Performance Metrics (Treatment vs. Control Groups)

Across the total study sample, the Treatment group outperformed the Control group on both primary and secondary metrics. This overall success was driven primarily by members in Stage B; differences in Stages A and C did not reach statistical significance, suggesting that the effectiveness of AI voice outreach is highly dependent on where persons are in their program journey (Table 1).

Table 1. Summary of performance metrics

Stage	Metric	Control	Treatment	Odds ratio	<i>P</i> -value
Total	Sample size	4,717	4,082	—	—
	Appt. completion (primary metric)	15 (0.32%)	44 (1.08%)	3.42	<.001
	Stage progression (secondary metric)	16 (0.34%)	56 (1.37%)	4.09	<.001
Stage A	Sample size	446	1,388	—	—
	Appt. completion (primary metric)	6 (1.35%)	21 (1.51%)	1.13	.503
	Acct. creation (secondary metric)	6 (1.35%)	32 (2.31%)	1.73	.146
Stage B	Sample size	2,345	1,437	—	—
	Appt. completion (primary metric)	1 (0.04%)	15 (1.04%)	24.71	<.001
	Appt. scheduling (secondary metric)	2 (0.09%)	16 (1.11%)	13.18	<.001
Stage C	Sample size	1,926	1,257	—	—
	Appt. completion (primary and secondary metric)	8 (0.42%)	8 (0.64%)	1.54	.269

3.2 Impact on Call Metrics (Treatment Group)

Call answer rates and opt-out rates were measured for the Treatment group; for brevity, we report across the entire sample. The overall answer rate was 23.99%, higher than reported industry benchmarks ranging from 2% to 20% [5, 6]. The overall opt-out rate was 0.12%, substantially lower than a reported 1.5% opt-out rate from interactive voice response (IVR) healthcare messaging [7]. All call transcripts were audited by an experienced human QA reviewer, with zero escalations for safety concerns identified.

3.3 Impact on Operational Efficiency

Benchmarked against historical human-agent outreach within this same population, the present deployment enabled a nine-fold increase in realized outbound call volume without additional staffing. For every 1,000 persons contacted, the platform saved approximately 75 h of human agent time; a roughly 70% efficiency gain.

4 Discussion

Taken together, these results suggest that AI voice agents are a promising complement to traditional digital channels for scaling personalized healthcare engagement.

With respect to key program behaviors, we observed a potential “sweet spot” within the program journey where the benefit of AI voice outreach is most effective; specifically, guiding members to schedule and complete a first appointment (Stage B), with smaller effects observed for earlier (Stage A) and later (Stage C) journey stages.

Additionally, the concrete savings in human agent time demonstrate that AI voice outreach can address a common operational gap: between the outreach that should happen, and what resource-constrained organizations can realistically deliver.

This deployment represents a step toward a *hybrid intelligence model* [8, 9]: AI that can handle high-volume, structured outreach to facilitate early steps in care, while clinicians focus on the clinical encounters themselves. Empirical evidence supports this division of labor: AI can match human performance for tasks like screening and triage [3, 4], yielding significant cost savings in healthcare delivery [10].

4.1 Limitations and Future Work

We note several limitations, largely driven by the nature of conducting a real-world evaluation of a convenience sample. First, the available study population had no eligibility constraints based on known GI needs or risk factors, which may explain the low observed rates for primary and secondary metrics.

Second, members in Stages B and C received concurrent SMS outreach; the present evaluation assessed the incremental impact of adding AI voice outreach, not a comparison of AI voice to human agents, or to no contact at all.

Additionally, the weekly cadence of data refreshes meant that some individuals in the Treatment group may have already completed a target action prior to being called. This would bias against a Treatment effect, making our estimates conservative.

Future work spans both platform development and continued evaluation. Technical priorities include (1) expanded integrations with client systems to simplify eligibility verification and facilitate real-time scheduling, (2) adaptive personalization using member interaction patterns, (3) multilingual capabilities, and (4) robust engagement metrics extracted from call transcripts, including dialogue progression milestones. We are also planning larger-scale studies with diverse populations and clinical contexts [11] to further validate the effectiveness of AI voice outreach and establish best practices for deployment.

As healthcare shifts from reactive episodic treatment to proactive continuous support, AI voice platforms provide the infrastructure to operationalize precision outreach at scale—extending personalized engagement to populations underserved due to resource constraints.

Acknowledgments. Our gratitude to Steven Kollars, Heather Max, and Carlyne Williams for their tireless dedication to the success of this project.

Disclosure of Interests Robert Ellis, Anmol Madan, and Sai Moturu are employees of Radiant-Graph. Sameer Berry, Paula Buzzard, Michael Goodman, and Sherene Philip are employees of Oshi Health.

Declaration of AI Assistance Technologies During Manuscript Preparation The authors used Anthropic's Claude for the refinement, correction, editing, and formatting of the manuscript to improve clarity of language. After using this technology, the authors reviewed the results and take full responsibility for the contents of the manuscript.

References

1. Laranjo, L., et al.: Conversational agents in healthcare: a systematic review. *J. Am. Med. Inform. Assoc.* **25**(9), 1248–1258 (2018). <https://doi.org/10.1093/jamia/ocy072>
2. Li, Y., et al.: Feasibility and effectiveness of artificial intelligence-driven conversational agents in healthcare interventions: a systematic review of randomized controlled trials. *Int. J. Nurs. Stud.* **143**, 104494 (2023). <https://doi.org/10.1016/j.ijnurstu.2023.104494>
3. Adams, S.J., Acosta, J.N., Rajpurkar, P.: How generative AI voice agents will transform medicine. *npj Digit. Med.* **8**, 353 (2025). <https://doi.org/10.1038/s41746-025-01776-y>
4. Mahajan, A., Powell, D.: Transforming healthcare delivery with conversational AI platforms. *npj Digit. Med.* **8**, 581 (2025). <https://doi.org/10.1038/s41746-025-01968-6>
5. Live Answer Rate. <https://saleshive.com/glossary/live-answer-rate/>. Last accessed 21 Jan 2026
6. State of the Call 2025. <https://www.hiya.com/state-of-the-call>. Last accessed 21 Jan 2026
7. Steiner, J.F., et al.: Factors associated with opting out of automated text and telephone messages among adult members of an integrated health care system. *JAMA Netw. Open.* **4**(3), e213479 (2021). <https://doi.org/10.1001/jamanetworkopen.2021.3479>
8. Liu, P., Zhang, J., Chen, S., Chen, S.: Human-AI teaming in healthcare: 1+1>2? *npj Artif. Intell.* **1**, 47 (2025). <https://doi.org/10.1038/s44387-025-00052-4>
9. Bellini, V., Badino, M., Maffezzoni, M., Bezzi, F., Bignami, E.: Evolution of hybrid intelligence and its application in evidence-based medicine: a review. *Med. Sci. Monit.* **29**, e939366 (2023). <https://doi.org/10.12659/MSM.939366>
10. Khanna, N.N., et al.: Economics of artificial intelligence in healthcare: diagnosis vs. treatment. *Healthcare.* **10**(12), 2493 (2022). <https://doi.org/10.3390/healthcare10122493>
11. Nadarzynski, T., et al.: Achieving health equity through conversational AI: A roadmap for design and implementation of inclusive chatbots in healthcare. *PLoS Digit. Health* **3**(5), e0000492 (2024). <https://doi.org/10.1371/journal.pdig.0000492>



Conformal Prediction for Risk-Controlled Medical Entity Extraction Across Clinical Domains

Manil Shrestha^(✉) and Edward Kim

Department of Computer Science, Drexel University, Philadelphia, USA
{ms5267, ek826}@drexel.edu

Abstract. Large Language Models (LLMs) are increasingly used for medical entity extraction, yet their confidence scores are often miscalibrated, limiting safe deployment in clinical settings. We present a conformal prediction framework that provides finite-sample coverage guarantees for LLM-based extraction across two clinical domains. First, we extract structured entities from 1,000 FDA drug labels across eight sections using GPT-4.1, verified via FactScore-based atomic statement evaluation (97.7% accuracy over 110,664 entities). Second, we extract radiological entities from MIMIC-CXR reports using the RadGraph schema with GPT-4.1 and Llama-4-Maverick, evaluated against physician annotations (entity F1: 0.81 to 0.84). Our central finding is that miscalibration direction reverses across domains: on well-structured FDA labels, models are underconfident, requiring modest conformal thresholds, while on free-text radiology reports, models are overconfident, demanding strict thresholds ($\hat{\tau}$ approaching 1.0). Despite this heterogeneity, conformal prediction achieves coverage close to the $(1 - \alpha)$ target on held-out test sets in both settings with manageable rejection rates (9–13%). These results demonstrate that calibration is not a global model property but varies across document types, extraction categories, and model architectures, motivating domain-specific conformal calibration for safe clinical deployment.

Keywords: Large Language Models · Entity Extraction · Conformal Prediction

1 Introduction

Extracting structured entities from unstructured medical text is critical for drug safety surveillance and clinical coding, and LLMs have shown strong performance on these tasks [11, 15]. However, LLM softmax probabilities are poorly calibrated [6, 7], providing no reliability guarantees for clinical deployment.

Conformal prediction (CP) addresses this by converting raw scores into acceptance decisions with finite-sample coverage guarantees under only an exchangeability assumption [2]. CP has been applied to medical imaging [4],

pathology coding [13], and EHR surveillance [5]. In our prior work [10], we applied CP to EHR extraction across 10,000 visits with thresholds validated against expert consensus [3]. Separately, LLM extraction has advanced on FDA labels [15] and radiology reports via the RadGraph schema [8,9], though calibration guarantees remain unaddressed.

We extend our conformal framework to these two domains using GPT-4.1 [1] and Llama 4 Maverick [12]. We restrict Llama-4-Maverick to the RadGraph task because it provides physician-annotated gold standards enabling direct cross-model comparison; FDA-label evaluation uses GPT-4.1 alone, building on our prior LLM-judge pipeline [10]. We find that miscalibration reverses across document types: on well-structured FDA labels, models are underconfident, while on free-text radiology reports, models are overconfident. Our framework adapts to both regimes, providing valid coverage guarantees and enabling domain-specific routing between automated acceptance and human review.

2 Methods

Both models return per-token log-probabilities. The pipeline has four steps.

Step 1: Extraction with Token Confidence. The LLM extracts entities from eight FDA label sections (indications, adverse reactions, drug interactions, contraindications, warnings, pregnancy, pediatric and geriatric use) and from radiology reports following the RadGraph schema [8] with anatomy (ANAT) and three observation types: definitely present (OBS-DP), absent (OBS-DA), and uncertain (OBS-U). For each entity span of m_e tokens with softmax probabilities $p_{e,t}$, we compute a span-level confidence via the geometric mean:

$$\hat{p}_e = \exp\left(\frac{1}{m_e} \sum_{t=1}^{m_e} \log p_{e,t}\right) \quad (1)$$

Step 2: Verification. For FDA labels, GPT-5-mini serves as an LLM-as-a-judge using the VeriFact framework [3], producing a fact-score from 0 to 3; only score 3 counts as correct ($y_e = 1$). We validated this verification approach against three-physician adjudication in our prior work [10], achieving expert-level agreement (87.3% vs. 88.5% inter-physician) on 13,290 propositions. For radiology reports, we match extracted entities against the RadGraph gold annotations.

Step 3: Nonconformity Score. We define $s_e = -\text{logit}(\hat{p}_e)$, mapping confidence from $(0, 1)$ to $(-\infty, +\infty)$ and spreading out the high-confidence region where most entities cluster.

Step 4: Split Conformal Calibration. For each category c , we partition verified entities 70/30 into calibration and test sets. On the calibration set of size n_c , we sort nonconformity scores and set the threshold:

$$\tau_c = s_{\lceil (1-\alpha)(n_c+1) \rceil} \quad (2)$$

Entities with $s_e \leq \tau_c$ are accepted; the rest are flagged for human review. By exchangeability, at least $(1-\alpha)$ of correct extractions are retained regardless of the model’s underlying calibration quality.

3 Results and Discussion

FDA Label Extraction. We extracted 110,664 entities from 1,000 FDA drug labels across eight sections using GPT-4.1, achieving 97.7% overall fact-score accuracy (per-section range: 74.3% for Pediatric Use to 100% for Geriatric Use). Calibration curves (Fig. 1) reveal that GPT-4.1 is *underconfident* on FDA labels: predicted confidence $\hat{p}$ consistently falls below empirical accuracy. ECE ranges from 0.022 (Adverse Reactions, nearly perfectly calibrated) to 0.211 (Pediatric Use, the sole exception, likely due to sparser safety language).

At $\alpha = 0.10$, conformal prediction achieves empirical coverage of 88.5% to 90.8% (close to the 90% target) with 9.2% to 11.5% rejection rates across all sections. The model’s underconfidence is advantageous here, keeping conformal thresholds modest, in contrast to our prior EHR work [10] where overconfidence forced near-maximum thresholds.

RadGraph Entity Extraction. We evaluated extraction on 100 MIMIC-CXR reports against test set of RadGraph gold annotations. With 5-shot prompting from dev set, GPT-4.1 achieves entity F1 of 0.826 and relation F1 of 0.502. Few-shot examples are critical: zero-shot recall drops 14.6 points and relation F1 nearly halves.

In stark contrast, both models are *overconfident* on radiology reports (Fig. 2), assigning near-certain probabilities to incorrect extractions. Llama achieves lower ECE than GPT-4.1 across all categories except OBS-U, which remains challenging due to hedging-language ambiguity. For OBS-DA, Llama requires $\hat{\tau} > 0.999$, reflecting high negation accuracy (ECE = 0.048). Rejection rates of 8.9–13.4% confirm conformal filtering flags a manageable fraction while maintaining target coverage (Table 1).

Implications and Conclusion. Our central finding is that LLM calibration direction reverses across clinical domains: underconfident on structured FDA labels, overconfident on free-text radiology reports. Conformal prediction adapts to both regimes automatically, with thresholds that are lenient where models are conservative and strict where they are overconfident. A key limitation is the requirement for token-level log-probabilities, which several frontier model APIs do not currently expose. Future work includes extending guarantees to black-box models via surrogate confidence estimation [14] and comparison with post-hoc calibration baselines.¹

¹ Declaration of AI Assistance: The authors used GitHub Copilot for writing or drafting manuscript content, and refinement or formatting of code reported in the submitted manuscript. After using this technology, the authors reviewed the results and take full responsibility for the contents of the manuscript.

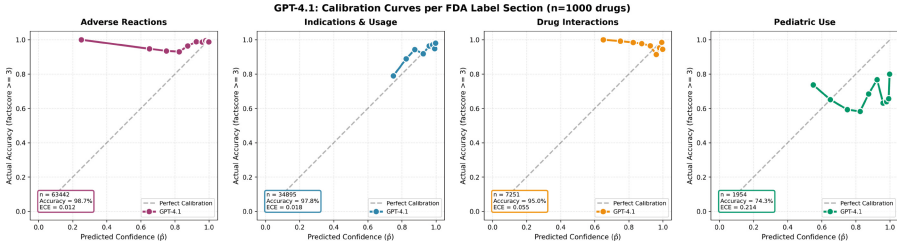


Fig. 1. Calibration curves for 4 sections from FDA labels extraction with GPT-4.1 ($n=1,000$ drugs). Predicted span confidence $\hat{p}$ (geometric mean of token probabilities) versus empirical accuracy. ECE ranges from 0.022 (Adverse Reactions) to 0.211 (Pediatric Use), motivating category-specific conformal thresholds.

Table 1. Per-category entity extraction and conformal prediction on 100 RadGraph test reports ($\alpha = 0.10$, 5-shot). $\hat{\tau}_c$ is the category-specific conformal threshold; entities with $\hat{p} < \hat{\tau}_c$ are flagged for review.

Category	GPT-4.1						Llama-4-Maverick					
	Prec.	Rec.	F1	ECE↓	$\hat{\tau}_c$	Rej.%	Prec.	Rec.	F1	ECE↓	$\hat{\tau}_c$	Rej.%
ANAT-DP	.834	.933	.880	.130	.763	8.9	.865	.893	.879	.085	.971	10.7
OBS-DP	.827	.769	.797	.123	.748	11.2	.905	.724	.804	.063	.941	11.2
OBS-DA	.833	.914	.871	.115	.782	8.9	.902	.900	.901	.048	>.999	10.1
OBS-U	.305	.648	.415	.528	.821	13.4	.444	.580	.502	.407	.898	9.6

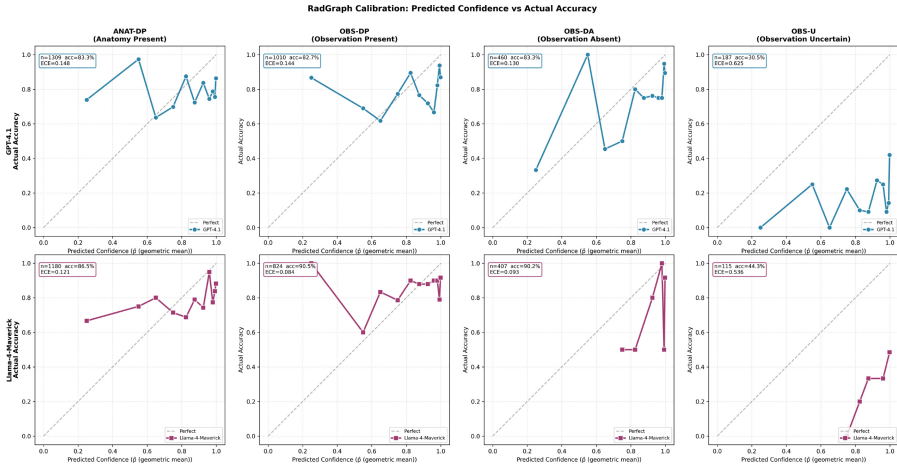


Fig. 2. Per-label calibration curves for GPT-4.1 and Llama-4-Maverick (5-shot, 100 CXR reports [9]). Predicted span confidence $\hat{p}$ (geometric mean of token probabilities) versus empirical accuracy across four entity categories. Llama-4-Maverick achieves lower ECE overall (0.085 vs 0.147), while OBS-U remains poorly calibrated for both models. Per-category ECE values are reported in Table 1.

References

1. Achiam, J., et al.: GPT-4 technical report. arXiv preprint [arXiv:2303.08774](https://arxiv.org/abs/2303.08774) (2023)
2. Angelopoulos, A.N., Bates, S.: Conformal prediction: a gentle introduction. *Found. Trends Mach. Learn.* **16**(4), 494–591 (2023)
3. Chung, P., et al.: VeriFact: verifying facts in LLM-generated clinical text with electronic health records. arXiv preprint [arXiv:2501.16672](https://arxiv.org/abs/2501.16672) (2025)
4. Fayyad, J., Alijani, S., Najjaran, H.: Empirical validation of conformal prediction for trustworthy skin lesions classification. *Comput. Methods Programs Biomed.* **253**, 108231 (2024)
5. Genari, J., Goedert, G.T.: Mining unstructured medical texts with conformal active learning. arXiv preprint [arXiv:2502.04372](https://arxiv.org/abs/2502.04372) (2025)
6. Geng, J., Cai, F., Wang, Y., Koepl, H., Nakov, P., Gurevych, I.: A survey of confidence estimation and calibration in large language models. In: *Findings of NAACL 2024*, pp. 6577–6595 (2024)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: *International Conference on Machine Learning*, pp. 1321–1330. PMLR (2017)
8. Jain, S., et al.: RadGraph: extracting clinical entities and relations from radiology reports. *PhysioNet* (2021)
9. Johnson, A.E., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**(1), 317 (2019)
10. Kim, E., Foty, R., Shrestha, M., Seyfert-Margolis, V.: Conformal prediction and verification of large language model extractions in EHR data. In: *Proceedings of the AAAI Symposium Series*, vol. 7, pp. 539–546 (2025)
11. Kim, E., Shrestha, M., Foty, R., DeLay, T., Seyfert-Margolis, V.: Structured extraction of real world medical knowledge using LLMs for summarization and search. In: *IEEE BigData 2024*, pp. 3421–3430. IEEE (2024)
12. Meta, A.: The Llama 4 herd: the beginning of a new era of natively multimodal ai innovation (2025)
13. Snyder, C., Brodsky, V.: Conformal prediction and large language models for medical coding. *Am. J. Clin. Pathol.* **162**, S171–S172 (2024)
14. Su, J., Luo, J., Wang, H., Cheng, L.: API is enough: conformal prediction for large language models without logit-access. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 979–995 (2024)
15. Wu, L., et al.: A framework enabling LLMs into regulatory environment for transparency and trustworthiness and its application to drug labeling document. *Regul. Toxicol. Pharmacol.* **149**, 105613 (2024)



Can LLMs Mimic Experienced Neurosurgeons in Real-World Spine Surgery Triage?

Mahvash Siavashpour^{1,2}(✉) , Erin McCabe¹ , Andrew Nataraj¹ ,
Niketh Pareek³ , Osmar Zaiane^{1,2} , and Douglas Gross¹

¹ University of Alberta, Edmonton, Canada
siavashp@ualberta.ca

² Alberta Machine Intelligence Institute, Edmonton, Canada

³ McGill University, Montreal, Canada

Abstract. Low back and neck pain are leading causes of disability, and some patients require surgical intervention to prevent long-term complications. Current triage workflows rely on MRI reports, questionnaires, and in-person assessment. However, many referred patients are ultimately managed non-surgically, contributing to long wait times and inefficient use of specialist resources. Automated decision support that distinguishes surgical from non-surgical cases could improve prioritization and care pathways.

This study investigates large language models (LLMs) for surgical triage using real-world data from the Spine Assessment Clinic in Edmonton, Canada. The dataset includes MRI reports and patient questionnaires, reflecting routine clinical inputs. We first evaluate conventional machine learning methods and find that these approaches perform poorly due to class imbalance, limited labelled data, and the complexity of radiology narratives. We then explore prompt-based use of open-source LLMs as an alternative to gradient-based training. LLMs outperform conventional models in this low-resource setting, particularly in unambiguous cases. Ablation analyses indicate that neither data modality alone is sufficient, but their combination yields the most reliable triage performance.

To the best of our knowledge, this work presents the first evaluation of LLMs for spine surgery triage on real-world clinical data. We benchmark multiple methods, examine multimodal integration strategies, and assess agreement with expert labels, highlighting both the promise and limitations of LLM-based clinical decision support.

Keywords: Large Language Models · Surgical Triage · Clinical Decision Support Systems

1 Introduction and Related Work

Low back and neck pain are common conditions, yet only a subset of patients require surgical intervention. In practice, many referred patients are managed

non-surgically, leading to unnecessary specialist consultations and prolonged wait times. At the Spine Assessment Clinic (SAC) in Edmonton, Canada, triage relies on diverse clinical inputs, including referral letters, MRI reports, patient questionnaires, and consultation findings. This motivates developing automated decision support systems to identify likely surgical candidates earlier. However, this task is challenging due to unstructured radiology narratives, high-dimensional and evolving questionnaire data, and limited, imbalanced labelled outcomes.

In recent years, machine learning has advanced clinical text processing [1]. However, triage using clinical text remains underexplored and often relies on curated datasets or narrow tasks. Prior work on the Spine Assessment Clinic triage showed that traditional NLP approaches can be outperformed by basic clinical variables, highlighting the difficulty of extracting surgical signals from radiology reports [3]. Multimodal approaches combining reports and questionnaires improve performance but still depend heavily on large labelled datasets [5]. LLMs can potentially offer an alternative by leveraging prior knowledge without task-specific training. Although recent studies examine LLMs for clinical reasoning and triage [2, 4], they focus mainly on emergency or synthetic settings, leaving their effectiveness for real-world surgical decision support unclear.

In this work, we address surgical triage for spine care using a real clinical dataset with limited samples and heterogeneous inputs. We evaluate prompting-based use of open-source LLMs and compare them directly with conventional machine learning approaches. This setting allows us to study the feasibility of privacy-preserving, locally deployed LLM decision support for a nuanced clinical task that has received little prior attention.

2 Method

2.1 Data

We use de-identified MRI reports from 408 Spine Assessment Clinic patients, each paired with a 65-item self-assessment questionnaire. Of these, 69 (16.9%) were surgical and 339 (83.1%) non-surgical. The questionnaire captures demographics, clinical symptoms, and disability (Oswestry Disability Index), among others. The outcome is the neurosurgeon’s final decision, made by one of five certified neurosurgeons. Decision codes were grouped into three categories: clear non-surgical (Codes 1–2), in-person assessments without surgery (Codes 3–4), and surgeon-involved decisions, including surgical (Codes 5–7) and non-surgical management (Codes 8–10). For consistency with clinical practices, decision codes 5, 6, and 7 are treated as surgical in the current work, while all the remaining codes are non-surgical.

2.2 LLM as Surgical Decision Support System

We propose a pipeline for leveraging the knowledge encoded in both the radiology reports and patient-completed questionnaires. We combine questionnaire and

report text with a prompt and use the LLM to classify patients as surgical or non-surgical. Here we describe the components of the pipeline:

To integrate questionnaire data into the model, each patient questionnaire is serialized into text following the `<field_name>: <field_value>` template. Questions with missing information are omitted from the serialized text. Radiology reports are processed by extracting the Findings and Impression sections [1], which are then combined with the serialized questionnaire to form different input representations for the LLM. To address the lack of fixed rules in spine surgery triage, this study evaluates multiple prompting strategies rather than relying on manually engineered prompts designed by medical experts, aiming to better align model predictions with expert decisions in a scalable way. For this purpose we evaluate three different prompting methods:

1. Plain Prompting: The model is provided with the patient’s data and asked to directly classify the patient as surgical or non-surgical.

2. Heuristic Prompting: The prompt asks the LLM to act as a neurosurgeon and apply heuristics, which are manually derived from the literature, to determine whether surgery is needed. These heuristics include symptom and pain patterns, and imaging findings.

3. Prompt Learning: In this approach, the LLM is given a set of ground-truth examples and asked to identify patterns in the data and generate a prompt that includes information about how to do a surgical triage. We then use this prompt along with the patient’s data to instruct a second LLM to make the final decision. Examples used in the prompt learning approach are selected randomly with an equal number of examples from each class.

Table 1. Qwen2.5-32B results compared to the baselines. Results are obtained under the best data combination for each model, averaged across folds (mean $\pm$ std).

Method	ROC AUC	Specificity	Sensitivity
Machine learning baseline			
Logistic Regression + BGE-M3	0.6172 $\pm$ 0.07	0.5958 $\pm$ 0.08	0.6385 $\pm$ 0.13
Fine-tuned RadBERT	0.5000 $\pm$ 0.00	1.0000 $\pm$ 0.00	0.0000 $\pm$ 0.00
Prompt-based LLM (Qwen2.5-32B)			
Plain Prompt	0.6681 $\pm$ 0.08	0.5219 $\pm$ 0.07	0.8143 $\pm$ 0.11
Heuristic Prompt	0.6918 $\pm$ 0.09	0.6133 $\pm$ 0.07	0.7703 $\pm$ 0.11
Prompt Learning	0.6172 $\pm$ 0.05	0.3629 $\pm$ 0.05	0.8714 $\pm$ 0.09

After constructing the prompt, the patient data and prompt are combined and fed into the LLM for classification. The model outputs a decision label along with a confidence score and key decision factors, improving interpretability and providing useful support for surgeons during triage.

3 Experimental Results

Table 1 reports the best-performing configuration for each model under its most informative data combination using 5-fold stratified cross-validation. The LLM temperature was set to 0 to eliminate stochasticity. Moreover, for brevity, we report only the strongest model per category, although more models were evaluated. The best results are obtained using the prompt-based method and when the model is instructed with explicit decision heuristics. Although prompt learning performs similarly to the baseline in ROC AUC, its high sensitivity suggests that letting the LLM craft decision rules from examples has resulted in a bias toward predicting the surgery class. In our experiments, limited training data and class imbalance prevented complex machine learning models, such as RadBERT [6], from learning robust patterns and resulted in them only predicting the majority class (negative), whereas simpler models could be trained more reliably. Additionally, Table 2 compares Qwen2.5-32B performance across three input modalities: MRI reports, questionnaires, and their combination. Questionnaire-only inputs show weak performance, suggesting limited predictive signal. Performance improves with radiology reports, but the best results come from combining both modalities, achieving the highest ROC AUC (0.69) and improved sensitivity. Overall, multimodal inputs are essential for accurate triage, and full reports outperform isolated sections.

Table 2. ROC AUC across input modalities. Abbreviations: Q = Questionnaire; RF = Report (findings); RI = Report (impression); R = Report (full); CF = Combined (findings); CI = Combined (impression); C = Combined (full).

	Q	RF	RI	R	CF	CI	C
ROC AUC	0.57	0.57	0.56	0.64	0.66	0.61	0.69

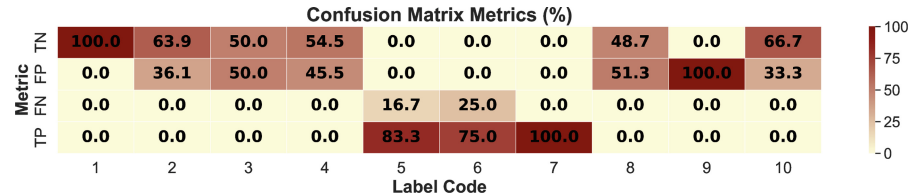


Fig. 1. Confusion matrix for Qwen2.5-32B (heuristic) on the full dataset. TN/FP/FN/TP denote true/false negatives/positives.

4 Discussion and Conclusions

After comparing the LLM results across patient categories in Fig. 1, we infer that performance is strongest for unambiguous surgical candidates (codes 5–7), indicating that the model reliably recognizes patients who ultimately require

imminent or urgent surgery. This supports the hypothesis that LLMs are effective at identifying clear surgical patterns. Performance is also strong for clear non-surgical cases decided without surgeon consultation (codes 1–2). However, the model performs poorly for borderline or uncertain clinical groups, which include patients seen by surgeons but managed conservatively or monitored over time. The results indicate that the model frequently predicts surgery when clinicians instead opt for observation or delayed surgical decision-making. This pattern implies that the model struggles with cases in which surgical candidacy depends on nuanced contextual factors, such as longitudinal response to treatment. The system therefore appears most suitable as a first-stage triage aid to help surgeons identify clearly surgical or clearly low-risk patients; it is intended only as a decision-support assistant, not a replacement for clinical judgment. Complex shared-decision cases still require expert review and may benefit from additional clinical context and more advanced prompt engineering.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Declaration of AI Assistance Technologies During Manuscript Preparation. The authors used ChatGPT (OpenAI) to paraphrase portions of the text. After using this technology, the authors have reviewed the results and take full responsibility for the content of the manuscript.

References

1. Han, Y., Yuan, Y., Mou, L.: A dual-view approach to classifying radiology reports by co-training. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 477–483. ELRA and ICCL, Torino, Italia, May 2024. <https://aclanthology.org/2024.lrec-main.42/>
2. Kim, S.H., et al.: Benchmarking the diagnostic performance of open-source LLMs in 1933 Eurorad case reports. *npj Digit. Med.* **8**(1), 97 (2025). <https://doi.org/10.1038/s41746-025-01488-3>
3. Krebs, B., et al.: Developing a triage predictive model for access to a spinal surgeon using clinical variables and natural language processing of radiology reports. *Eur. Spine J.* **35**(3), 1330–1343 (2026). <https://doi.org/10.1007/s00586-023-07552-4>
4. Lee, J., et al.: From promising capability to pervasive bias: assessing large language models for emergency department triage. *arXiv preprint arXiv:2504.16273* (2025)
5. Siavashpour, M., McCabe, E., Nataraj, A., Pareek, N., Zaiane, O., Gross, D.: A multimodal deep learning ensemble framework for building a spine surgery triage system. *Stud. Health Technol. Inf.* **329**, 1824–1825 (2025). <https://doi.org/10.3233/SHTI251233>
6. Yan, A., et al.: RadBERT: adapting transformer-based language models to radiology. *Radiol. Artif. Intell.* **4**(4), e210258 (2022). <https://doi.org/10.1148/ryai.210258>



Imitation Learning for Clinical Decision Support in Pediatric ECMO

Fateme Golivand Darvishvand¹(✉), Michael Skinner¹, Saurabh Mathur²,
Ameet Soni³, Phillip Reeder⁴, Kristian Kersting², Lakshmi Raman⁴,
and Sriraam Natarajan¹

¹ University of Texas at Dallas, Dallas, USA

fateme.golivanddarvishvand@utdallas.edu

² Technische Universität Darmstadt, Darmstadt, Germany

³ Swarthmore College, Swarthmore, USA

⁴ University of Texas Southwestern Medical Center, Dallas, USA

Abstract. Pediatric critical care is a dynamic, high-stakes process involving constant monitoring and adjustments in life-saving treatments. Modeling these interventions is crucial for effective decision support. In pediatric Extracorporeal Membrane Oxygenation (ECMO), where data are scarce and clinical dynamics are complex, we frame clinical decision-making as imitation learning from observational data, in which actions are not directly observed. We consider TabPFN, a recent transformer-based approach for tabular data, and traditional baselines including XGBoost and Multi-Layer Perceptrons on real-world pediatric ECMO data to learn the action models. We find that the TabPFN-based approach consistently outperforms these classical baselines, supporting its use as a strong clinician-behavior baseline for pediatric ECMO decision support.

Keywords: Pediatric ECMO · Imitation Learning · Clinical Decision Support · Tabular Transformers

1 Introduction

Managing patients on life support requires clinicians to make continuous, high-stakes adjustments to machine settings in response to rapidly changing physiological signals. However, automated decision support remains limited, particularly in specialized domains like pediatric Extracorporeal Membrane Oxygenation (ECMO) [5], a method used to support critically ill children with cardiac or respiratory failure. Building decision support systems for this domain is complicated by the incomplete and difficult-to-formalize nature of pediatric-specific protocols, alongside the ethical and practical impossibility of experimentation on critically ill children.

We acknowledge support from NIH award R01NS133142.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
P. Andreev et al. (Eds.): AIME 2026, LNAI 16749, pp. 377–382, 2027.
https://doi.org/10.1007/978-3-032-30813-9_70

We address this challenge of learning to manage pediatric ECMO by framing it as an imitation learning (IL) [4, 7] task; here, an AI agent learns a policy that maps observations to actions by mimicking expert behavior. Imitation learning from ECMO trajectories presents three key challenges, namely irregular clinical observation intervals, a concurrent, high-dimensional action space without explicit annotations, and severe data scarcity. We resolve these by aggregating trajectories into hourly windows, using expert knowledge to infer latent clinician actions, and employing TabPFN, a pretrained tabular transformer, as the base policy learner. We empirically evaluate this method on a small dataset of 78 pediatric ECMO trajectories, and find that the TabPFN-based learner outperforms traditional baselines such as a multilayer perceptron (MLP) and gradient-boosted trees, demonstrating the promise of tabular foundation models and IL for clinical decision support in this challenging setting¹.

2 Learning to Act in Pediatric ECMO

We aim to learn policies to manage children on ECMO life support by mimicking expert behavior in historical trajectories. We define the task as follows:

Given: Observational pediatric ECMO patient trajectories.

Task: Learn a treatment policy that imitates clinician behavior.

We address this task using Imitation Learning (IL) [4]. Unlike reinforcement learning [9], which requires active interaction, offline IL uses expert demonstrations already available in the historical clinical records, making it more suitable for pediatric ICU settings. To formalize pediatric ECMO as IL, we first derive latent clinical actions from the patient trajectories using expert knowledge. We then decompose the imitation learning task into a factorized, hierarchical policy model, predicting interventions independently per action knob and separating the identification of a change from the determination of its direction. Finally, we instantiate the imitation learner using the TabPFN tabular foundation model and compare it with instantiations based on traditional machine learning models.

Each trajectory is represented as a sequence of state-action pairs, (s_t, a_t) . The state s_t is constructed from four primary data streams: hemodynamics, laboratory test results, ECMO circuit variables, and ventilator settings. For each selected variable, we use both the mean hourly value and its delta relative to the preceding hour to capture short-term temporal trends. The action a_t is modeled as a multi-dimensional vector $(a_t^1, \dots, a_t^K)$, allowing for the representation of simultaneous adjustments across K distinct “knobs”. Since ECMO trajectories lack explicit intervention labels, we infer each knob-specific action k by monitoring its delta over a 1-h window:

$$a_t^k = \begin{cases} \text{Increase,} & x_{t+1}^k - x_t^k > \delta_k, \\ \text{Decrease,} & x_{t+1}^k - x_t^k < -\delta_k, \\ \text{Same,} & \text{otherwise,} \end{cases}$$

¹ Code is available at https://github.com/fateme-gd/ImitationLearning_ECMO.git.

where δ_k is a physician-defined threshold. This converts physiologic trajectories into supervised demonstrations while filtering small fluctuations that are unlikely to reflect meaningful intervention. This state-action definition allows us to make the first-order Markov assumption [8], approximating clinician behavior as the conditional distribution of a decision at time t given only the current state instead of the full history: $p(A_t | s_t)$.

To address the high-dimensional action space and the dominance of the “Same” action, we adopt a factorized, hierarchical policy structure. We train independent predictors for each knob, which allows the model to recommend concurrent interventions while avoiding the combinatorial explosion of a joint action space of 3^K possible composite actions. Each knob-specific head employs a two-stage process: first, a binary classifier identifies a change; second, conditional on a change, a classifier determines its direction.

We instantiate this framework using the TabPFN 2.5 tabular transformer [3]. TabPFN encodes a broad tabular prior from its extensive pretraining on millions of synthetic datasets. This enables it to perform probabilistic inference in context, producing calibrated predictive probabilities in a single forward pass without requiring dataset-specific gradient-based training or hyperparameter tuning. We compare this instantiation with ones based on a 3-layer feedforward neural network (MLP) trained with gradient-based optimization [2] and a gradient-boosted model (XGBoost) consisting of 200 shallow trees [1].

3 Evaluation

We evaluate the generalization performance, calibration, and alignment with clinician decisions of the three imitation learners using a small pediatric cohort.

Dataset. We obtained 78 pediatric ECMO trajectories from Children’s Medical Center, Dallas; these excluded cases with congenital heart disease due to the complexity of specialized management. The cohort had an average age of 4.66 years and an average trajectory length of 215 h. The state vector consists of 48 features, encompassing 23 physician-selected clinical variables across hemodynamics, ECMO circuit configurations, ventilator settings, and laboratory results; their one-hour deltas; the ECMO type; and an active-support indicator. To account for pediatric developmental variance, heart rate and mean arterial pressure are age-normalized. Action labels are derived using the physiological thresholds defined in Table 1².

Table 1. Actionable features (knobs) and physician-defined action thresholds. A discrete action is identified when the absolute change within a 60-minute window exceeds the specified threshold.

Knob	Threshold
Arterial PO ₂	25 mmHg
Arterial PCO ₂	5 mmHg
SpO ₂	5%
FiO ₂	10%

² A complete description of the state and action space is provided in <https://github.com/fateme-gd/ImitationLearning-ECMO/blob/master/y.pdf>.

Metrics. Given the inherent class imbalance, we quantify the performance of the imitation learners using balanced accuracy and macro-F1 scores for each action head, alongside aggregate macro-precision and macro-recall. Additionally, we quantify model calibration using Expected Calibration Error (ECE) [6], where lower values indicate that the model’s predicted confidence aligns more closely with its empirical accuracy. We compute these metrics using leave-one-out (LOO) cross-validation: in each fold, one patient trajectory is held out for testing, and all remaining patients are used for training. We apply the same LOO protocol and identical two-stage pipeline for all methods.

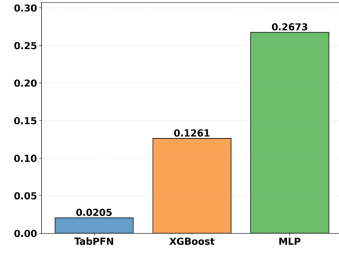


Fig. 1. Expected Calibration Errors of imitation learners based on TabPFN, XGBoost, and MLP.

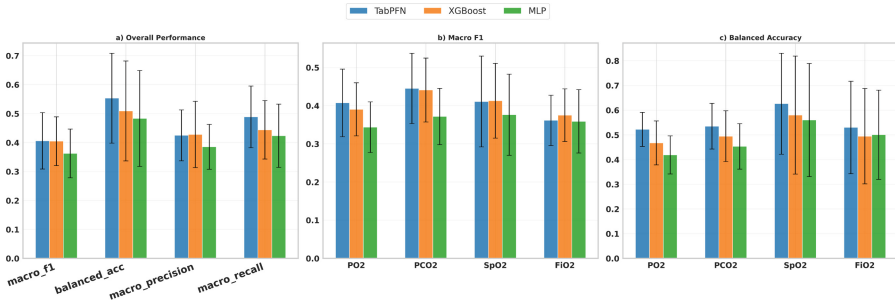


Fig. 2. Action selection performance of imitation learners based on TabPFN, XGBoost, and MLP. Metrics computed using leave-one-out cross-validation. (a) Aggregate performance. (b–c) Per-action-head performance. Values show mean $\pm$ standard deviation.

Results. We now present our results, organized around three questions.

Q1: Do the learned policies generalize to unseen data? TabPFN achieves the highest mean balanced accuracy and macro-F1 across the majority of the actions (Fig. 2). The consistent improvement in balanced accuracy suggests that TabPFN is more resilient to the class imbalance inherent to this setting and generalizes more effectively to held-out patients.

Q2: Are the policies calibrated? Beyond raw accuracy, clinical decision support models must be well-calibrated to be trustworthy. TabPFN consistently achieves a lower ECE than the MLP and XGBoost baselines (Fig. 1). This indicates that the posterior probabilities produced by TabPFN’s in-context inference closely match its empirical accuracy, providing reliable confidence estimates.

Q3: Where do the learned policies disagree with the clinicians? Beyond quantitative metrics, we investigate the specific physiological conditions under which the learned policies diverge from human expert decisions. To identify

these regions of the state space, we extracted the most frequent features across states with the highest model-expert disagreement by training shallow decision trees on the residual errors of their predicted probabilities (Table 2).

Oxygenation-related signals, specifically FiO_2 and SpO_2 , consistently define the boundaries where model predictions deviate from clinician actions. However, the nature of these disagreements varies by model type: TabPFN disagreements are most often associated with FiO_2 , and SpO_2 , whereas MLP disagreement more often involves PO_2 and lactate. This suggests that the models deviate from clinician actions in different regions of the state space and motivates future expert review of high-disagreement states.

Table 2. Most frequent features appearing in high disagreement states, across all knobs for each model.

Feature	MLP	TabPFN	XGBoost
FiO_2	2	5	5
SpO_2	1	5	3
pH	2	3	2
PO_2	3	2	1
PCO_2	2	3	1
Lactate	3	1	1

4 Conclusion

We presented an Imitation Learning framework to model clinical decision-making in pediatric ECMO. Our framework maps patient states to concurrent clinical interventions by inferring action labels from patient trajectories and employing a factorized, hierarchical policy architecture. Our empirical evaluation finds that TabPFN consistently outperforms traditional gradient-based baselines, suggesting that it can serve as an effective clinician-behavior baseline in challenging critical care settings. There are two key directions for future work. First, while we identify regions of model-clinician disagreements, future work should analyze whether they correlate with patient outcomes, such as risk of neurological injury. Second, while our policies accurately replicate clinician behavior, future work should explore policy improvement by incorporating expert-defined rewards.

References

1. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: KDD (2016)
2. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep Learning, vol. 1. MIT Press Cambridge (2016)
3. Hollmann, N., et al.: Accurate predictions on small data with a tabular foundation model. *Nature* **637**(8045), 319–326 (2025)
4. Hussein, A., Gaber, M.M., Elyan, E., Jayne, C.: Imitation learning: a survey of learning methods. *ACM Comput. Surv. (CSUR)* **50**(2), 1–35 (2017)
5. Lin, J.C.: Extracorporeal membrane oxygenation for severe pediatric respiratory failure. *Respir. Care* **62**(6), 732–750 (2017)
6. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using Bayesian binning. In: AAAI, vol. 29 (2015)
7. Natarajan, S., Joshi, S., Tadepalli, P., Kersting, K., Shavlik, J.: Imitation learning in relational domains: a functional-gradient boosting approach. *IJCAI* (2011)

8. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons (2014)
9. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT (2018)



Towards Multidisciplinary Summarization of Hospital Stays: Efficient Sentence-Level Clinical Section Categorization

Baris Karacan¹(✉), Vaibhav Bhargava¹, Barbara Di Eugenio¹, Natalie Parde¹, Mary Khetani¹, Yu-Shan Tseng¹, Vanessa Barbosa¹, Julie Vignato², Lindsey Knake², Rajashree Dahal¹, Emily Spellman², Danielle Hitzel², Janine Petitgout², Kristi Haughey², Amanda Karstens², Brianna Clarahan², Rachel Dawson², Lauren Boyd⁴, Mackenzie Weis³, Angie Tipton³, Jaewon Bae², Catherine K. Craven³, Karen Dunn Lopez², and Andrew D. Boyd¹

¹ University of Illinois Chicago, Chicago, USA
bkarac3@uic.edu

² University of Iowa, Iowa City, USA

³ University of Missouri, Columbia, USA

⁴ Loyola College, Chennai, India

Abstract. Effective “all-team” summarization in high-complexity settings like the Neonatal Intensive Care Unit (NICU) requires aggregating insights from diverse disciplines (physicians, nurses, therapists) spread across hundreds of clinical free-text notes. Simply pooling heterogeneous text often leads to incoherent outputs. Structured summarization therefore first requires accurate categorization of sentence-level section labels across multi-source notes. This pilot study introduces a clinical section categorization pipeline using supervised fine-tuning (SFT) of large language models (LLMs). We adapted two Llama-3 models (8B and 70B) to MedSecId, a corpus of 2,002 MIMIC-III (Adult ICU) notes annotated with clinical section headers, achieving in-domain Macro F1 scores above 92% for both models. To evaluate cross-domain generalization, we assessed model capacity (8B vs. 70B) and quantization on a gold-standard dataset of 227 sentence-level spans derived from three multidisciplinary NICU summaries. Experimental results suggest a potential scale-dependent transfer effect: while SFT produced only marginal changes for the 8B model, it substantially improved the 70B model, increasing Macro F1 by 7%. Notably, the quantized fine-tuned 70B model outperformed its full-precision baseline while substantially reducing computational requirements. These findings suggest that sufficient model capacity is critical for preserving semantic flexibility during cross-domain clinical transfer and that efficient quantized adaptation can enable structured section-label modeling for downstream summarization.

Keywords: Clinical NLP · Large Language Models · Supervised Fine-Tuning · Clinical Section Categorization

1 Introduction

Critical care in high-complexity environments like the Neonatal Intensive Care Unit (NICU) relies on synchronized expertise across diverse disciplines, including physicians, nurses, and therapists. This collaborative model produces unbalanced Electronic Health Records (EHR), where extensive physician documentation can overshadow lower-frequency contributions from specialized services. Standard summarization approaches risk losing these nuanced signals by “pooling” heterogeneous text. To address this challenge, we propose a section-aware retrieval pipeline that categorizes sentence-level spans from multi-source clinical notes, preserving multidisciplinary structure for downstream summarization.

Clinical section header identification has evolved from rule-based matching [1] to context-aware supervised language models [2,3]. However, supervised approaches typically require extensive domain-specific annotation and often struggle to generalize to specialized clinical settings [4]. To address this limitation, we investigate whether large language models (LLMs) can transfer structured section representations across domains. Specifically, we evaluate two Llama-3 models [5] (8B and 70B), fine-tuned on adult ICU data, for cross-domain generalization to neonatal care. Our findings suggest a scale-dependent transfer pattern: while both models perform comparably in-domain, the quantized 70B model more effectively preserves structured section constraints under domain shift, achieving more balanced cross-domain performance compared to the smaller variant.

2 Methodology

2.1 Datasets and Clinical Section Schema

To investigate cross-domain transfer, we study sentence-level clinical section categorization using two datasets spanning adult ICU and neonatal intensive care (NICU), where each span is assigned to a predefined section/header category.

Source Domain (Adult ICU): We leveraged MedSecId [6], comprising 2,002 MIMIC-III notes with annotated headers. To promote cross-domain generalization, annotations were distilled into 17 high-prevalence categories (e.g., *History of Present Illness*, *Hospital Course*), yielding 1,562 notes split 80/10/10. This curation reduces overfitting to facility-specific headers and improves transferability. The data were formulated as a chat-based instruction-tuning task with 134,640 sentence-level spans (107,777 train, 13,876 val, 12,987 test), where each instance includes a system instruction, contextualized input sentence (with adjacent context), and ground-truth label.

Target Domain (NICU): To evaluate zero-shot transfer, we constructed a NICU pilot set of 227 sentence-level spans from three de-identified multidisciplinary “all-team” discharge summaries. These summaries were produced by a multidisciplinary clinical team (physicians, nurses, therapists) and used to derive section labels. Each sentence was reviewed by two clinicians using a consensus-derived schema, with disagreements resolved via adjudication; formal

agreement metrics (e.g., κ) were not computed given the pilot setting and iterative schema refinement. The schema includes 25 fine-grained categories spanning discharge sections (e.g., *Brief Admission Summary*), discipline summaries (*PT*, *OT*, *SLP*), care coordination (e.g., *Feeding-Nutrition*, *Family-Social*), and organ-system hospital course (e.g., *Respiratory*, *Cardiovascular*). This representation only partially overlaps with adult ICU headers and introduces greater granularity, providing a stringent test of cross-domain generalization. The study was approved by the University of Illinois Chicago IRB.

Task Example: Given a sentence-level span from a NICU discharge summary, e.g., “*Patient was intubated in delivery room and placed on high frequency jet ventilator.*,” the model aims to predict the corresponding section label (here, *Hospital Course Respiratory*). In contrast, “*27kcal Enficare ad lib feeding with slow flow nipple*” is assigned to *Feeding-Nutrition*. The task assigns each span to a clinical section category rather than generating text.

2.2 Model Architectures and Efficient Fine-Tuning

We fine-tuned Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct models using QLoRA [7], reducing model weight memory by approximately 75% relative to 16-bit storage. This enabled a full 70B model replica to fit within the 80GB memory budget of an A100 GPU, allowing data-parallel training across four GPUs. Both models were quantized to 4-bit Normal Float (NF4) precision, with LoRA adapters [8] ($r = 16, \alpha = 32$) attached to all linear projections, including MLP *gate*, *up*, and *down* layers, rather than restricting adaptation to attention modules. This substantially increased trainable parameter coverage and allowed modification of both attention patterns and feed-forward representations. Following standard scaling practice, we used a learning rate of $2e^{-4}$ for the 8B model and $1e^{-4}$ for the 70B model. Both models converged stably with minimal train/validation loss gaps (< 0.08), suggesting effective generalization without evidence of overfitting. Training used 1 epoch, effective batch size 16, bf16 precision, and LoRA dropout 0.05 for both models.

3 Results

3.1 In-Domain Performance (MedSecId)

Before evaluating cross-domain transfer, we first assessed the instruction-tuning pipeline on the source domain. Both architectures demonstrated strong alignment with the clinical section schema. As shown in Table 1, the 8B and 70B models achieved nearly identical weighted F1-scores of 0.94 on the held-out test set ($N = 12,987$ spans). Importantly, Macro F1-scores (0.920.93) closely matched the weighted metrics, indicating that performance was consistent across both high- and low-frequency categories rather than being driven primarily by prevalent classes. For example, the low-frequency section header *Review of Systems* ($N = 56$ spans; 0.43% of the test set) achieved F1-scores between 0.81 and 0.84, demonstrating meaningful generalization even under limited class support. The

close agreement between weighted and macro averages establishes a robust in-domain baseline. Notably, the larger 70B model did not exhibit a statistically meaningful advantage over the 8B variant under domain-specific supervision, suggesting that parameter scale alone does not substantially improve schema internalization in the source setting.

Table 1. In-Domain Evaluation on MedSecId Test Set ($N = 12,987$ spans)

Model	Macro Average			Weighted Average		
	Precision	Recall	F1	Precision	Recall	F1
Llama-3.1-8B	0.94	0.92	0.92	0.94	0.94	0.94
Llama-3.3-70B	0.94	0.92	0.93	0.94	0.94	0.94

3.2 Cross-Domain Transfer to NICU

Evaluation on the NICU Pilot Set ($N = 227$ spans) revealed a divergence in generalization capabilities (Table 2). Although both architectures achieved comparable in-domain performance (0.94 weighted F1), their ability to transfer section-label supervision to neonatal care differed. Llama-3.3-70B (SFT) achieved the strongest overall performance, reaching a weighted F1 of 0.60, a 4-point improvement over its base version. More notably, its Macro F1 improved from 0.52 to 0.59. This suggests that the 70B model may have sufficient parameter capacity to internalize the structured constraints of the clinical schema while maintaining the semantic flexibility required to recognize specialized, low-prevalence NICU section categories not observed during fine-tuning. In contrast, Llama-3.1-8B demonstrated only modest transfer gains. While weighted F1 improved slightly from 0.47 to 0.50, Macro F1 remained essentially unchanged (0.53 to 0.52). This near-parity suggests that fine-tuning primarily reinforced performance on higher-frequency classes without meaningfully improving minority-category generalization. Unlike the 70B model, the 8B architecture appears constrained in its ability to translate structured supervision into cross-domain improvements. Overall, these results suggest that larger-capacity models may be more effective at translating structured supervision into balanced cross-domain performance. Errors for the 70B model were most pronounced in Caregiver Education categories (e.g., SLP), with low recall despite high support, reflecting domain mismatch: these sections encode clinician instruction and intervention (e.g., “*Assistance with breast and bottle feeding with cueing and readiness*”) and were often misclassified into broader adult ICU categories such as *Hospital Course*, which lack explicit representations of caregiver-directed or feeding-related content.

4 Conclusion and Future Work

This study shows that supervised fine-tuning can adapt LLMs to fine-grained clinical section categorization, but cross-domain transfer from adult ICU docu-

Table 2. Cross-Domain Evaluation on NICU Pilot Set ($N = 227$ spans).

Model	Version	Macro Average			Weighted Average		
		Precision	Recall	F1	Precision	Recall	F1
Llama-3.1-8B	Base	0.54	0.62	0.53	0.53	0.50	0.47
	SFT	0.61	0.55	0.52	0.66	0.52	0.50
Llama-3.3-70B	Base	0.57	0.57	0.52	0.65	0.58	0.56
	SFT	0.64	0.69	0.59	0.70	0.62	0.60

mentation to multidisciplinary NICU summaries remains challenging. The 70B model showed numerically stronger transfer than the 8B model, but this pattern should be interpreted cautiously: the NICU Pilot Set contains only 227 sentence-level spans from three multidisciplinary discharge summaries, and a McNemar exact test was not statistically significant ($p = 0.24$). The study is also limited by the absence of non-LLM baselines such as ClinicalBERT.

Developing larger NICU evaluation sets is resource-intensive because section labels in “all-team” summaries require manual review of heterogeneous physician, nursing, and therapy documentation, where relevant information is distributed across discipline-specific notes rather than consolidated into standardized sections. Future work will expand the NICU set, add non-LLM baselines, and evaluate whether section classifiers can support downstream summarization by providing more targeted contextual grounding for LLM-based discharge summaries. Overall, our findings suggest that larger models may offer more robust cross-domain gains, but stronger evidence will require larger annotated datasets and broader baseline comparisons.

References

1. Denny, J.C., Miller, R.A., Johnson, K.B., Spickard, I.A.: Development and evaluation of a clinical note section header terminology. In: AMIA Annual Symposium Proceedings, p. 156 (2008)

2. Zhang, F., Laish, I., Benjamini, A., Feder, A.: Section classification in clinical notes with multi-task transformers. In: Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI), pp. 54–59 (2022)

3. Saleh, M., Baghdadi, S., Paquelet, S.: TocBERT: medical document structure extraction using bidirectional transformers (2024). [arXiv:2406.19526](#)

4. Karacan, B., Di Eugenio, B., Thornton, P.: Bridging the domain divide: supervised vs. zero-shot clinical section segmentation from MIMIC-III to obstetrics. In: Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pp. 2594–2607. European Language Resources Association (ELRA), (2026). <https://doi.org/10.63317/4ktoypuohtci>

5. Dubey, A., et al.: The Llama 3 herd of models (2024). [arXiv:2407.21783](#)

6. Landes, P., Patel, K., Huang, S.S., Webb, A., Di Eugenio, B., Caragea, C.: A new public corpus for clinical section identification: MedSecId. In: Proceedings of the 29th International Conference on Computational Linguistics (COLING), pp. 3709–3721 (2022)

7. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: efficient fine-tuning of quantized LLMs. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 36, pp. 10088–10115 (2023)
8. Hu, E.J., et al.: LoRA: low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)



Evaluating Tile-Level Feature Embeddings for Attention-Based MIL in Placental Histopathology

Saleh Abou-Alwan¹✉, Afsoon Khodaei¹, Dina El Demellawy²,
Shannon Bainbridge³, and Adrian D. C. Chan¹

¹ Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada
salehaboualwan@cmail.carleton.ca

² Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

³ Faculty of Health Sciences, University of Ottawa, Ottawa, ON, Canada

Abstract. Maternal Vascular Malperfusion (MVM) lesions are placental abnormalities associated with increased lifelong cardiovascular disease risk in mothers and are typically identified using high-resolution whole-slide images (WSIs). Manual annotation of MVM lesions is time-consuming and subjective, motivating weakly supervised approaches such as attention-based multiple instance learning (MIL) for WSI-level classification. However, the influence of feature extractor pretraining, particularly domain alignment and dataset scale on downstream MIL performance remains misunderstood. This study compares tile-level feature representations extracted from ImageNet-pretrained models (ResNet-18, ResNet-50, ViT-B, ViT-L, and DeiT-III) and pathology-specific foundation models (Prov-GigaPath, H-Optimus-1, Hibou-B, and Hibou-L) within a fixed attention-based MIL framework for MVM classification. Prov-GigaPath achieved the highest ROC-AUC (73.33%), followed by H-Optimus-1 (72.07%). Prov-GigaPath also produced higher accuracy, precision, and F1-score, while H-Optimus-1 demonstrated higher recall (sensitivity). Among ImageNet-pretrained models, ViT-L performed best (ROC-AUC 69.81%). Overall, this study demonstrates that most pathology-aligned foundation models produced more discriminative tile-level representations, resulting in improved downstream MIL classification performance compared to non-domain-specific counterparts.

1 Introduction

Cardiovascular disease (CVD) is a leading cause of mortality among women and has been associated with pregnancy complications such as preeclampsia and fetal growth restriction [1]. Maternal vascular malperfusion (MVM), a placental pathology that has been associated with these complications, has been identified as a potential indicator for screening of long-term cardiovascular risk [2]. Although MVM lesions could be identified using whole slide images (WSIs) consisting of placental tissue, manual assessment is time-consuming, laborious and prone to intra and inter-rater variability [3], highlighting the need for automating the detection of MVM lesions. Pretrained foundation models leverage large-scale data to extract transferable features for downstream tasks. The use of

an attention-based multiple instance learning (MIL) has been proven to be effective for classifying histopathological lesions such as MVM [4]. However, the impact of feature extractor pretraining regimen, especially with regards to dataset type, scale and model capacity, remains misunderstood. The objective of this research study is to provide a controlled comparison of CNN and ViT-based foundation models pretrained on natural images (e.g., ImageNet), compared to large-scale histopathology datasets used as feature extractors for placental MVM classification, using an attention-based MIL framework.

2 Materials and Methods

2.1 Dataset

The dataset used for this study consisted of 540 H&E-stained WSIs, acquired from the Children’s Hospital of Eastern Ontario (CHEO). WSIs were scanned at a 20× magnification (0.5 μm/pixel) and manually assessed by a perinatal pathologist screening for MVM lesions, including multiple MVM related pathological features. Each MVM lesion was assigned an integer score reflecting lesion presence (0: absent, 1: present), scores were aggregated across each WSI to obtain a cumulative MVM score. WSIs with cumulative scores >2 were labeled as MVM+ and WSIs with cumulative scores ≤2, were labeled as MVM−. The dataset used in this study consists of 447 MVM− cases representing 82.7% of the total dataset, and 93 MVM+ cases representing 17.2% of the total dataset (Table 1).

Table 1. Distribution of WSI by MVM Lesion Score

	MVM−		MVM+		
Score	0	2	3	4	5
# of Cases	321	126	72	21	0
% of Score	59.44%	23.33%	13.33%	3.88%	0%

2.2 Image Preprocessing and Tile Extraction

WSIs were down sampled from 0.5 μm/pixel to 2.5 μm/pixel. Tissue regions were segmented using Otsu thresholding [5] to exclude background, and Macenko stain normalization was applied to standardize color appearance prior to tile extraction.

2.3 Feature Extraction

Foundational models were used as feature extractors to extract tile-level features, which were then input into a fixed attention-based MIL framework for classification. This study used two sets of foundational models: 1) CNN (ResNet-18 [6] and ResNet-50 [6]) and

ViT (ViT-B [7], ViT-L [7], DeiT-III [8]) variants pretrained on a natural image dataset (ImageNet), and 2) ViT variants pretrained on pathology-specific datasets (Hibou-B [9], Hibou-L [9], Prov-GigaPath [10], and H-Optimus-1 [11]). Pathology-specific models were pretrained using self-supervised learning on a large scale histopathology specific dataset. For instance, Prov-GigaPath was trained on 171,000 WSIs, resulting in 1.3 billion image tiles, while the H-Optimus-1 model was trained on 1 million WSIs across diverse tissue types. Additionally, the Hibou models were trained on up to 1.2 billion tiles with different histopathology datasets. In contrast, ImageNet-pretrained models were trained on natural image datasets, significantly differing from histopathology images in both structure and texture.

2.4 Attention-Based Multiple Instance-Learning (MIL)

An attention-based multiple instance learning (MIL) framework was employed for WSI classification using tile-level representations, similar to [4]. For each WSI, extracted tile embeddings were grouped in a labeled bag, and each instance was transformed into a lower-dimensional latent representation using fully connected layers with ReLU activation and dropout. Attention scores were then computed for each instance and normalized via a SoftMax function to obtain attention weights, which were used to aggregate tile-level features into bag-level representation. The aggregated representation was subsequently passed through a linear classifier to produce final predictions.

2.5 Training and Evaluation

Class imbalance was addressed through augmentation of the minority (MVM+) class using rotation, horizontal and vertical flipping, and scaling. Models were trained using 10-fold stratified cross-validation, with an 80/20 train and validation split within each fold. Hyperparameters were selected based on validation ROC-AUC, and performance was evaluated using pooled predictions across folds.

3 Results and Discussion

Table 2 summarizes the overall classification performance metrics on the pooled results, illustrating the difference in performance between non-domain specific and domain specific foundational models.

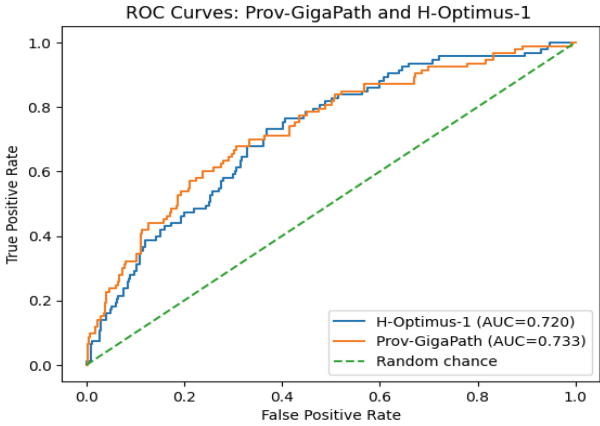


Fig. 1. ROC curves of H-Optimus-1 and Prov-GigaPath models

Prov-GigaPath features, the MIL classifier achieved higher ROC-AUC, accuracy, precision, and specificity, while H-Optimus-1 yielded higher recall and F1-score, indicating complementary strengths between these pathology-specific models. Although the Prov-GigaPath achieved the highest pooled ROC-AUC, a paired Wilcoxon signed-rank test on fold-level ROC-AUC showed no statistically significant difference between the two models ($p = 0.062$), with ROC curves (Fig. 1) further suggesting comparable performance. Overall, most pathology-aligned foundation models produced more effective tile-level representations than ImageNet-pretrained counterparts, supporting the importance of domain-specific pretraining. Consistent with prior work [12, 13], performance differences are influenced by a combination of pretraining domain, dataset scale, and model capacity. This is reflected in our results, where larger models (e.g., ResNet-50 vs. ResNet-18, ViT-L vs. ViT-B) generally outperformed smaller counterparts. However, Hibou-L underperformed Hibou-B despite having more parameters, indicating that increased model capacity alone doesn't guarantee improved performance. Despite being pretrained on large-scale pathology datasets, Hibou-B and Hibou-L consistently underperformed compared to Prov-GigaPath and H-Optimus-1. One possible explanation is the aggressive stain and geometric augmentations used during Hibou pretraining, which may suppress subtle color texture and morphological cues relevant to MVM lesions. In contrast, Prov-GigaPath and H-Optimus-1 employ more conservative augmentation strategies which preserve tissue morphology, potentially producing more discriminative features.

Despite promising results, several limitations should be acknowledged. The dataset was acquired a single institution and annotated by a single pathologist, which may limit generalizability. Additionally, the dataset is imbalanced, although augmentation was applied to the minority class, these transformations may not sufficiently increase sample diversity and could introduce redundancy. Furthermore, while models achieved reasonable ROC-AUC performance, sensitivity remained low, indicating that a substantial portion of MVM+ cases may be missed, which is a limitation for potential clinical screening applications.

Table 2. Overall Classification Performance Using Pooled Cross-Validation Predictions

Model	ROC-AUC	Accuracy	Precision	Sensitivity (Recall)	Specificity	F1-Score
ResNet-18	57.32%	75.32%	27.32%	18.11%	82.32%	23.21%
ResNet-50	67.49%	76.85%	20.10%	24.33%	87.97%	21.32%
ViT-B	64.82%	77.33%	45.83%	34.67%	86.21%	38.12%
ViT-L	69.81%	79.81%	35.59%	34.11%	89.52%	33.19%
DeIT-III	61.41%	76.26%	15.32%	10.18%	92.79%	12.01%
Hibou-B	68.49%	77.89%	48.24%	27.89%	88.44%	32.21%
Hibou-L	64.21%	80.19%	36.73%	16.12%	92.18%	21.10%
H-Optimus-1	72.01%	79.44%	39.12%	33.33%	89.03%	35.84%
Prov-GigaPath	73.32%	81.82%	45.45%	26.88%	93.26%	33.36%

4 Conclusion

This study employed an attention-based MIL framework for weakly supervised classification of MVM in placental WSIs. Pathology-aligned foundation models, particularly Prov-GigaPath and H-Optimus-1, produced more discriminative tile-level representations than ImageNet-pretrained counterparts, highlighting the importance of domain-specific pretraining, dataset scale, and model capacity for histopathological classification. Future work will aim to improve sensitivity and generalizability through different data balance strategies such as oversampling of MVM+ bags and class-weighted loss, as well as evaluation on multi-institutional datasets. Additional directions include exploring alternative weakly supervised frameworks (e.g., CLAM), multi-class severity predictions and attention map visualization to enhance interpretability.

Acknowledgements. We would like to acknowledge that this study was supported CIHR and NSERC agencies. Authors would also like to acknowledge that ChatGPT-5.2 was used for grammar enhancement.

References

1. Brosens, I., Pijnenborg, R., Vercruysse, L., Romero, R.: The ‘great obstetrical syndromes’ are associated with disorders of deep placentation. *Am. J. Obstet. Gynecol.* **204**(3), 193–201 (2011). <https://doi.org/10.1016/j.ajog.2010.08.009>
2. Honigberg, M.C., Zekavat, S., Aragam, K., et al.: Long-term cardiovascular risk in women with hypertension during pregnancy. *J. Am. Coll. Cardiol.* **74**(22), 2743–2754 (2019). <https://doi.org/10.1016/j.jacc.2019.09.052>
3. Liang, P.-I., Lin, W.-C., Wen, M.-C., et al.: Learning more from the inter-rater reliability of interstitial fibrosis assessment beyond just a statistic. *Sci. Rep.* **13**, 13260 (2023). <https://doi.org/10.1038/s41598-023-40221-6>

4. Khodaei, A., Chan, A.D.C., Ukwatta, E., Bainbridge, S.: Classification of whole slide images for the presence of maternal vascular malperfusion lesions using attention-based, weakly supervised deep learning. In: Proceedings of the IEEE International Instrumentation and Measurement Technology Conference (I2MTC), pp. 1–6 (2024). <https://doi.org/10.1109/I2MTC60896.2024.10560610>
5. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979). <https://doi.org/10.1109/TSMC.1979.4310076>
6. O’Shea, K., Nash, R.: An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*. (2015)
7. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021)
8. Touvron, H., Cord, M., Jégou, H.: DeiT III: revenge of the ViT. In: *Computer Vision – ECCV 2022*, Tel Aviv, Israel, pp. 516–533 (2022)
9. Nechaev, D., Pchelnikov, A., Ivanova, E.: Hibou: a family of foundational vision transformers for pathology. *arXiv preprint arXiv:2406.05074*. (2024)
10. Xu, H., Usuyama, N., Bagga, J., et al.: A whole-slide foundation model for digital pathology from real-world data. *Nature*. **630**, 181–188 (2024). <https://doi.org/10.1038/s41586-024-07441-w>
11. Biopitimus: H-Optimus-1: A Foundation Model for Histopathology. [Online]. (2024)
12. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 12104–12113 (2022)
13. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3344–3354 (2023)



Evaluating a Locally Deployed 20-Billion Parameter Large Language Model for Automated Abstract Screening in Systematic Reviews

Paulo Henrique Moreira Melo¹(✉), Elena Guadagno², and Dan Poenaru^{1,2}

¹ Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada
paulo.moreiramelo@mail.mcgill.ca

² Harvey E. Beardmore Division of Pediatric Surgery, The Montreal Children's Hospital,
McGill University Health Centre, Montreal, QC, Canada

Abstract. Systematic reviews require extensive manual abstract screening. Large language models (LLMs) can automate this, but cloud-based solutions raise privacy and reproducibility concerns. As a feasibility study, we evaluated a locally deployed 20B-parameter LLM (`gpt-oss:20b`) using sensitivity-enhanced prompting across three systematic reviews (16,646 abstracts). After blind expert adjudication of all discordant cases, sensitivity was 100% (SR1, 95% CI 89.6–100%), 95.7% (SR2, 95% CI 85.8–98.8%), and 85.7% (SR3, 95% CI 76.2–91.8%). The LLM identified 11 human screening errors and screened $4.7\times$ faster. Performance was strongest for technology-focused topics with objective criteria. We recommend deploying LLMs as second screeners alongside human reviewers.

Keywords: systematic review · large language model · abstract screening · artificial intelligence · local deployment

1 Introduction

Systematic reviews (SRs) require at least two independent reviewers to screen all abstracts, with a third adjudicating disagreements [1, 2]. With search strategies often identifying thousands of abstracts, this process is profoundly resource-intensive and increasingly unsustainable given the exponential growth of research output [3].

LLMs have shown promise for automating SR screening [4–6, 19]. Sanghera et al. found LLM ensembles achieved superior sensitivity versus human reviewers across 23 Cochrane reviews [7, 13]. Performance depends on criteria clarity [8], and multiple studies report accuracy at or above 90% [9, 14, 16]. Prompt design is critical: instructing models to include uncertain cases can elevate sensitivity from 62% to 95% [10, 15, 17].

However, cloud-based APIs raise concerns about data privacy, cost, and reproducibility [8, 11, 20]. Local deployment of open-source models addresses these barriers [12]. As a feasibility study, we evaluate a locally deployed 20B-parameter LLM across three SRs using sensitivity-enhanced prompting, with blind expert adjudication of all discordant cases.

Table 1. Initial LLM screening performance (before expert adjudication).

Review	n	Human Incl.	AI Incl.	FN	FP
SR1: AI in Ped. Surgery	3,350	31	102	1	72
SR2: LLMs in EHR	4,326	49	534	5	486
SR3: Parental Stress	8,970	79	66	14	1
Total	16,646	159	702	20	559

Table 2. Final LLM screening performance after expert adjudication. Sensitivity reported with Wilson 95% confidence intervals.

Review	Rel.	Sens. (95% CI)	Spec.	FN	FP	Hum. Err
SR1: AI in Ped. Surgery	33	100% (89.6–100)	97.9%	0	69	3
SR2: LLMs in EHR	47	95.7% (85.8–98.8)	88.7%	2	485	4
SR3: Parental Stress	77	85.7% (76.2–91.8)	99.99%	11	0	4
Overall	157	91.7% (86.3–95.1)	96.7%	13	554	11

2 Methods

We deployed `gpt-oss:20b` (OpenAI, August 2025) locally via Ollama [12] and screened abstracts from three SRs: SR1—AI in pediatric surgical pathology (n = 3,350); SR2—LLMs in electronic health records (n = 4,326); SR3—parental stress/caregiver burden in surgically treated children (n = 8,970). Total: 16,646 abstracts.

Each abstract was processed with its title and the SR’s inclusion/exclusion criteria. The model returned a JSON object with a binary decision and reasoning. The prompt included the instruction: “When in doubt, **INCLUDE** the abstract for further review,” prioritizing sensitivity since false negatives are irreversible whereas false positives proceed to full-text screening [7, 10]. The full prompt template is available in the code repository.

All discordant cases underwent blind expert adjudication by a senior researcher who was unaware which decision came from the LLM or human. Performance metrics included sensitivity, specificity, false negatives (FN), false positives (FP), and human errors. We report Wilson 95% confidence intervals (CIs) for sensitivity.

3 Results

Table 1 shows pre-adjudication results. The LLM had 20 apparent FN and 559 FP.

Expert adjudication revised these findings substantially (Table 2). In SR1, the single apparent FN was adjudicated as a correct exclusion, yielding 100% sensitivity; 3 FPs were relevant studies humans had missed. In SR2, 3/5 FNs were correct exclusions (2

true FN); 1 FP was a correct inclusion. In SR3, 3/14 FNs were correct exclusions (11 true FN); the 1 FP was a correct inclusion. Across all reviews, expert adjudication identified 11 human screening errors that the LLM had correctly classified.

The 11 SR3 misses were predominantly qualitative or mixed-methods studies in which parental stress was reported through subjective narrative descriptors rather than named instruments or explicit outcome statements. Several involved chronic or non-surgical pediatric conditions in which caregiver burden was a secondary theme, making relevance contingent on nuanced reading of context rather than keyword cues. This pattern is consistent with the lower sensitivity observed for psychosocial outcomes with subjective inclusion criteria.

The LLM screened SR1 in 5.58 h versus 26 h for human reviewers ($4.7\times$ faster; ~ 6 s/abstract).

4 Discussion and Conclusion

The locally deployed LLM achieved high sensitivity for technology-focused SRs (100%, 95.7%) but lower for psychosocial outcomes (85.7%), consistent with findings that objective criteria yield better LLM performance [8, 14]. Bidirectional errors—11 human misses caught by the LLM, 13 LLM misses caught by humans—demonstrate complementarity, supporting hybrid workflows over full automation [13, 16]. Expert adjudication was essential; without it, human decisions would have been treated as ground truth, distorting the evaluation. Local deployment via Ollama addressed privacy and reproducibility concerns while delivering substantial time savings [18, 19].

As a feasibility study, this work has several limitations. Performance was domain-dependent and outputs were binary without confidence scores. Evaluation was restricted to abstract-level screening at a single institution and to a single model family, so prompt behavior and absolute performance may not generalize to other open-weights LLMs. We did not benchmark against existing screening tools or alternative LLMs; comparative evaluation across models and against established tools is needed before claims of broader practical guidance can be made. Confidence intervals on sensitivity are wide for SR1 and SR2 given the small number of relevant abstracts. The 13 missed studies reinforce the necessity of human oversight. We recommend deploying LLMs as second screeners with expert adjudication of disagreements. Future work should develop confidence-calibrated outputs, conduct cross-model and cross-tool benchmarking, and establish standardized evaluation across domains [19, 21].

Acknowledgments. Declaration of AI Assistance The author(s) used Claude (Anthropic) for assistance with manuscript formatting and for supporting the development of the screening pipeline code. After using this technology, the first author reviewed and edited all outputs and takes full responsibility for the contents of the manuscript.

Data and Code Code and full prompt template: https://github.com/PauloHenriqueMelo/Systematic_Review_Screening_LLM. Data available on request.

References

1. Chandler, J., et al.: *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane (2022)
2. Centre for Reviews and Dissemination: *Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care*. University of York, York (2009)
3. Bastian, H., Glasziou, P., Chalmers, I.: Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med.* **7**(9), e1000326 (2010)
4. Murton, M., et al.: Using large language models for data extraction in systematic reviews. *Cochrane Evid. Synth. Methods.* **3**(6), e70058 (2025)
5. Bayani, A., et al.: Automating scoping reviews with open-source large language models. *Int. J. Med. Inform.* **204**, 106048 (2025)
6. Khraisha, Q., et al.: Can large language models replace humans in systematic reviews? *Res. Synth. Methods.* **15**(4), 616–626 (2024)
7. Sanghera, R., et al.: High-performance automated abstract screening with large language model ensembles. *J. Am. Med. Inform. Assoc.* **32**(5), 893–904 (2025)
8. Delgado-Chaves, F.M., et al.: Transforming literature screening: the emerging role of large language models in systematic reviews. *Proc. Natl. Acad. Sci. USA.* **122**, e2411962122 (2025)
9. Guo, W., et al.: Evaluating the effectiveness of large language models in abstract screening. *Syst. Rev.* **13**(1), 219 (2024)
10. Gargari, O.K., et al.: Use of ChatGPT for performing medical literature screening: prompt engineering for increased inclusivity. *BMJ Evid. Based Med.* **29**(1), 69–76 (2024)
11. Wang, Y., et al.: Automated paper screening for clinical reviews using large language models. *J. Med. Internet Res.* **26**, e48996 (2024)
12. Ollama: Run Large Language Models Locally. <https://ollama.ai>. (2025)
13. Oami, T., Okada, Y., Nakada, T.A.: Performance of a large language model in screening citations. *JAMA Netw. Open.* **7**(7), e2420496 (2024)
14. Tran, V.T., et al.: Sensitivity and specificity of using GPT-3.5 turbo models for title and abstract screening in systematic reviews. *Ann. Intern. Med.* **177**(6), 791–799 (2024)
15. Matsui, K., et al.: Human-comparable sensitivity of large language models in identifying eligible studies through title and abstract screening. *J. Med. Internet Res.* **26**, e52758 (2024)
16. Issaiy, M., et al.: Methodological insights into ChatGPT's screening performance in systematic reviews. *BMC Med. Res. Methodol.* **24**(1), 78 (2024)
17. Syriani, E., David, I., Kumar, G.: Screening articles for systematic reviews with ChatGPT. *J. Comput. Lang.* **80**, 101287 (2024)
18. van de Schoot, R., et al.: An open source machine learning framework for efficient and transparent systematic reviews. *Nat. Mach. Intell.* **3**, 125–133 (2021)
19. Lieberum, T., et al.: The emergence of large language models as a tool in literature reviews. *J. Am. Med. Inform. Assoc.* **32**(6), 1071–1081 (2025)
20. Oami, T., Okada, Y., Nakada, T.A.: Optimal large language models to screen citations for systematic reviews. *Res. Synth. Methods.* **16**, 859–875 (2025)
21. Akinseloyin, T., et al.: A question-answering framework for automated abstract screening using large language models. *J. Am. Med. Inform. Assoc.* **31**(9), 2086–2099 (2024)
22. Dennstädt, F., et al.: Title and abstract screening for literature reviews using large language models. *Syst. Rev.* **13**(1), 188 (2024)



Contextualized Interpretation of Machine-Learning Predictions in Rare Disease Omics: Integrating SHAP, Biomedical Knowledge, and Language Models

Daniel Daza^{1,3(✉)}, Yorrick Jaspers², Alberto Bernardi³, Luca Costabello³,
Christophe Gueret³, Michael Cochez^{1,4}, Marc Engelen², Stephan Kemp²,
and Martijn Schut²

¹ Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

² Amsterdam University Medical Centers, Amsterdam, The Netherlands

³ Accenture Labs, Dublin, Ireland
d.dazacruz@vu.nl

⁴ ELLIS Institute Finland and Åbo Akademi University, Turku, Finland

Abstract. Machine-learning models are increasingly applied in preclinical research, yet it remains difficult to interpret how input features influence predictions. This is especially true for rare diseases with complex biological mechanisms and when the analyzed data are high-dimensional. Several methods for interpreting the influence of variables in prediction models provide only feature-level insights and are limited by insufficient integration with broader biomedical knowledge. To address these limitations, we present SHAP-RAG: a framework for contextualized interpretation of preclinical prediction models. The framework augments SHAP-based feature attribution with complementary contextual information, including relationships between influential features, integration of external knowledge bases and the scientific literature, and a natural-language summary generated by a large language model to synthesize these elements. We implement the framework in a tool and evaluate it for understanding the prediction of adrenal insufficiency in patients with the rare disease of adrenoleukodystrophy from lipidomic data. Preliminary qualitative feedback from clinicians indicates that the framework supports interpretability with the integration of external knowledge, and generated summaries contribute to a deeper understanding of model predictions.

Keywords: Explainable AI · preclinical machine learning · rare diseases

1 Introduction

Machine-learning (ML) models are increasingly used in biomedical research, but interpreting their predictions remains difficult, particularly in high-dimensional

omics settings [1, 10]. In rare diseases, this challenge is amplified by limited sample sizes and complex biology [8]. Feature-attribution methods such as SHAP [5] can identify variables that drive predictions, yet isolated importance scores rarely provide enough context for biological interpretation or hypothesis generation [7].

We address this gap with SHAP-RAG: a systems-oriented framework for *contextualized interpretation* of rare-disease omics models. We combine feature attribution in machine learning models with three kinds of context that experts often need in practice: computational context about how features act together in the model, biochemical context from curated resources, and scholarly context from prior studies. This systems contribution matters in rare-disease omics because experts need more than a ranked feature list: they need a structured way to connect model signals to known biology and existing evidence, and to inspect whether apparently important features form coherent biochemical patterns rather than isolated artifacts. We instantiate the framework for lipidomic prediction of adrenal insufficiency in adrenoleukodystrophy (ALD) and evaluate it as a proof of concept through a use case and a qualitative expert assessment.

2 SHAP-RAG

Inspired by retrieval-augmented generation (RAG) [4], we investigate systems for explaining model predictions using natural language grounded in information retrieved from reliable sources. The framework takes a trained clinical prediction model as its starting point, and its aim is to support clinicians and disease experts in constructing an integrated, knowledge-rich view of the model features identified as influential for prediction. This is achieved through three stages of context enrichment detailed next.

Computational Context. We compute instance-level SHAP values and rank features by mean absolute SHAP to identify the most influential signals. For a selected feature, we then compute Spearman correlations between its SHAP values and those of the remaining features, highlighting groups of variables whose contributions co-vary across samples.

Knowledge Context. We enrich influential features with external knowledge. In the lipidomics use case, features are mapped to RefMet-style lipid identities, after which the framework retrieves biochemical annotations and linked studies from Metabolomics Workbench [9] and pathway information from KEGG [3]. The resulting interface combines ranked features, per-sample SHAP behavior, top correlated lipids, biochemical annotations, and retrieved evidence in a single view.

Scholarly Context. To complement curated database knowledge, the framework retrieves scientific articles that mention the selected features. These references provide contextual evidence from prior experimental and clinical studies, indicating the biological processes, disease contexts, or experimental settings in which the features have been previously investigated.

In the last step the framework feeds the collected context for each selected feature into a large language model (LLM). The prompt instructs the model to generate a cautious, structured summary covering: (i) the feature type and likely biological role; (ii) how its SHAP importance and SHAP-level correlations relate to domain questions; and (iii) how pathways and literature contextualize potential mechanisms. The LLM synthesizes only the provided information, without introducing new hypotheses. The resulting summary links model behavior to plausible biological interpretations in a form accessible to clinicians and disease experts. The full prompt is provided in the supplementary material, and an open source implementation of our complete framework is available at <https://github.com/dfdazac/shaprag>. We deploy it as a web-based, user-facing application that presents the contextualized information generated at each step of the pipeline. An excerpt of this interface is shown in Fig. 1.

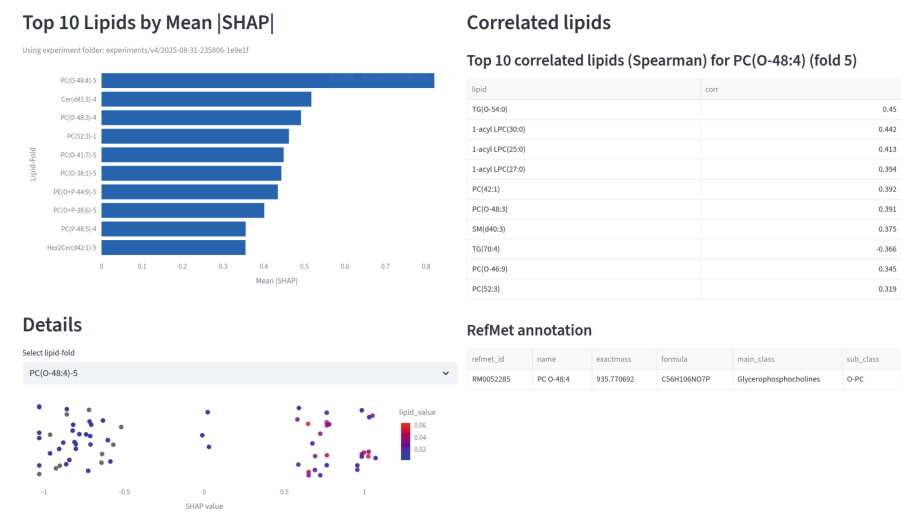


Fig. 1. Interface excerpt showing ranked SHAP signals, detailed inspection of a selected lipid-fold, correlated lipids, and biochemical context.

3 ALD Use Case and Expert Assessment

ALD is a rare disorder with disrupted lipid metabolism and clinically relevant adrenal insufficiency [6]. We apply the framework to an XGBoost model trained

for predicting the outcome of adrenal insufficiency, based on lipidomic data from prior work on ALD biomarkers [2]. The predictive pipeline uses lipid features and age, nested cross-validation, and SHAP export for downstream interpretation; in this use case, the final model achieves a 5-fold ROC AUC of 0.78. In the supplementary material, we provide the full model-development details, including feature selection, tuning, and fold-wise performance.

We conducted a qualitative assessment of the proposed framework by a clinician with experience in the research and treatment of adrenoleukodystrophy. The expert emphasized the utility of the framework’s visualization tools in interpreting model predictions for ALD from a clinical perspective. These tools provide an intuitive overview of the associations between various lipid species and clinical outcomes, allowing clinicians to quickly grasp which lipid groups are most strongly linked to disease progression and model outputs. This “feel” for the data is particularly valuable in clinical settings, where understanding the biological relevance of model predictions is essential. Importantly, the lipid patterns and pathway associations highlighted by the system were found to be in alignment with established biological knowledge of ALD and adrenal insufficiency [2].

The framework effectively identified very long-chain fatty acid-containing lipid species and their associated pathways as being relevant both to the diagnosis of ALD and to clinical features such as the development of adrenal failure. This concordance with current understanding reinforces the credibility of the model’s outputs. However, the expert noted that, at this stage, the framework has not revealed novel relationships or insights that would not have been apparent from standard statistical analyses. While the associations identified are biologically plausible, the system has yet to provide unexpected findings that could advance understanding of ALD pathophysiology.

The main limitation is thus the lack of new insights, although the framework is considered promising as a proof of principle. The expert suggested that the framework could benefit from the integration of additional datasets, including genomics and diverse outcome measures. Such expansion may enable the discovery of previously unrecognized links and provide deeper clues into the biochemical pathways underlying ALD. Overall, while the current system is useful for visualizing and confirming known associations, its potential for generating novel insights will depend on further development and broader data integration.

4 Discussion and Conclusion

This work contributes an integrative framework and interactive tool for contextualizing SHAP-based explanations in rare-disease omics. Its main value is practical: it helps experts move from isolated feature rankings to a structured view that combines model behavior, biochemical metadata, literature-linked evidence, and optional narrative synthesis. In settings such as rare-disease biomarker research, this kind of evidence organization can be valuable even when the methodological novelty lies in integration rather than in a new attribution algorithm.

The current evidence is limited to one disease, one omics modality, one task, and one qualitative expert assessment. More broadly, the computational layer of

the framework is modality-agnostic, but the knowledge layer is domain-specific and would require different mappings and databases for transcriptomics, proteomics, or poorly characterized diseases, where external knowledge may be sparser and feature normalization may be less straightforward. Formal evaluation of generalization, usefulness, and LLM reliability remains future work, together with comparisons against baseline explanation workflows and broader user studies.

References

1. Adlung, L., Cohen, Y., Mor, U., Elinav, E.: Machine learning in clinical decision making. *Med* **2**(6), 642–665 (2021)
2. Jaspers, Y.R.J., et al.: Lipidomic biomarkers in plasma correlate with disease severity in adrenoleukodystrophy. *Commun. Med.* **4**(1), 1–16 (2024)
3. Kanehisa, M., Goto, S.: KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**(1), 27–30 (2000)
4. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural. Inf. Process. Syst.* **33**, 9459–9474 (2020)
5. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: Guyon, I., Luxburg, U.v., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 4–9 December 2017, Long Beach, CA, USA, pp. 4765–4774 (2017). <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
6. Moser, H.W., Mahmood, A., Raymond, G.V.: X-linked adrenoleukodystrophy. *Nat. Clin. Pract. Neurol.* **3**(3), 140–151 (2007)
7. Saraswat, D., et al.: Explainable AI for Healthcare 5.0: Opportunities and Challenges. *IEEE Access* **10**, 84486–84517 (2022). <https://doi.org/10.1109/ACCESS.2022.3197671>. <https://ieeexplore.ieee.org/abstract/document/9852458>
8. Schieppati, A., Henter, J.I., Daina, E., Aperia, A.: Why rare diseases are an important medical and social issue. *The Lancet* **371**(9629), 2039–2041 (2008)
9. Sud, M., et al.: Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **44**(D1), D463–D470 (2016)
10. Yang, W., et al.: Survey on explainable AI: from approaches, limitations and applications aspects. *Human-Centric Intell. Syst.* **3**(3), 161–188 (2023)



Annotating Motivational Interviewing Transcripts Using Large Language Models

Dinesh Rivankar¹, Tanmay Tulsidas Verlekar^{1(✉)}, and Abhijit Nadkarni^{2,3}

¹ Department of Computer Science and Information Systems, BITS Pilani K. K. Birla, Goa Campus, Goa 403726, India

tanmayv@goa.bits-pilani.ac.in

² Centre for Global Mental Health, Department of Population Health, London School of Hygiene and Tropical Medicine, London, UK

³ Addictions and Related Research Group, Sangath, Goa, India

Abstract. Annotating motivational interviewing (MI) transcripts by experts can be a time and resource-intensive task. The use of LLMs allows scalable, automated labelling of MI transcripts and can serve as an important tool for assessing the quality of intervention. An advantage of using LLMs over traditional deep learning models is their ability to perform inference without training, using only a few representative examples. This paper explores multiple LLMs for classifying clients' utterances as change, sustain, or neutral talk. The paper follows the Motivational Interviewing Skills Code v1, a coding system developed for motivational interviewing to create an instruction set for LLMs. It then evaluates various prompting strategies, including zero-shot, few-shot prompting and retrieval-augmented generation (RAG). The results highlight the significance of selecting few-shots with matching sub-types for successful classification. The proposal also highlights the need to effectively capture various contexts, including target behaviour, the counsellor's speech that elicits the client's utterance, and talk sub-types. The work is evaluated on the publicly available AnnoMI dataset. Results report that the GPT-OSS-120B with RAG achieves an improvement of 0.20 in the macro F1 score over the state-of-the-art. The results also show that RAG retrieval is equivalent to that of human experts.

Keywords: LLM · Text Annotation · Motivational Interviewing

1 Introduction

MI focuses on goal-oriented communication between a counsellor and a “client” (the person receiving MI) to strengthen the client's motivation and commitment to change their behaviour [7]. Clients' speech can be broadly divided into change talk, indicating a client's willingness to change; sustain talk, indicating a client's resistance to change; and neutral talk, which includes responses that are neither change talk nor sustain talk. Counsellors often use the frequency of change, sustain and neutral talk to estimate clinical outcomes [12].

Recent advancements in deep learning (DL) and natural language processing have enabled the automatic annotation of MI transcripts [13]. The work presented in [13] fine-tunes the Bidirectional Encoder Representations from Transformers (BERT) model to perform classification of input text sequences as change, sustain and neutral talk. Work such as [8] classifies input utterances as change talk or non-change talk using a recurrent neural network (RNN) trained on private datasets. While others focus on estimating therapist empathy using DL models such as an RNN [3]. Advancements in the use of DL models for detecting change, sustain, and neutral talk have been limited primarily by a lack of training data. To our best knowledge, the Anno-MI [14] is the only publicly available dataset containing labelled MI transcripts.

This paper explores the use of large language models (LLMs) to classify client utterances in MI as change, sustain, or neutral using prompts. Its contribution includes: Comparing LLMs, such as QwQ 32B, GPT OSS 120B, Gemini 2.0 Flash and Gemma 3 27B to assess their ability to detect change, sustain and neutral talk; Investigating the influence of adding the counsellor's speech as additional context for the LLM; Investigating the influence of prompting strategies such as zero-shot, few-shot and retrieval augmented generation (RAG) on the performance of the LLMs; Investigating the ability of RAG to retrieve relevant few-shot examples, with matching subtypes for the LLM. The paper reports its results by evaluating the proposal on the Anno-MI [14] dataset. The results motivate the use of LLMs for annotating the language of change in MI.

2 Proposal

The proposal assigns the LLM a persona of an MI expert and provides it with an instruction set encoding the Motivational Interviewing Skills Code (MISC) v1 [6] for a target behaviour. For every query, the prompt [11] first restates explicit definitions of change talk, sustain talk, and neutral talk in terms of movement toward or away from the target behaviour, and then presents the current client utterance. The LLM is instructed to output exactly one label: change, sustain, or neutral. This zero-shot prompting strategy uses only these definitions and the current conversation snippet, without any labelled examples, and relies on the LLM's knowledge to assign the label.

Classifying client utterances in MI requires understanding and associating these utterances with the defined target behaviour, such as reducing consumption of psychoactive substances. Change talk can occur through problem recognition, expressing feelings of concern, expressing the desire or intent to change, or expressing optimism about the possibility of change, which are its sub-types. Sustain talk represents client responses that favour maintaining the status quo for the identified target behaviour. Its subtypes include arguing with the therapist, interrupting the therapist, denying responsibility, ignoring the therapist's direction, and changing the subject. A client's utterance not associated with the target behaviour is neutral talk. Expert human annotators traditionally label by meticulously reviewing session transcripts. There is a high degree of subjectivity

in the tasks, which require experts to review utterances in the context of the target behaviour and the counsellor’s speech that induced it, making the tasks challenging. The proposal next considers the few-shot prompting to incorporate the context of the target behaviour and the counsellor’s speech in the client’s utterances. It includes examples of the client’s utterances and the counsellor’s speech pair, along with the corresponding label and the target behaviour in the prompt. The LLM uses these examples to understand the association of the utterances between the target behaviour and the counsellor’s speech to improve its response. To further improve the LLM’s performance, the proposal suggests selecting examples with a match subtype. This task can be performed using human experts. To automate this task, the proposal uses retrieval augmented generation (RAG). RAG retrieves few-shot examples with matching subtypes using cosine similarity between the client’s speech embedding and the embeddings of examples in an external knowledge base, which contain labelled utterances, the corresponding counsellor’s speech and the relevant target behaviour. The embeddings are generated using the Gemini-embedding-001 [5]. It is a multi-lingual text embedding model that maps input text to a 3072-dimensional vector. The proposal retrieves the labelled utterances and the counsellor’s speech pairs with the highest similarity scores as the few-shot examples for the LLM.

The proposal considers Gemini 2.0 Flash [4] and Gemma 3 27B [2], which focus on a broad number of generative tasks, such as comprehension and summarisation. Being trained on large, diverse text corpora, they are capable of processing natural language effectively. They, however, struggle with complex, multi-step logical reasoning, required for the client utterance classification. To ensure accurate association with the target behaviour, the counsellor’s speech and subtype, the paper also considers LLMs such as QwQ 32B [10] and GPT OSS [9]. It generates a chain of thought, which is an explicit, step-by-step breakdown of the complex problem. It allows these LLMs to possess critical analysis and complex decision-making capabilities. Their prompts are available in [11].

3 Evaluation

The proposal is evaluated on the Anno-MI [14] dataset. It contains 133 MI demonstration transcripts. Each transcript is associated with a target behaviour. The transcripts are split into the client’s utterances and the counsellor’s (therapist’s) speech, with the client’s utterances labelled as change, sustain, or neutral. The splitting of transcripts results in 6725 labelled individual utterances, with 4217 labelled as neutral, 1668 as change, and 840 as sustain. The evaluations are performed according to the experimental setup described in [14], which involves 5-fold cross-validation. Since the proposal doesn’t involve training any models, the training folds are used to select the few-shot examples and create the external knowledge base for the RAG. The temperature is set to 0 for all LLM models, with the remaining hyperparameters set to their default values. Results are reported in Table 1 in terms of macro-F1 score. The class imbalance in the dataset motivates selecting the macro F1 score for evaluation.

Table 1. F1 scores of the proposal with respect to different LLMs and prompts.

Prompting strategy	QwQ 32B F1 score				GPT-OSS-120B F1 score				Gemma 3 27B IT F1 score				Gemini 2.0 Flash F1 score			
	Change	Neutral	Sustain	Macro	Change	Neutral	Sustain	Macro	Change	Neutral	Sustain	Macro	Change	Neutral	Sustain	Macro
Zero-shot (without counsellor speech)	0.64	0.79	0.59	0.67	0.63	0.81	0.58	0.67	0.61	0.74	0.58	0.64	0.60	0.79	0.64	0.68
Zero-shot	0.57	0.79	0.64	0.67	0.65	0.81	0.66	0.71	0.65	0.76	0.62	0.68	0.62	0.81	0.66	0.69
Few-shot(random)	0.47	0.80	0.57	0.61	0.48	0.82	0.57	0.62	0.49	0.79	0.54	0.61	0.58	0.78	0.59	0.65
Few-shot(using human experts)	0.48	0.79	0.55	0.61	0.52	0.81	0.58	0.64	0.57	0.77	0.62	0.65	0.58	0.79	0.60	0.66
RAG(Proposed)	0.59	0.80	0.64	0.68	0.70	0.83	0.71	0.75	0.65	0.81	0.61	0.69	0.61	0.79	0.61	0.67

The results in Table 1 indicate that when using only the client’s utterance for classification, the macro F1 remains between 0.64 and 0.68 across the considered LLMs. Including the counsellor’s speech as the cause for the client’s utterance slightly improves the LLMs’ performance. This minor gain indicates that simply appending the counsellor’s speech is insufficient for LLMs to classify the client’s behaviour. A significant improvement in performance is reported across all LLMs when prompted with few-shot examples. When three examples per class are randomly sampled from the training set, they allow the LLMs to form associations between the target behaviour, the client’s utterances and the counsellor’s speech. However, randomly selecting examples can lead to a mismatch between subtypes. To validate this claim, a second few-shot prompt is used, in which a human expert handpicks the examples for the classification task, ensuring the subtype matches the client’s utterance. The prompt improved the macro F1 score of QwQ 32B to 0.68, GPT-OSS-120B to 0.67, Gemma 3 27B IT to 0.65 and Gemini 2.0 Flash to 0.66. To automate the example selection process, the proposal uses RAG, resulting in the final macro F1 scores reported in Table 1, with QwQ 32B achieving a macro F1 of 0.68, GPT-OSS-120B 0.75, Gemma 3 27B IT 0.69 and Gemini 2.0 Flash 0.67. For GPT-OSS-120B, this corresponds to an improvement of 0.07 over the human expert-selected few-shot. In the case of Gemini 2.0 Flash, the expert selected few-shot and RAG have similar performance with a macro F1 of 0.66 and 0.67, respectively. The experts were tasked to retrieve the few-shot based on individual utterances. They expressed low confidence in matching the subtype without the conversation history. Thus, expert-selected few-shot had higher recall for neutral talk. The proposed RAG has better recall for change and sustain talk -see Table 1. Thus, it can be concluded that the RAG is better equipped to retrieve few-shot with matching subtype for change and sustain talk, which is more significant for MI applications.

The results in Table 1 also show inconsistent performance across all LLMs, with Gemma 3 27B having the worst macro F1 score of 0.69 with RAG. It can be attributed to the fact that Gemma 3 27B is an open-source model designed to run efficiently on various resource-constrained devices. It trades accuracy for efficiency, customisation, and local deployment. In contrast, GPT-OSS-120B is a high-capacity proprietary model that benefits most from RAG, reaching a macro

F1 of 0.75. Gemini 2.0 Flash and QwQ 32B fall between these two LLMs. The results reported in Table 1 are, however, significantly better than the state-of-the-art BERT model, presented in [1, 14], which obtains F1 scores of 0.51, 0.74, and 0.39 for change, neutral, and sustain talk, respectively, and a macro F1 of 0.55. The fine-tuning forces the BERT model to focus on the overrepresented neutral talk in the unbalanced AnnoMI dataset. In contrast, the proposal reports a more balanced result across the three classes. To obtain equivalent results, the work in [1] uses a multimodal model operating on text, audio, and faces.

4 Conclusions

This paper explores the use of LLMs for annotating MI transcripts. The LLM models classify a client’s utterance as change, sustain, or neutral talk. It achieves this by associating the client’s utterance with the target behaviour using the instruction set. It then uses RAG to select examples that are similar to the client’s utterance with respect to its subtype. The result is an LLM capable of automatically annotating MI transcripts, achieving a macro F1 score of 0.75 on the Anno-MI dataset. While the results are promising, they indicate that the LLM struggles to match the subtype given the limited context of an individual client’s utterance. Thus, future work will focus on explicitly building subtype context for LLMs using conversational history.

References

1. Galland, L., Pelachaud, C., Pecune, F.: Seeing and hearing what has not been said: a multimodal client behavior classifier in motivational interviewing with interpretable fusion. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), pp. 1–9. IEEE (2024)
2. Gemma Team: Gemma 3 Technical report. [arXiv:2503.19786](https://arxiv.org/abs/2503.19786) (2025)
3. Gibson, J., et al.: A deep learning approach to modeling empathy in addiction counseling. *Commitment* **111**(2016), 21 (2016)
4. Google: Gemini 2.0 flash [large language model]. [arXiv:2507.06261](https://arxiv.org/abs/2507.06261) (2025)
5. Google DeepMind: Gemini embedding model: Gemini-embedding-001 (2025)
6. Miller, W.R., Moyers, T.B., Ernst, D., Amrhein, P.: Manual for the motivational interviewing skill code (MISC). Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico (2003)
7. Miller, W.R., Rollnick, S.: Motivational interviewing: Helping people change. Guilford press (2012)
8. Nakano, Y.I., Hirose, E., Sakato, T., Okada, S., Martin, J.C.: Detecting change talk in motivational interviewing using verbal and facial information. In: Proceedings of the 2022 International Conference on Multimodal Interaction, pp. 5–14 (2022)
9. OpenAI: GPT-OSS-120b & GPT-OSS-20b model card. [arXiv:2508.10925](https://arxiv.org/abs/2508.10925) (2025)
10. Qwen Team: Qwen3 Technical report. [arXiv:2505.09388](https://arxiv.org/abs/2505.09388) (2025)
11. Rivankar, D.: <https://github.com/dineshrivankar/MI-transcript-annotation>
12. Rubak, S., Sandbæk, A., Lauritzen, T., Christensen, B.: Motivational interviewing: a systematic review and meta-analysis. *Br. J. Gen. Pract.* **55**(513), 305–312 (2005)

13. Wu, Z., Balloccu, S., Kumar, V., Helaoui, R., Reforgiato Recupero, D., Riboni, D.: Creation, analysis and evaluation of ANNOMI, a dataset of expert-annotated counselling dialogues. *Future Internet* **15**(3), 110 (2023)
14. Wu, Z., et al.: Anno-mi: a dataset of expert-annotated counselling dialogues. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6177–6181. IEEE (2022)



Forecasting Gait Dynamics with Foundation Models

Federico Colelli Riano¹ , Alberto Malovini², Armando Coccia³ ,
Federica Amitrano³ , Giovanni D'Addio³ , and Riccardo Bellazzi¹

¹ Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy

federico.colelliriano01@universitadipavia.com

² Laboratory of Medical Informatics and Artificial Intelligence, Istituti Clinici Scientifici Maugeri IRCCS, Pavia, Italy

³ Bioengineering Unit of Telese Terme Institute, Istituti Clinici Scientifici Maugeri IRCCS, Telese Terme, Italy

Abstract. Foundation Models (FMs) are emerging as promising tools in health-care thanks to their ability to anticipate signal trends and provide early indications useful for clinical evaluation and decision support. In this study, FMs were used for short-term forecasting of inertial signals acquired during walking, adopting a multivariate approach on the triaxial accelerometer components. FMs (Moirai, Lag-Llama, Chronos) were used in both zero-shot (ZS) and fine-tuning (FT) modes, comparing their performances with baselines of widely used traditional Machine Learning (ML) models (XGBoost, LSTM, DeepAR). The results show that FMs effectively capture the dynamics of walking inertial signals and, especially after FT, offer competitive performance compared to baselines, with greater stability over longer horizons. The use of FMs for short-term predictions in gait analysis may open the way for their integration into multidisciplinary rehabilitation contexts thanks to the ability to forecast gait anomalies that might be useful to assess patient condition.

Keywords: Foundation Model · Time Series · Gait Analysis · Wearable Device

1 Introduction

Walking is a fundamental human activity requiring coordinated control of the musculoskeletal and nervous systems. In clinical contexts, gait analysis is essential for monitoring disorders that affect ambulation, such as Parkinson's disease or trauma related impairments; the use of wearable devices in gait analysis allows real-time acquisition of inertial signals, which can be processed to extract spatial and temporal parameters including: Gait Cycle Time (GCT) which is the time between two consecutive steps of the same foot; stance which is time of foot contact during gait cycle; swing which is the time foot-off contact during gait [1]. Artificial Intelligence (AI) can be used to process these signals, enabling short term forecasting to detect variations and potential anomalies, thereby supporting clinicians during rehabilitation monitoring. In this

context, Foundation Models (FMs) are particularly promising, as they offer strong generalisation capabilities even with limited domain specific data and can model temporal dynamics by generating plausible short horizon trajectories [2]. The aim of this study is to evaluate the feasibility of using FMs for short term prediction of accelerometric gait signals from which it is possible to capture walking dynamics in order to provide clinical insight such as gait abnormalities and risk of fall.

2 Materials and Methods

In this study, walking sessions were recorded using an e-textile sock: “Sensoria Smart Socks” by Sensoria Health (Redmond, WA, USA). The device is connected to an acquisition unit that integrates an inertial measurement unit (IMU) for triaxial signals (i.e., corresponding to x-axis, y-axis and z-axis) collection with a sampling rate frequency of 50 Hz; signals are measured in g ($1\text{ g} = 9,8\text{ m/s}^2$). Eight participants without gait impairments were recruited; each wore the sock on the left foot and walked 10 m along straight path, repeating the trial three times. In Fig. 1 the wearable device used in this study.



Fig. 1. The sensorised sock worn. The blue circle highlights the core. The red arrows illustrate the triaxial acceleration components. The x-axis and z-axis correspond to the base plane, while the y-axis indicates the orthogonal component.

2.1 Foundation Models

The FMs considered in this study were Chronos, Lag-Llama, and Moirai. Chronos is introduced as a family of general-purpose models for multi-domain forecasting, designed to simplify modelling pipelines and deliver ready-to-use predictions even on previously unseen data [3]. Lag-Llama is an open-source initiative that emphasises generalisation to novel contexts by employing a lightweight and easily adaptable configuration, built upon a pre-trained base derived from diverse collections of time series [4]. Moirai is developed to operate across heterogeneous scenarios and sampling frequencies using a unified framework, favouring immediate generalisation while enabling refinement through lightweight fine-tuning (FT) [5]. All FMs rely on large-scale pre-training on

broad and heterogeneous datasets, with the aim of achieving strong performance in zero shot (ZS) mode. Additionally, they are designed to support rapid performance improvements when FT is applied where necessary.

2.2 Baseline Models

The Machine Learning (ML) models considered as baselines in this work were: eXtreme Gradient Boosting (XGB), Long Short-Term Memory (LSTM) and DeepAR. XGB consists of a tree-based gradient boosting framework that flexibly models nonlinear relationships and variable interactions, often yielding competitive performance across benchmarking tasks [6]. LSTM consists of a recurrent neural network designed to capture long-term dependencies, thereby mitigating the gradient problems typical of training over extended time horizons. It is a popular choice for sequential tasks, such as those involving physiological or sensory signals [7]. Finally, DeepAR consists of a probabilistic forecasting approach based on recurrent autoregressive networks trained jointly across sets of related series; it produces predictive distributions on the future [8].

2.3 Evaluation Approach

FMs were evaluated both in ZS and FT mode, following the operational guidelines provided by each developer. ML models were assessed using 10 - fold cross-validation with a randomised hyperparameter search (20 configurations per fold). Sessions were grouped by subject, applying an 80%/20% split to ensure a strict separation between training and test sets and to prevent any form of data leakage. Forecasts were evaluated at 2 s and 4 s horizons. To characterise the information content of each component during the step, a multivariate approach was adopted to jointly predict the three axial acceleration components. This allowed estimation of the regression performance on each axis and facilitated comparison across models. Model performance was quantified using Mean Absolute Error (MAE), to express the average error on the measurement scale; Root Mean Squared Error (RMSE), to penalise larger prediction errors; and Coefficient of Variation (CV), to evaluate the relative deviation from each subject's reference signal. The data preparation, model execution, and statistical analysis were conducted in Python (v. 3.13).

3 Results

Figure 2 shows the forecasting performance across the three inertial axes for the two prediction horizons. Chronos FT consistently emerges as the most accurate and robust model, significantly outperforming traditional baselines and the other FMs. The bar plots highlight two key findings. First, Chronos FT drastically reduces both absolute errors and inter-subject variability. As visually supported by the lower CV above the bars, the FT version maintains high stability and reliability even when the time horizon is extended to 4 s. Second, while most baselines and ZS models suffer a performance degradation on the y-axis, Chronos FT successfully mitigates this complexity, ensuring consistently low errors across all axes. Finally, Lag Llama performs worst among the FMs.

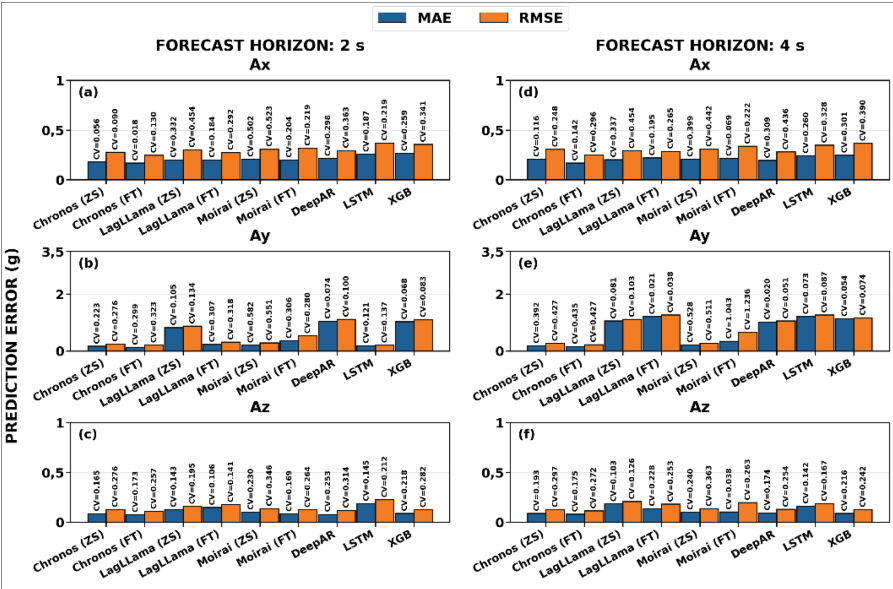


Fig. 2. Barplots of the forecast horizon for 2 s x-axis (a), y-axis (b), z-axis (c) and 4 s x-axis (d), y-axis (e) and z-axis (f).

4 Conclusions

This study demonstrates the feasibility of applying FMs to short-horizon forecasting of triaxial inertial signals acquired during walking, supporting their potential role in gait analysis and rehabilitation-oriented monitoring. Across both prediction horizons, Chronos FT emerges as the most accurate and consistent model overall, providing the lowest errors across axes and maintaining stable performance when moving from 2 s to 4 s. When considering axis-specific behaviour, the y-axis tends to show higher average prediction errors when results are aggregated across models, suggesting that this component can be comparatively more challenging to forecast. A plausible explanation is related to the sensor reference frame and placement, as the axis aligned with the lower-limb profile may combine gravitational and stabilisation components, introducing greater variability compared with the other axes acting mainly on the support plane [9, 10]. Future work should involve validation on larger and more heterogeneous populations, as well as the assessment of broader prediction horizons and clinically oriented endpoints, to better characterise the practical utility of these models in forecasting deviations from gait baselines and supporting rehabilitation.

References

1. Tao, W., Liu, T., Zheng, R., Feng, H.: Gait analysis using wearable sensors. *Sensors*. **12**(2), 2255–2283 (2012). <https://doi.org/10.3390/s120202255>

2. Awais, M., et al.: Foundation models defining a new era in vision: a survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **47**(4), 2245–2264 (2025). <https://doi.org/10.1109/TPAMI.2024.3506283>
3. Ansari, A.F., et al.: Chronos: learning the language of time series. (2024). <https://doi.org/10.48550/arXiv.2403.07815>
4. Liu, C., et al.: Moirai 2.0: when less is more for time series forecasting. (2025). <https://doi.org/10.48550/arXiv.2511.11698>
5. Rasul, K., et al.: Lag-llama: towards foundation models for probabilistic time series forecasting. (2023). <https://doi.org/10.48550/arXiv.2310.08278>
6. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794 (2016). <https://doi.org/10.1145/2939672.2939785>
7. Zhao, R., Wang, J., Yan, R., Mao, K.: Machine health monitoring with LSTM networks. In: *2016 10th International Conference on Sensing Technology (ICST)*, pp. 1–6 (2016). <https://doi.org/10.1109/ICSENST.2016.7796266>
8. Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T.: DeepAR: probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* **36**(3), 1181–1191 (2020). <https://doi.org/10.48550/arXiv.1704.04110>
9. van Hees, V.T., et al.: Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PLoS One.* **8**(4) (2023). <https://doi.org/10.1371/journal.pone.0061691>
10. Leporace, G., Batista, L.A., Nadal, J.: Prediction of 3D ground reaction forces during gait based on accelerometer data. *Res. Biomed. Eng.* **34**, 211–216 (2018). <https://doi.org/10.1590/2446-4740.06817>



Sequential Counterfactual Inference for Temporal Clinical Data: Addressing the Time Traveler Dilemma

Jingya Cheng, Alaleh Azhir, Jiazi Tian, and Hossein Estiri^(✉)

Massachusetts General Hospital, Boston, MA, USA
hestiri@mgh.harvard.edu

Abstract. Counterfactual inference enables clinicians to ask “what if” questions about patient outcomes, but standard methods assume feature independence and simultaneous modifiability, which could violate the assumptions of longitudinal clinical data. We introduce the **Sequential Counterfactual Framework**, which respects temporal dependencies in electronic health records by distinguishing immutable features (chronic diagnoses) from controllable features (lab values) and modeling how interventions propagate through time. Applied to 2,723 COVID-19 patients (383 Long COVID heart failure cases, 2,340 matched controls), we demonstrate that 38–67% of patients with chronic conditions would require biologically impossible counterfactuals under naive methods. We identify a cardiorenal cascade ($\text{CKD} \rightarrow \text{AKI} \rightarrow \text{HF}$) with relative risks of 2.27 and 1.19 at each step, illustrating temporal propagation that sequential but not naive counterfactuals can capture. Our framework transforms counterfactual explanation from “what if this feature were different?” to “what if we had intervened earlier, and how would that propagate forward?”, yielding clinically actionable insights grounded in biological plausibility. Code is available at https://github.com/clai-group/sequential_counterfactual_inference.git.

Keywords: Counterfactual inference · Temporal EHR · Long COVID · Explainable AI · Clinical decision support

1 Introduction

Counterfactual reasoning is central to clinical decision-making. When adverse outcomes occur, clinicians ask what might have happened under different circumstances, such as earlier treatment or better-controlled comorbidities. Machine learning enables this at scale: given a predictive model f , counterfactual explanation methods identify minimal feature changes that would alter the predicted outcome [1, 9]. For example, lowering HbA1c from 8.2% to 6.5% might reduce predicted heart failure risk from 73% to 41%, turning abstract risk scores into actionable targets.

J. Cheng and A. Azhir—Equal contribution (co-first authors).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
P. Andreev et al. (Eds.): AIME 2026, LNAI 16749, pp. 415–420, 2027.
https://doi.org/10.1007/978-3-032-30813-9_77

Standard methods, however, treat features as independent and simultaneously modifiable [2, 3, 11]. Clinical data are inherently *temporal*: a diagnosis in one year influences laboratory values and complications in subsequent years. This mismatch creates the **Time Traveler Dilemma**: models may generate biologically implausible counterfactuals such as “remove diabetes to reduce heart failure risk,” ignoring that chronic diseases persist and carry downstream consequences. Empirically, we find that 57% of patients in our cohort would require at least one implausible change under naïve methods.

We propose the **Sequential Counterfactual Framework** to address these limitations, integrating (i) a feature taxonomy distinguishing immutable, controllable, and intervention features; (ii) a temporal dependency graph encoding how conditions propagate over time; and (iii) plausibility constraints that any generated counterfactual must satisfy.

2 Background

Counterfactual explanations identify minimal changes to an input that would alter a model’s prediction, providing recourse-oriented interpretability [1, 9, 10]. Formally, given a classifier $f : \mathcal{X} \rightarrow \{0, 1\}$ and input $\mathbf{x}$ with $f(\mathbf{x}) = 1$, a counterfactual seeks $\mathbf{x}'$ with $f(\mathbf{x}') = 0$ minimizing $\lambda(f(\mathbf{x}') - y')^2 + d(\mathbf{x}, \mathbf{x}')$ [1]. Most methods implicitly assume features are independently and simultaneously modifiable. This assumption is problematic in clinical data, where variables are temporally ordered, biologically constrained, and interdependent; cross-sectional feature representations can yield implausible scenarios such as reversing chronic diagnoses or altering downstream consequences without upstream causes. Temporal models for EHRs, including recurrent and transformer-based architectures [5, 6], encode time-dependent structure. Frameworks such as TLDR [7, 8] partition observations relative to an index event to distinguish antecedent conditions from downstream manifestations. However, these temporal representations have not been incorporated into counterfactual generation. We address this gap by integrating temporal structure and biological plausibility constraints into counterfactual inference for longitudinal clinical data.

3 Methods

We applied the framework to patients from a large academic medical center with documented COVID-19 infection. Cases were patients who developed heart failure (ICD-10: I50.x) within 365 days, attributed to Long COVID [13]; controls were matched patients who recovered without Long COVID complications. Patient trajectories were structured using the TLDR temporal learning framework [7], implemented via the MLHO package in R, which partitions observations into three periods relative to the index infection: HISTORY (pre-infection), PAST (between first and last infection), and LAST (post-infection, before outcome). Across periods, 223 clinical concepts were extracted, including

diagnoses (ICD-10), medications (RxNorm/VA class), and laboratory abnormalities (LOINC with `_H/_L/_A` suffixes indicating high/low/abnormal values), yielding binary feature vectors for each concept–period combination. A gradient boosting machine (GBM; `MLH0` package, 5-fold cross-validation) was trained on concatenated temporal features to predict Long COVID-associated heart failure, achieving AUROC 0.88 (95% CI: 0.84–0.91).

Feature Taxonomy. Features are partitioned into three disjoint classes: **Immutable** ($\mathcal{I}$): chronic diagnoses that cannot be counterfactually reversed once present (e.g., E11 diabetes, I10 hypertension, N18 CKD, I50 heart failure). **Controllable** ($\mathcal{C}$): acute events and lab abnormalities that can change in response to intervention (e.g., Glucose_H, Creatinine_H, N17 AKI). **Intervention** ($\mathcal{R}$): actionable treatments (e.g., Lisinopril, Insulin, Metoprolol, loop diuretics).

Temporal Dependency Graph. A directed graph $G = (V, E)$ encodes feature–timepoint dependencies. An edge $(i, t) \rightarrow (j, t')$ is included if $P(x_j^{t'} = 1 \mid x_i^t = 1) / P(x_j^{t'} = 1 \mid x_i^t = 0) > \gamma = 2.0$. Key dependency types are: persistence ($dx_{i,h} \rightarrow dx_{i,l}$), medication→lab, lab→diagnosis, and diagnosis→complication cascades.

Plausibility Constraints. A counterfactual $\tau' = (\mathbf{h}', \mathbf{s}', \mathbf{l}')$ is *plausible* if it satisfies: **(P1) Immutability**: if $x_i^t = 1$ for $i \in \mathcal{I}$, then $x_i^{t'} = 1$ for all $t' \geq t$; **(P2) Temporal Coherence**: any change to a feature i requires a directed path in G from some intervention $j \in \mathcal{R}$ at an earlier time; **(P3) Conditional Plausibility**: $P(\mathbf{l}' \mid \mathbf{h}', \mathbf{s}', \text{interventions}) > \epsilon$ estimated from the empirical distribution. Sequential counterfactuals are generated by applying an intervention $\mathbf{r}$ at HISTORY and propagating its effects forward through G via learned conditional distributions, then enforcing P1–P3. A naïve counterfactual bypasses this propagation and directly modifies LAST features.

4 Results

The cohort comprised 2,723 patients: 383 (14.1%) Long COVID heart failure cases and 2,340 (85.9%) matched controls. Median time from infection to heart failure was 109 days (IQR: 67–156).

Temporal Persistence of Chronic Conditions. Chronic conditions in HISTORY persist to LAST at dramatically higher rates than new-onset diagnoses: diabetes, $P(E11_l = 1 \mid E11_h = 1) = 0.673$ vs. 0.050 (13.5 $\times$); hypertension, 0.520 vs. 0.091 (5.7 $\times$); CKD, 0.379 vs. 0.030 (12.6 $\times$). These 6–14 $\times$ persistence ratios are the empirical basis for P1 constraints.

Plausibility Violation Analysis. Table 1 quantifies naïve counterfactual violations. **P1 violations** (immutability) were dominant, affecting 54.4% of patients ($n = 1,481$): 95.6% of patients with hypertension in LAST had it in HISTORY, 96.0% for diabetes, and 87.6% for CKD. **P2 violations** (temporal coherence) affected 12.0% ($n = 328$): among patients with elevated glucose in LAST, 60.7% had no insulin in HISTORY; among patients with post-COVID AKI, 66.8% had no ACE inhibitor in HISTORY. **Combined**, 57.3% of patients ($n = 1,561$) would require at least one implausible counterfactual under naïve methods. The 4.5:1 ratio of P1 to P2 violations indicates that chronic disease persistence, not missing intervention pathways, is the primary barrier.

Table 1. Naïve counterfactual plausibility violations.

Violation type	n	%
<i>Patient-level (at least one feature affected)</i>		
P1: Immutability (chronic dx persists)	1,481	54.4
P2: Temporal coherence (no intervention path)	328	12.0
Any violation	1,561	57.3
<i>Feature-level P1 (diagnosis persistence, History $\rightarrow$ Last)</i>		
Hypertension (I10)	1,120/1,172	95.6
Type 2 Diabetes (E11)	834/869	96.0
CKD (N18)	346/395	87.6
<i>Feature-level P2 (missing intervention pathway)</i>		
Elevated glucose, no insulin in History	202/333	60.7
Post-COVID AKI, no ACE inhibitor in History	155/232	66.8

Cardiorenal Cascade. Among patients without prior AKI, $P(AKI_l = 1 \mid CKD_h = 1) = 0.068$ vs. 0.030 (**RR**=2.27); among all patients, $P(HF \mid AKI_l = 1) = 0.164$ vs. 0.138 (**RR**=1.19), yielding: $CKD_h \xrightarrow{RR=2.27} AKI_l \xrightarrow{RR=1.19} HF$. A naïve counterfactual proposing “remove AKI_last” ignores this upstream cause. Sequential counterfactuals instead propose renoprotective intervention at HISTORY (e.g., ACE inhibitor) and propagate its downstream effect on AKI risk and ultimately heart failure.

Confounding by Indication. Among diabetic patients, insulin-treated patients had *worse* glucose control ($P(\text{Glucose_H}_t = 1) = 0.193$ vs. 0.139 without insulin), reflecting that insulin is prescribed to sicker patients (CKD: 51.6% vs. 22.9% ; prior HF: 61.8% vs. 35.7%). This illustrates why medication counterfactuals require modeling the confounding structure: a naïve change to insulin status would mislead without this context.

5 Discussion and Conclusion

We introduced the Sequential Counterfactual Framework for longitudinal clinical data. Its core contribution is not a new optimization method but a principled set of constraints (immutability, temporal coherence, and conditional plausibility) that any counterfactual must satisfy to be clinically actionable. The cardiorenal cascade illustrates how the framework shifts the clinical question from “what feature should be different?” to “when and how should we intervene, and what propagates forward?”, aligning XAI with actual clinical reasoning [12].

Limitations. Results are based on single-center observational data; generalizability across institutions remains to be validated. The current evaluation is primarily descriptive (violation rates, conditional RRs) rather than a direct empirical comparison against existing counterfactual baselines such as DiCE or FACE [4], a limitation we plan to address in future work through head-to-head plausibility and predictive-shift metrics. The dependency graph encodes associations, not established causal relationships [14]; a validated causal graph could replace it without changing the constraint machinery. Additional limitations include binary feature representations, laboratory missingness ambiguity, and potential unmeasured confounders.

Conclusion. Standard counterfactual methods affect 57% of patients in our Long COVID cohort with biologically impossible or temporally incoherent scenarios. The Sequential Counterfactual Framework resolves this by grounding counterfactual inference in the temporal structure of clinical data, producing explanations that are both plausible and actionable.

References

1. Wachter, S., et al.: Counterfactual explanations without opening the black box: Automated Decisions and the GDPR. *Harvard J. Law Technol.* **31**(2), 841–887 (2018)
2. Mothilal, R.K., et al.: Explaining machine learning classifiers through diverse counterfactual explanations. In: *FAT*, pp. 607–617 (2020)
3. Karimi, A.H., Schölkopf, B., Valera, I.: Algorithmic recourse: from counterfactual explanations to interventions. In: *FAccT*, pp. 353–362 (2021)
4. Poyiadzi, R., et al.: FACE: Feasible and actionable counterfactual explanations. In: *AAAI/ACM AIES*, pp. 344–350 (2020)

5. Choi, E., et al.: RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In: NeurIPS, pp. 3504–3512 (2016)
6. Li, Y., et al.: BEHRT: transformer for electronic health records. *Sci. Rep.* **10**(1), 7155 (2020)
7. Cheng, J., Hgel, J., et al.: Temporal Learning with Dynamic Range (TLDR) for modeling recurrent exposure and treatment outcomes. *Sci. Rep.* (2026)
8. Estiri, H., et al.: Predicting COVID-19 mortality with electronic medical records. *NPJ Digit. Med.* **4**(1), 15 (2021)
9. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2019)
10. Verma, S., Dickerson, J., Hines, K.: Counterfactual explanations for machine learning: a review. [arXiv:2010.10596](https://arxiv.org/abs/2010.10596) (2020)
11. Mahajan, D., Tan, C., Sharma, A.: Preserving causal constraints in counterfactual explanations for machine learning classifiers. [arXiv:1912.03277](https://arxiv.org/abs/1912.03277) (2019)
12. Ghassemi, M., et al.: The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **3**(11), e745–e750 (2021)
13. Davis, H.E., et al.: Long COVID: major findings, mechanisms and recommendations. *Nat. Rev. Microbiol.* **21**(3), 133–146 (2023)
14. Pearl, J.: *Causality: Models, Reasoning and Inference*, 2nd edn. Cambridge University Press (2009)



Explainable Case-Based Diagnosis and Risk Prediction in Oncology

Isabelle Bichindaritz¹  and Leszek Kotula²

¹ SUNY Oswego, Oswego, NY 13126, USA
ibichind@oswego.edu

² SUNY Upstate Medical University, Syracuse, NY 13202, USA

Abstract. In medical domains, precision medicine highlights the importance of diagnostic, prognostic, and therapeutic choices tailored to each patient, based on the multimodal data available. In particular, so-called confounding factors such as demographic and socio-economic factors affect the outcome, so that their inclusion in a predictive model is of paramount importance. In precision oncology, genetic data play a key role, along with confounding factors, so that large amounts of such data is available to researchers to help decipher the association between multifactorial variables and diagnosis, prognosis, and therapy. This project demonstrates that a case-based approach can be advantageous in the handling of confounding factors and in explainability, not to mention model efficiency. The effectiveness of the case-based approach is demonstrated on both diagnostic and prognostic tasks in comparison with state-of-the-art statistical and machine learning models while affording precise user explanations at the case level.

Keywords: Case-based Reasoning · Survival Analysis · Precision Medicine

1 Introduction

To improve cancer diagnosis, risk stratification, and therapeutics, this article presents a case-based reasoning (CBR), diagnostic and prognostic ML model in oncology, O-CBR, proposing user explanations at the case level. This research affords:

- 1) **Multitask modeling:** the CBR system provides decision support recommendations in several tasks, including diagnosis and risk stratification, through classification and survival analysis models.
- 2) **Multidimensional data modeling:** the CBR system analyzes data from different types and dimensions, including demographic and socioeconomic data, and several types of genetic data (transcriptomic and epigenetic).
- 3) **Effective and efficient modeling:** the CBR system is at least as effective as state-of-the-art ML models used in this domain, including statistical models.
- 4) **Explainability:** the CBR system provides case-level explanations to users by design, through the discovery of genetic biomarkers, motifs and biological pathways.

2 Background

O-CBR combines the rich experience of CBR systems in medicine, enhanced with survival analysis, multitask capabilities, and explainability by design [1].

Case-Based Reasoning. Is a machine learning methodology related to analogical reasoning and cognitive modeling, with many applications in medical decision-making [2, 3]. CBR proceeds by first selecting the case(s) that is/are the most similar to the new problem (retrieval step) and by adapting it/them, taking into account the differences between the retrieved case(s) and the new problem (reuse/adaptation step).

Survival Analysis. Survival analysis is a well-known task aiming at predicting how likely an event is to happen over time. Major developments that contributed to this field are the Kaplan-Meier survival curve [4] and the Cox Proportional Hazard model [5].

Explainable AI. In medical domains, human-competitive interpretability has been shown to be crucial for the adoption of clinical decision support systems [6, 7]. Instead of after the fact explanation methods [8], XAI researchers stress the importance of ML models that are interpretable by design [9]. Recent user studies have demonstrated that showing a retrieved case to users is the most important factor affecting users' confidence in observed levels of goodness, convincingness, and trust ("Cases are king" [10]) and encourage the exploration of example-based explanations for XAI [11].

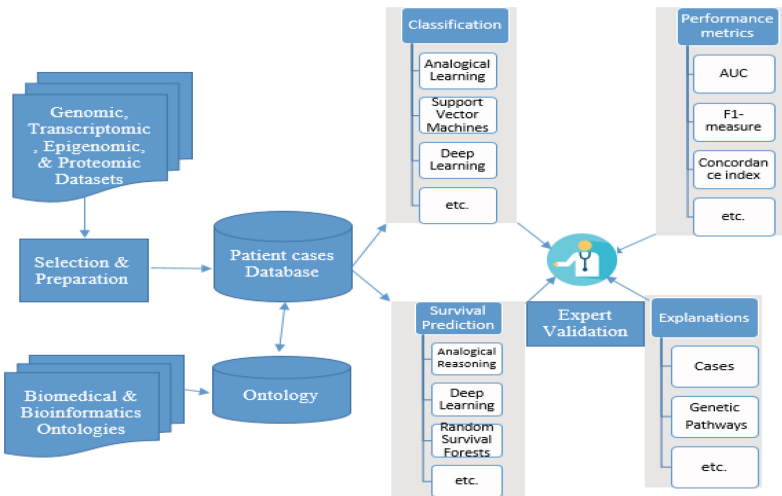


Fig. 1. O-CBR architecture.

3 Methods

Figure 1 show the system architecture Due to the high-dimensionality of the cases, up to 485,000 features for DNA methylation, feature selection was applied in each task.

3.1 Diagnosis

Feature Selection. The feature selection process can be described as a *multiscale and biologically significant reduction*. After batch effects removal, differentially methylated regions (DMR) were selected, and features further reduced [12, 13].

Retrieval. The retrieval process can be described as a two-stage retrieval tailored to each sample as *an optimal number of similar cases is retrieved* for each test case, using a confidence metric. During the two-stage process, memorized cases are retrieved by first finding cases with similar confounding factors. Then, the pool of retrieved cases is further refined using a Euclidean distance matrix from the other features.

3.2 Risk Prediction

If diagnosis is the first step in caring for a cancer patient, personalized therapy takes closely the severity risk under advisement.

Feature Selection. The feature selection process starts in the same manner as described in 3.1, as a *multiscale and biologically significant reduction*. To locate a subset of features highly specific to the overall survival of the cancer samples, multivariate Cox regression was applied to determine a prognostic score for each feature.

Retrieval. To establish prognostic groups for survival, a multivariate Cox regression was constructed using each DNA methylation feature as a covariate in order to test for each probe's contribution to survival. To find risk groups, the beta coefficient *BetaCoeff* for each feature j , among n , was multiplied by the DNA methylation β value *MethylBetaValue* and these values were then summed together to find a single prognostic score for each sample (see Eq. 1). Samples were ranked by their prognostic scores and divided into equal-sized low, medium and high risk groups.

$$PrognosticScore = \sum_{j=1}^n (BetaCoeff * MethylBetaValue) \quad (1)$$

A case-based retrieval framework was designed to automatically determine the appropriate number of retrieved cases for each test case, through a confidence metric. To better adjust for both the contribution of the prototypes and the potential power of each memorized case c at forming a solution for an unsolved case, we derived Eq. 2.

$$Confidence(c) = 1/2dist(c, P) + \frac{d(c, m1) + \dots + d(c, mk)}{k} \quad (2)$$

The confidence metric is one half of the distance d from case c to its prototype P , added to the average distance of case c to each memorized case m . The predicted survival time for a test case would then be the mean survival time of each retrieved case.

Adaptation. Once cases were retrieved and their average survival duration was assigned to the test case, a case adaptation phase was attempted. Case adaptation is the ability to morph training cases to reflect a test case more closely. The distance between a retrieved and a test case was added or subtracted to the retrieved case survival.

Explanations. In a CBR system, explanations may take several forms. The first form is to show the retrieved cases, with commonalities and differences with the test case. We also map genetic regions into pathways through gene set enrichment to display the main pathways involved in the commonalities and differences with the test case.

4 Results

The system’s effectiveness was experimentally tested on methylation data for the classification task and on multimodal methylation and transcriptomic data for the risk prediction task. The preferred performance metric for classification is the *F1 score*, particularly valued in binary classification where one class, the presence of disease, is more important. The preferred metric for survival analysis is the concordance index. For both metrics, higher values represent better performance, 1 being perfect performance.

4.1 Datasets

Classification Task. Breast cancer data was downloaded from The Cancer Genome Atlas (TCGA). After data preprocessing, 782 cancer samples, 96 normal samples and 364,464 DNA methylation probes were used, then reduced to 8,772 features.

Survival Analysis Task. The performance of the proposed method of survival analysis was studied on four cancer DNA methylation and gene expression datasets. We selected breast cancer (BRCA), Glioma cohort cancer (GBMLGG) and Pan-kidney cohort cancer (KIPAN) from TCGA (see Table 1).

Table 1. Gene and clinical characteristics in three cancer types.

Characteristics	BRCA	BRCA	GBMLGG	KIPAN
Genetic data type	Gene	DNA	DNA	DNA
Patient no.	Expression 761	Methylation 763	Methylation 650	Methylation 654
Gene no.	1,151	8,725	20,116	20,533
Selected features	23	70	586	749
Survival status	512	514	434	500
Living				
Deceased	249	249	216	154
Follow up (days)	1–1234	1–1234	1–481	3–5925

4.2 Experimental Results

Classification Task. O-CBR performance was determined with and without feature selection using 10-fold cross validation and compared with k NN using $1/d$ as distance weight and an optimal k of 10; Random Forest with 500 iterations; Support Vector Machine; and Naïve Bayes. All results indicate that integrating clinical covariates improves performance. The two-stage O-CBR outperformed all algorithms on the F1-score using all 8,772 features with high balanced accuracy over 95%. The results were close, but show the utility of clinical covariates and CBR for cancer classification.

Survival Analysis Task. We compared O-CBR with state-of-the-art survival prediction models: LASSO [14], and Random Survival Forests (RSF) [15]. For the sake of fairness, the same selected feature set was used in all 10-fold cross-validation tests.

Table 2. Survival performance by concordance index.

Methods	BRCA Gene expression	BRCA Methylation	GBMLGG Methylation	KIPAN Methylation
LASSO	0.5	0.5	0.6215333	0.6574272
RSF	0.4656382	0.4660381	0.5629102	0.7084548
O-CBR	0.7537651	0.6795273	0.6708844	0.7039824

Table 2 shows the O-CBR performed significantly better, at $p = .05$ level, than the other two methods on BRCA, GBMLGG and KIPAN datasets. On KIPAN dataset, RSF performed slightly better, but not significantly so.

5 Conclusion

In this paper we presented and evaluated a multitask CBR system for oncology providing decision support for diagnosis and risk severity prediction from multidimensional data that is both effective and explainable. Our future research directions will focus on extending the system to additional tasks, datasets, dimensions, and to comparison with additional machine learning models such as Gradient Boosting and deep learning [16], based on LSTM or transformer models [17]. Once fully optimized, field test studies are planned to evaluate the system acceptance and usefulness in clinical settings.

Acknowledgments. We want to thank the National Institutes of Health for their funding of this research and the TCGA Research Network: <https://www.cancer.gov/tcga>.

References

1. Weber, R.O., Johs, A.J., Goel, P., Silva, J.M.: XAI is in trouble. *AI Mag.* **45**(3), 300–316 (2024)

2. Bichindaritz, I., Marling, C.: Case-based reasoning in the health sciences: what's next? *Artif. Intell. Med.* **36**(2), 127–135 (2006)
3. Montani, S.: Exploring new roles for case-based reasoning in heterogeneous AI systems for medical decision support. *Appl. Intell.* **28**, 275–285 (2007)
4. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete samples. *J. Am. Stat. Assoc.* **73**(282), 457–481 (1958)
5. Cox, D.R.: Regression models and life-tables. *J. R. Stat. Soc. Ser. B Methodol.* **34**(2), 187–202 (1972)
6. Van Woensel, W., et al.: Explainable clinical decision support: towards patient-facing explanations for education and long-term behavior change. In: *International Conference On Artificial Intelligence in Medicine*, pp. 57–62. Springer International Publishing, Cham (2022)
7. Leveson, N.G.: The therac-25: 30 years later. *Computer.* **50**(11), 8–11 (2017)
8. Bai, S., Nasir, S., Khan, R.A., Meyer, A., Konik, H.: Breast cancer diagnosis: a comprehensive exploration of explainable artificial intelligence (XAI) techniques. *ArXiv:2406.00532*. (2024)
9. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**(5), 206–215 (2019)
10. Gates, L., Leake, D., Wilkerson, K.: Cases are king: a user study of case presentation to explain CBR decisions. In: *International Conference on Case-Based Reasoning*, pp. 153–168. Springer Nature Switzerland, Cham (2023)
11. Saeed, W., Omlin, C.: Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. *Knowl.-Based Syst.* **263**, 110273 (2023)
12. Anaissi, A.: Case-Base Retrieval of Childhood Leukaemia Patients Using Gene Expression Data. University of Technology Sydney (Australia) (2013)
13. Li, S., Harner, J., Adjero, D.A.: Random KNN feature selection-a fast and stable alternative to random forests. *BMC Bioinf.* **12**(1), 450 (2011)
14. Tibshirani, R.: The lasso method for variable selection in the Cox model. *Stat. Med.* **16**, 385–395 (1997)
15. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008)
16. Katzman, J.L., et al.: DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 24 (2018)
17. Bichindaritz, I., Liu, G., Bartlett, C.: Integrative survival analysis of breast cancer with gene expression and DNA methylation data. *Bioinformatics.* **37**, 2601–2608 (2021)



Multi-view Lumbar MRI for Therapy-Oriented Decision Support System in Low Back Pain

Ruben Piperno¹, Luca Bacco¹(✉), Giuseppe Francesco Papalia²,
Fabrizio Russo², Girolamo Maltese², Gianluca Vadalà², Rocco Papalia²,
Leandro Pecchia^{1,2}, and Mario Merone^{1,2}

¹ Research Unit of Intelligent Technology for Health and Wellbeing, Department of Engineering, Università Campus Bio-Medico di Roma, Via Alvaro del Portillo 21, 00128 Rome, Italy

l.bacco@unicampus.it

² Fondazione Policlinico Universitario Campus Bio-Medico, Via Alvaro del Portillo 200, 00128 Rome, Italy

Abstract. Therapy selection for symptomatic low back pain remains challenging since clinical decisions require patient-level synthesis of heterogeneous imaging findings. While AI methods for lumbar MRI mainly target diagnostic or structure-specific tasks, therapy-oriented decision support from multi-view acquisitions remains underexplored. We curated a retrospective multi-view lumbar MRI dataset of 294 patients annotated with three therapeutic pathways: *conservative*, *regenerative*, and *surgical*. Each examination includes five sagittal slices and up to three axial slices at clinically relevant levels. We propose a patient-level model based on a shared 2D slice encoder and attention-based aggregation of sagittal and axial evidence. In 5-fold cross-validation, the proposed model achieved $85.03 \pm 4.14\%$ macro-F1, outperforming sagittal-only and axial-only variants. Ablations show that sagittal views provide the main global context, axial slices add complementary information, and pre-training improves performance. These results support the feasibility of therapy-oriented multi-view MRI decision support for low back pain.

Keywords: Decision Support System · Lumbar MRI · Low Back Pain

1 Introduction

Low back pain (LBP) is a leading cause of disability worldwide and represents a major clinical and socioeconomic burden [1]. Selecting the most appropriate therapeutic pathway is often complex, since symptomatic patients may require conservative rehabilitation, regenerative procedures, or surgery depending on the integration of clinical and imaging findings [2]. Lumbar MRI is routinely used in this process, as it provides detailed information on intervertebral discs, neural structures, stenosis, and paraspinal tissues. However, degenerative findings are

often multi-level and may also appear in asymptomatic subjects, limiting the direct translation of isolated MRI abnormalities into treatment decisions [3].

Artificial Intelligence (AI) methods for lumbar spine MRI have shown promising results, but most studies address diagnostic or structure-specific tasks, such as disc degeneration grading, stenosis assessment, or paraspinal muscle quantification [4]. Although these approaches support automated image interpretation, they generally predict radiological findings rather than patient-level therapeutic decisions. Moreover, many pipelines rely primarily on sagittal MRI, whereas clinical assessment often combines sagittal overview with axial level-specific information. The relative contribution of sagittal and axial views for therapy-oriented prediction remains insufficiently characterized.

To address this gap, we curated a retrospective multi-view lumbar MRI dataset of symptomatic LBP patients annotated with therapy-oriented labels: *conservative*, *regenerative*, and *surgical*. We propose a patient-level decision support system that encodes sagittal and axial slices with a shared 2D backbone and fuses slice-level evidence through patient-level aggregation. The main contributions of this work are: (i) a multi-view lumbar MRI dataset with therapy-oriented labels, planned for public release with documentation and standardized splits; (ii) a patient-level model for therapeutic pathway prediction from routine MRI; and (iii) an ablation analysis assessing the role of sagittal and axial views, aggregation strategy, and pre-training.

2 Materials and Methods

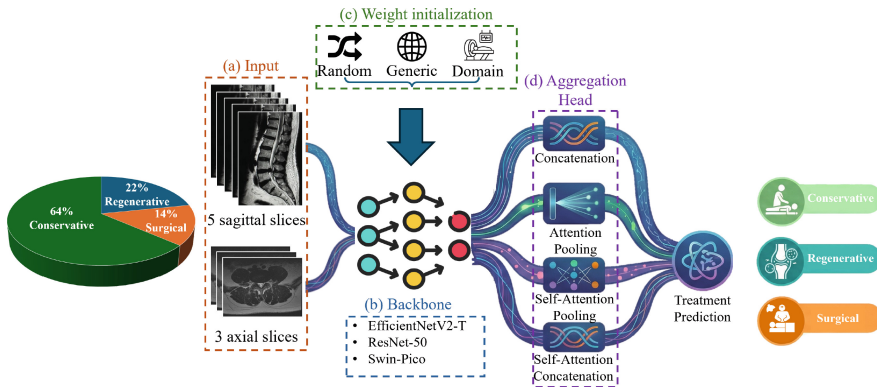


Fig. 1. Overview of the proposed multi-view MRI pipeline for patient-level therapeutic pathway prediction.

2.1 Dataset

We curated a retrospective cohort of $n = 294$ symptomatic LBP patients undergoing routine lumbar spine MRI. Each patient was assigned a therapy-oriented

label derived from the documented pre-operative clinical decision: *conservative* ($n = 187$, 63.6%), *regenerative* ($n = 64$, 21.8%), or *surgical* ($n = 43$, 14.6%). The regenerative class includes minimally invasive biologic procedures, the conservative class includes rehabilitation and non-invasive management, and the surgical class includes decompression and/or stabilization procedures. Labels reflected the final clinical indication recorded in the patient pathway: since labels encode real-world therapeutic decisions, they should be interpreted as pathway-level clinical targets rather than purely image-derived ground truth.

MRI exams included sagittal and axial T2-weighted TSE acquisitions from 1.5T Siemens scanners, consistent with prior data from the same clinical source [4]. Each patient was represented by five canonical sagittal slices and up to three axial slices at L3-L4, L4-L5, and L5-S1, when available. For the multi-view setting, sagittal and axial slices were concatenated in a fixed order. To mitigate class imbalance, offline patient-level augmentation was applied to surgical cases during training only, keeping augmented replicas within the same split/fold as the original patient.

2.2 Framework

Our approach formulates therapeutic pathway prediction as a patient-level multi-class classification problem from multi-view lumbar MRI. As illustrated in Fig. 1, the pipeline consists of four main components:

(a) Input Representation. Each examination is represented as a set of 2D slices. We use five canonical sagittal slices for the sagittal-only setting, up to three axial slices at clinically relevant levels for the axial-only setting, and their fixed-order concatenation for the multi-view setting. Missing axial levels are handled through padding and a slice-validity mask.

(b) Backbone Encoder. All slices are processed by a shared 2D backbone to obtain slice-level embeddings. We screened EfficientNetV2-T [5], ResNet-50 [6], and a lightweight Swin variant [7]. EfficientNetV2-T showed the most stable performance across modalities in preliminary hold-out screening and was therefore selected for the main cross-validation experiments.

(c) Weight Initialization. We compared random initialization, generic natural-image pre-training, and domain-relevant radiology/MRI pre-training. Domain-relevant checkpoints included RadImageNet [8] and a lumbar MRI checkpoint [9] when compatible with the given backbone. RGB-pretrained models were adapted to single-channel MRI inputs through a lightweight input conversion layer when required.

(d) Patient-Level Aggregation. Slice-level embeddings are fused into a patient representation using one of four heads: concatenation, attention pooling, self-attention pooling, or self-attention followed by concatenation. The final classifier predicts one of the three therapeutic pathways.

All evaluations were performed at patient level to avoid subject overlap across splits. Macro-F1 was used as the primary metric because of class imbalance. A preliminary hold-out screening was used only for backbone-family selection and not for final performance reporting; after fixing the backbone, the selected pipeline was evaluated using stratified 5-fold cross-validation. Models were trained with cross-entropy loss and AdamW optimization, selecting the best checkpoint according to validation loss. One-factor-at-a-time ablations were then performed on input views, aggregation head, and pre-training strategy.

3 Results

In the preliminary hold-out screening, EfficientNetV2-T showed the most stable performance across modality settings and was selected for the 5-fold cross-validation experiments. As shown in Table 1, the proposed multi-view model with attention pooling and generic pre-training achieved $85.03 \pm 4.14\%$ macro-F1. Removing axial slices reduced performance by 7.75 points, indicating that axial levels provide complementary information when combined with sagittal context. Removing sagittal slices caused a larger drop of 32.83 points, confirming that sagittal views provide the main patient-level evidence for therapy prediction. Attention pooling outperformed concatenation and self-attention variants, suggesting that explicit slice weighting is beneficial under variable axial availability. Generic pre-training provided the most reliable initialization: random initialization decreased macro-F1 by 10.85 points, while the domain checkpoint improved over random initialization but remained less stable than generic pre-training.

Table 1. Ablation study in 5-fold cross-validation using EfficientNetV2-T [5]. Values are macro-F1 (% , mean $\pm$ std). Δ is w.r.t. base model.

Setting	Macro-F1 (%)	Δ
<i>Proposed model</i>		
<i>Multi-view, AttnPool, Pre-training</i>	85.03 ± 4.14	-
<i>(a) View ablation</i>		
- <i>Axial</i>	77.28 ± 7.75	-7.75
- <i>Sagittal</i>	52.20 ± 6.78	-32.83
<i>(b) Head ablation</i> (– AttnPool)		
+ <i>Concat</i>	76.58 ± 12.49	-8.45
+ <i>SA-Pool</i>	76.18 ± 8.57	-8.85
+ <i>SA-Concat</i>	74.94 ± 6.12	-10.09
<i>(c) Pre-training ablation</i> (– Generic Pre-training)		
+ <i>Random initialization</i>	74.18 ± 6.98	-10.85
+ <i>Domain pre-training</i>	77.13 ± 11.44	-7.90

4 Conclusions

We presented a multi-view lumbar MRI decision support system for patient-level therapeutic pathway prediction in symptomatic LBP. The proposed model combines sagittal and axial evidence and achieved $85.03 \pm 4.14\%$ macro-F1 in 5-fold cross-validation. Ablation results indicate that sagittal views provide the main global context, axial slices add complementary level-specific information, and attention pooling with generic pre-training yields the most robust configuration. Since treatment selection also depends on clinical variables beyond MRI, the proposed system should be interpreted as imaging-based decision support rather than a standalone treatment recommendation tool. Future work will focus on larger multi-center validation, integration of clinical variables, and interpretability tools to support clinical deployment.

Acknowledgments. Authors were supported by the European Union - Next Generation EU - NRRP M6C2 - Investment 2.1 Enhancement and strengthening of biomedical research in the NHS, project n. PNRR-MAD-2022-12376692_VADALA' - CUP F83C22002470001.

Declaration of AI assistance technologies during manuscript preparation. The authors used ChatGPT (OpenAI) to support manuscript writing and editing, and NotebookLM (Google) to assist in creating illustrative figures. After using these technologies, the authors reviewed the results and take full responsibility for the contents of the manuscript.

References

1. Knezevic, N.N., Candido, K.D., Vlaeyen, J.W.S., Van Zundert, J., Cohen, S.P.: Low back pain. *The Lancet* **398**(10294), 78–92 (2021)
2. D’Antoni, F., et al.: Artificial intelligence and computer aided diagnosis in chronic low back pain: a systematic review. *Int. J. Environ. Res. Public Health* **19**(10), 5971 (2022)
3. Brinjikji, W., et al.: MRI findings of disc degeneration are more prevalent in adults with low back pain than in asymptomatic controls: a systematic review and meta-analysis. *Am. J. Neuroradiol.* **36**(12), 2394–2399 (2015)
4. Giaccone, P., et al.: Artificial intelligence based intervertebral discs grading: from unsupervised detection to explainable PFIRRMANN classification. *Biomed. Signal Process. Control*, **121**, 110177 (2026)
5. Tan, M., Le, Q.V.: Efficientnetv2: smaller models and faster training. In: *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 10096–10106 (2021)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016)
7. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022 (2021)

8. Mei, X., Liu, Z., Robson, P.M., Marinelli, M., Huang, Y., Chandra, A., et al.: Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiol. Artif. Intell.* **4**(5), e210315 (2022)
9. Batra, A., Gumber, A., Kumar, A.: M-scan: a multistage framework for lumbar spinal canal stenosis grading using multi-view cross attention (2025). [arXiv:2503.01634](https://arxiv.org/abs/2503.01634)



A Multimodal Wearable Framework for Predicting Epileptic Seizures

Agnese Bonfigli¹ , Alberto Maria Di Giacinto¹ , Flavia Giordani¹ ,
Leandro Pecchia^{1,2} , Mario Merone^{1,2} , and Luca Bacco¹

¹ Research Unit of Intelligent Technology for Health and Wellbeing, Department of Engineering, Università Campus Bio-Medico di Roma, Via Alvaro del Portillo 21, 00128 Rome, Italy

l.bacco@unicampus.it

² Fondazione Policlinico Universitario Campus Bio-Medico, Via Alvaro del Portillo 200, 00128 Rome, Italy

Abstract. Epileptic seizures remain largely unpredictable in daily life, especially for drug-resistant patients, reducing autonomy and safety. Wearable sensing can support continuous monitoring, but seizure prediction remains challenging because pre-ictal patterns are subtle, patient-dependent, and easily confounded by artifacts. We propose a wearable-oriented multimodal framework integrating behind-the-ear EEG (bte-EEG), electrocardiography (ECG), and electromyography (EMG). Continuous recordings are segmented into 20-s windows and represented through handcrafted features, with additional feature selection for EMG. A Multi Layer Perceptron classifier is trained under a subject-independent split and evaluated at seizure level over a 30-min prediction horizon. Across unimodal and multimodal settings, bte-EEG provides the strongest single-modality performance, whereas ECG alone is more prone to false positives. Trimodal fusion achieves the most favorable operating point, reducing false alarms while maintaining high sensitivity. These results support the feasibility of multimodal wearable systems designed to raise an alert when a seizure is expected within the next 30 min.

Keywords: Seizure prediction · Wearable multimodal biosignals · Multimodal fusion

1 Introduction

Epilepsy is a chronic neurological disorder characterized by recurrent seizures, whose unpredictability can severely affect safety, autonomy, and quality of life, especially in drug-resistant patients [1, 2]. Reliable short-term seizure prediction could enable timely warnings and personalized interventions, but its translation to realistic clinical and wearable settings remains challenging [3]. Unlike detection, which targets the ictal event, prediction requires identifying subtle pre-ictal changes before seizure onset. Reported performance is therefore strongly

influenced by the definition of the pre-ictal period, the prediction horizon, the windowing strategy, and the handling of false alarms.

In continuous monitoring, high sensitivity alone is not sufficient: frequent false positives may reduce clinical usefulness, user trust, and long-term adherence [9]. Wearable systems can support long-term monitoring outside controlled environments, but they must remain robust to limited channel availability, motion artifacts, and signal non-stationarities. Behind-the-ear electroencephalography (bte-EEG) offers a more wearable alternative to conventional EEG, while peripheral biosignals can provide complementary information. Electrocardiography (ECG) may reflect autonomic changes associated with seizure dynamics, whereas electromyography (EMG) may capture neuromuscular activity relevant under wearable conditions [5, 10].

Previous wearable-oriented studies have explored multimodal machine-learning pipelines based on central and peripheral signals [6]; however, the added value of EMG remains less consistently quantified, especially in fully wearable configurations and in comparisons that separate unimodal, bimodal, and trimodal contributions. A further limitation is the limited comparability of prediction results across studies. Definitions of the pre-ictal period, seizure prediction horizon, seizure occurrence period, and alarm post-processing can substantially change what is counted as a correct prediction and how false positives are measured [7–9]. This is particularly relevant in continuous monitoring, where window-level outputs must be converted into actionable warnings. For this reason, the sensitivity–alert–burden trade-off should be treated as a central evaluation criterion. Motivated by these gaps, this work investigates a **fully wearable multimodal system** targeting seizures within a *30-min prediction horizon*. The study contributes by: evaluating a setting designed for continuous use; comparing unimodal, bimodal, and trimodal configurations to quantify modality complementarity; and adopting an evaluation strategy that explicitly accounts for windowing choices and false-alarm behavior.

2 Materials and Methods

This section describes the implementation and operational design of the proposed pipeline illustrated in Fig. 1.

2.1 Dataset and Preprocessing

The study used the publicly available SeizeIT2 dataset, collected in an Epilepsy Monitoring Unit from patients with drug-resistant epilepsy [4]. Seizure onset and offset times were annotated by clinical staff on scalp video-EEG and temporally aligned with the wearable recordings. Three synchronized modalities were considered: two-channel bte-EEG, one-channel ECG, and one-channel EMG from the deltoid muscle, selected as a proximal site less prone to continuous motion artifacts than distal muscles. Labels were assigned according to a predefined seizure prediction horizon (SPH). A **window** was labelled as pre-ictal when

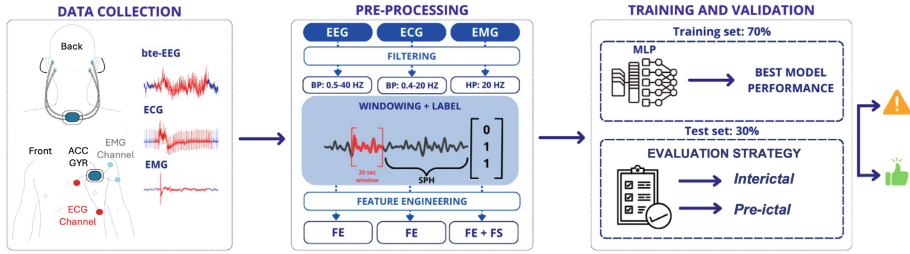


Fig. 1. Overview of the proposed wearable framework. Bte-EEG, ECG, and EMG recordings are segmented into time windows, transformed into handcrafted features, and classified by an MLP under a seizure-level evaluation protocol.

seizure onset occurred after the selected horizon, with a tolerance of ± 2 min; otherwise, eligible windows were labelled as interictal. Interictal samples were required to be at least 30 min away from ictal events to reduce contamination from seizure-related activity [7]. EEG and ECG windowing followed [6], while the EMG-specific strategy was based on [10].

Preprocessing and Feature Engineering. For each subject, recordings were reconstructed into continuous modality-specific time series and filtered to reduce wearable-related artifacts: 0.5–40 Hz for EEG, 0.4–20 Hz for ECG, and high-pass filtering at 20 Hz for EMG. Signals were segmented into 20-s windows with 50% overlap. To reduce post-ictal confounding, segments within 60 min after seizure termination were excluded unless they overlapped with a subsequent seizure. Each window was represented by handcrafted features selected from prior work on wearable and multimodal biosignals [6,10]. The full representation included 72 features, as detailed in Table 1. EMG feature selection was then applied.

Table 1. Summary of handcrafted feature groups extracted from each modality. For bte-EEG, features are computed per channel and concatenated.

Modality	Feature groups	N
bte-EEG	Time-domain statistics, spectral power, and spectral entropy features computed on two channels	36
ECG	RR-interval statistics and heart-rate variability features in time and frequency domains	23
EMG	Time-domain amplitude, activation, complexity, and statistical descriptors	13
Total		72

2.2 Model and Experimental Design

An MLP classifier was used for window-level binary classification. The network consists of six fully connected hidden layers with 128, 256, 256, 128, 64, and 32 neurons. Each layer is followed by batch normalization, ReLU activation, and

dropout to stabilize training and reduce overfitting. The output layer provides a probabilistic estimate of the pre-ictal state. The pipeline was first evaluated across multiple SPH values to select the prediction horizon. Dedicated models were then trained for all unimodal, bimodal, and trimodal configurations. Data were split by subject into 70% training and 30% test sets to assess generalization to unseen individuals. Class imbalance was addressed by undersampling the majority class in the training set.

Evaluation Strategy. Test samples were generated using a class-specific protocol. Interictal windows were extracted after excluding, for each annotated seizure, both the ictal interval and the 30 min preceding onset, corresponding to the selected SPH. This avoided contamination from pre-ictal activity. Twenty-second windows were then sampled from the remaining recording segments and treated as interictal. For pre-ictal evaluation, each seizure was assessed over a 30-min prediction interval. Given a seizure with onset time t_i and duration D_i , 20-s windows were evaluated from $t_i - \text{SPH}$ to $t_i + D_i - \text{SPH}$, with a 1-min step. A seizure was considered correctly predicted if at least one evaluated window was classified as pre-ictal.

Table 2. MLP performance across sensing configurations. Metrics are percentages. **Bold** indicates the best unimodal and bimodal results; for FP, lower is better. **Bold underlined** values mark trimodal improvements over both groups.

Metric (%)	Unimodal			Bimodal			Trimodal
	EEG	ECG	EMG	EEG+ECG	EEG+EMG	ECG+EMG	EEG+ECG+EMG
F1-score	44.0	10.0	42.0	48.0	30.0	30.0	<u>69.0</u>
Accuracy	67.0	11.0	64.0	77.0	39.0	39.0	<u>96.0</u>
Sensitivity (TP)	98.7	100.0	92.4	98.7	92.4	92.4	94.9
False-positive rate (FP)	33.2	90.2	36.1	23.3	61.5	61.5	<u>3.6</u>

3 Results and Discussion

Table 2 reports the performance of the proposed framework. Metrics include macro F1-score, accuracy, seizure-level sensitivity (TP), and false-positive rate (FP), jointly capturing prediction capability and alert burden under continuous monitoring.

Overall Performance and Multimodal Fusion. The trimodal configuration provides the most balanced operating point, achieving the highest F1-score and accuracy, the lowest false-positive rate, and high sensitivity. This is particularly relevant for continuous monitoring, where practical usability depends not only on detecting upcoming seizures but also on limiting unnecessary alerts. The improvement over unimodal and bimodal settings suggests that combining EEG with peripheral correlates provides complementary information and stabilizes predictions under wearable conditions.

Modality Contribution. Among single modalities, EEG provides the best balance between sensitivity and false positives, confirming its central role in pre-ictal discrimination. EMG shows intermediate performance, suggesting useful neuromuscular cues but higher sensitivity to wearable variability. ECG reaches high sensitivity but produces many false positives, likely because autonomic changes are not seizure-specific. In bimodal settings, EEG-based pairs outperform ECG+EMG, supporting EEG as the anchor signal. EEG+ECG achieves the strongest bimodal balance, whereas ECG+EMG is the weakest combination, indicating that peripheral signals alone provide less stable interictal/pre-ictal separation.

4 Conclusion

This study investigated short-term prediction in a fully wearable setting by integrating bte-EEG, ECG, and EMG within a feature-based MLP pipeline. Using 20-s windows, a 30-min horizon, and a subject-independent split, the framework enabled a controlled comparison of unimodal, bimodal, and trimodal configurations while explicitly considering the sensitivity–false-alarm trade-off. Results showed that bte-EEG was the most informative single modality, whereas ECG alone was associated with a higher false-positive rate. Multimodal fusion improved robustness, and the trimodal configuration achieved the best operating point by maintaining high sensitivity while substantially reducing false alarms.

These findings support the feasibility of wearable multimodal early-warning systems. Future work should validate the approach on larger and more diverse cohorts, improve robustness to artifacts and signal non-stationarities, and investigate personalized calibration to stabilize alarm burden in long-term use.

Declaration

of AI Assistance Technologies During Manuscript Preparation. ChatGPT (OpenAI) was used to support manuscript editing. The authors reviewed all outputs and take full responsibility for the final content.

References

1. Fisher, R.S., Acevedo, C., Arzimanoglou, A., et al.: ILAE official report: a practical clinical definition of epilepsy. *Epilepsia* **55**(4), 475–482 (2014)
2. Kwan, P., Arzimanoglou, A., Berg, A.T., et al.: Definition of drug resistant epilepsy: consensus proposal by the ad hoc task force of the ILAE commission on therapeutic strategies. *Epilepsia* **51**(6), 1069–1077 (2010)
3. Kuhlmann, L., Lehnertz, K., Richardson, M.P., Schelter, B., Zaveri, H.P.: Seizure prediction - ready for a new era. *Nat. Rev. Neurol.* **14**(10), 618–630 (2018)
4. Bhagubai, M., et al.: SeizeIT2: wearable dataset of patients with focal epilepsy. *Sci. Data* **12**, 1228 (2025)
5. Bhagubai, M., Vandecasteele, K., Swinnen, L., et al.: The power of electrocardiography in semi-automated seizure detection in addition to two-channel behind-the-ear electroencephalography. *Bioengineering* **10**(4), 491 (2023)

6. Zambrana-Vinaroz, D., Vicente-Samper, J.M., Manrique-Cordoba, J., et al.: Wearable epileptic seizure prediction system based on machine learning techniques using electrocardiography, photoplethysmography and electroencephalography signals. *Sensors* **22**(23), 9372 (2022)
7. Hazarika, N., Chen, J., Tsoi, A.C., Sergejew, A.: On the proper selection of preictal period for seizure prediction. *Epilepsy Behav.* **46**, 158–166 (2015)
8. Wu, X., et al.: Seizure prediction using deep learning: definitions of seizure prediction horizon and seizure occurrence period. *Front. Neurosci.* (2023)
9. Batista, J., Pinto, M.F., Tavares, M., et al.: EEG epilepsy seizure prediction: the post-processing stage as a chronology. *Sci. Rep.* **14**, 407 (2024)
10. Zhang, J., Swinnen, L., Chatzichristos, C., et al.: Multimodal wearable EEG, EMG and accelerometry measurements improve the accuracy of tonic-clonic seizure detection. *Physiol. Meas.* **45**(6), 065005 (2024)



Derived Neural Markers of Creativity: Debiased Weighted Phase Lag Index for Neurocognitive State Recognition

Morteza Zangeneh Soroush^{1,2} , Daniel Massicotte² , and Wei-Ping Zhu¹ 

¹ Université du Québec à Trois-Rivières, Trois-Rivières, Canada

morteza.zangenehsoroush@concordia.ca

² Concordia University, Montreal, Canada

Abstract. Creativity emerges through dynamic transitions across multiple cognitive states, and its neural mechanisms remain insufficiently understood. This study investigates brain network dynamics using an EEG-based functional connectivity framework based on the debiased Weighted Phase Lag Index (DwPLI). Participants performed a modified Torrance Test of Creative Thinking (TTCT-F) designed to elicit four cognitive states: idea generation (IDG), idea evolution (IDE), idea rating (IDR), and rest (RST). From the resulting connectivity networks, interpretable graph-theoretic features—node strength and betweenness centrality—were extracted, and statistically significant, discriminative features were identified through a structured feature selection pipeline. DwPLI-derived features demonstrated strong statistical significance and revealed robust neural markers, primarily within frontal–parietal networks, for differentiating cognitive states. Classification results showed high performance across models, with SVM achieving up to ~93% accuracy under ten-fold validation and ~90% under one-subject-out validation, indicating strong generalizability. The findings demonstrate that creative cognition is governed by the interplay between large-scale network integration and hub-based information routing. This work provides an efficient, interpretable, and scalable framework for cognitive-state recognition, with implications for real-time neurocognitive applications.

Keywords: Creativity · EEG · DwPLI · Brain Functional Connectivity

1 Introduction

Creativity is the capacity to generate novel, effective, and non-obvious ideas or solutions [1], emerging from dynamic interactions among higher-order cognitive processes such as memory, perception, learning, and reasoning across multiple phases [2]. Design is a goal-directed process aimed at transforming existing situations into more desirable outcomes [3], where design creativity involves generating novel and practically viable solutions while addressing ill-defined problems under contextual constraints [4, 5]. Early linear models of creativity [2, 6] fail to capture the dynamic nature of creativity [7], which is now viewed as a nonlinear, iterative process involving cycles of generation, refinement,

and evaluation [8]. Neuroscientific evidence supports this perspective, showing that creativity arises from dynamic brain processes and transitions across cognitive states [9–11]. From a systems viewpoint, cognition emerges from nonlinear interactions within distributed neural networks, producing time-varying connectivity patterns [12]. However, most prior work relies on channel-based analyses, overlooking inter-regional interactions, with inconsistent findings across cognitive states [11]. To address this gap, this study investigates EEG-based functional connectivity and extracts graph-based features to classify cognitive states of design creativity.

2 Materials and Methods

2.1 Dataset

This study employs an EEG dataset developed in prior work [13], subsequently used in several studies [9, 10, 14], and later publicly released with comprehensive documentation [15]. The dataset comprises EEG recordings from 28 right-handed graduate engineering students (4 females, aged 22–35 years), all in good health and without any reported neurological or psychiatric conditions. The experimental protocol is based on a modified figural version of the Torrance Test of Creative Thinking (TTCT-F). Each participant completed three independent trials. Each trial consisted of sequential phases: an initial resting state (RST), Idea Generation (IDG), Idea Evolution (IDE), Idea Rating (IDR), and a final resting period. Rest phases (3 min, eyes closed) provided baseline before and after the task sequence. During the IDG phase (≤ 3 min), participants completed the given sketch to generate an initial idea. In the IDE phase (≤ 3 min), participants revisited the same stimulus to produce a substantially different solution. In the IDR phase (≤ 3 min), participants evaluated the cognitive effort associated with their ideation using NASA-TLX-based self-assessment measures. All tasks were conducted in a loosely controlled, self-paced framework, allowing participants to follow natural creative workflows. EEG signals were recorded using a 64-channel Brain Vision actiCHamp system (500 Hz sampling rate). EEG signals were preprocessed and segmented into task-related phases.

2.2 Brain Functional Connectivity; Debiased Weighted Phase Lag Index

Connectivity was estimated by debiased weighted Phase Lag Index (DwPLI):

$$DwPLI = \frac{\sum_{j=1}^N \sum_{i \neq j} \text{Im}(S_j) \text{Im}(S_i)}{\sum_{j=1}^N \sum_{i \neq j} |\text{Im}(S_j) \text{Im}(S_i)|} \quad (1)$$

where S_j denotes the cross-spectrum between two signals at epoch j and $\text{Im}(S_j)$ is the imaginary component of the cross-spectrum [16]. Two local graph-based features were extracted. Node strength (S) reflects the overall level of functional integration, while betweenness centrality (B) captures the node's importance in information transfer.

2.3 Feature Selection, Classification, and Validation

Connectivity analysis generated a high-dimensional feature space across all the EEG channels. Our feature selection includes outlier correction, normalization, and removal of correlated features, and sequential forward/backward selection. We trained and evaluated multiple classifiers—k-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF). KNN used Euclidean distance with $k = 10$; SVM employed a third-degree polynomial kernel; and RF was configured with 100 trees and a maximum depth of 10 to control complexity.

3 Results

3.1 Statistical Analysis

ANOVA results (Table 1) indicate that features derived from DwPLI are highly significant across EEG channels (Ch; 64 channels), cortical regions (A; frontal (F), parietal (P), temporal (T), central (C), and occipital (O) lobes), and hemispheres (H; left (L) and right (R)). Feature rankings were determined based solely on statistical significance.

Table 1. Significant features identified by the proposed feature selection framework

Feature Type	#	Ch	A	H	p-value	Partial η^2
Strength (S)	1	TP7	T	L	0.0005	0.2354
	2	P1	P	L	0.0006	0.2674
	3	C4	C	R	0.0012	0.1475
	4	AF7	F	L	0.0039	0.0931
	5	FC4	F	R	0.0047	0.0754
Betweenness Centrality (B)	1	FC4	F	R	0.0001	0.3892
	2	P7	P	L	0.0067	0.0638
	3	AF4	F	R	0.0094	0.0257

Table 1 shows that a compact set of DwPLI-derived features—primarily from frontal and parietal regions—significantly differentiates the four cognitive states. Node strength and betweenness centrality provide meaningful neural markers of design creativity, capturing global integration and hub-based routing, respectively. FC4 emerges as a key hub, particularly in betweenness centrality, highlighting the role of right fronto-central regions, with additional contributions from channels such as C4 linked to motor planning, attention, and cognitive control. Eight significant features (five strength, three centrality) were identified, suggesting a greater role of connectivity strength. Strength features are widely distributed across lobes and hemispheres, whereas centrality features are more localized, mainly in the right frontal region.

3.2 Classification

The selected features were used to classify the cognitive states. Tables 2 and 3 present the classification results. Model performance was evaluated by ten-fold and one-subject-out cross-validation using accuracy (ACC), recall (REC), specificity (SPC), precision (PRC), F1-score (F1SC), and Cohen’s kappa (KPA). Under ten-fold cross-validation, SVM yielded the best results (93.22 ± 3.37 for accuracy). Consistently high precision, recall, and specificity indicate balanced and reliable classification across classes. It shows that the proposed method is effective and discriminates the brain dynamics effectively. Under the one-subject-out scheme, performance decreased across all models, while preserving the same ranking. SVM remained the top performer (90.58 ± 5.47 for accuracy), demonstrating strong subject-independent generalization. The higher performance observed under ten-fold cross-validation reflects partial subject overlap between training and testing sets, whereas the one-subject-out scheme evaluates true cross-subject generalization. The resulting performance gap highlights inter-subject variability. It should be considered in future studies.

Table 2. Classification accuracy (%) of cognitive states under ten-fold cross-validation.

Classifier	ACC	KPA	F1SC	PRC	REC	SPC
LR	82.87 ± 4.29	82.65 ± 5.21	82.96 ± 4.46	87.26 ± 3.97	82.78 ± 4.49	85.13 ± 1.98
KNN	80.37 ± 6.83	79.34 ± 7.19	80.24 ± 5.33	84.75 ± 4.11	80.19 ± 5.67	83.01 ± 2.62
RF	88.96 ± 4.21	87.74 ± 5.31	88.12 ± 5.19	90.32 ± 4.39	87.56 ± 3.07	90.92 ± 1.36
SVM	93.22 ± 3.37	90.76 ± 3.33	93.05 ± 3.27	94.25 ± 2.18	93.46 ± 3.19	94.86 ± 1.29

Table 3. Classification accuracy (%) of cognitive states under one-subject-out cross-validation.

Classifier	ACC	KPA	F1SC	PRC	REC	SPC
LR	71.74 ± 7.54	71.26 ± 7.34	72.21 ± 6.56	75.58 ± 5.26	71.26 ± 6.34	74.55 ± 2.26
KNN	73.69 ± 8.24	72.28 ± 8.23	72.65 ± 7.47	76.57 ± 5.65	72.39 ± 7.21	76.37 ± 3.87
RF	83.69 ± 6.17	82.26 ± 7.32	80.12 ± 7.47	85.32 ± 6.22	81.54 ± 4.28	84.28 ± 2.49
SVM	90.58 ± 5.47	89.43 ± 5.29	90.47 ± 5.48	90.52 ± 4.26	90.04 ± 5.22	91.37 ± 2.43

4 Discussion

4.1 Brain Functional Connectivity Method and Neural Markers

DwPLI is a phase-based bias-corrected connectivity measure derived from the imaginary cross-spectrum, reducing volume conduction effect [16]. Topographic patterns reveal neural insights: IDG shows fronto-temporal integration with centrality in right temporal–parietal and centro-frontal hubs; IDE shifts to bilateral parietal–temporal regions with stable parietal–central hubs; IDR localizes in parietal and fronto-central areas with frontal–parietal and occipital hubs; and RST shows reduced frontal engagement with temporal strength and parietal centrality.

4.2 Computation Complexity

Experiments were conducted on a standard workstation (Intel® Core™ i3, 8 GB RAM, MATLAB R2023b). Feature extraction required $\sim 10.6 \pm 0.6$ s per trial, with inference taking $\sim 85.2 \pm 7.1$ ms per sample. The full pipeline took about 75 min. The framework has the potential to be employed in continuous cognitive state monitoring.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Runco, M.A., Jaeger, G.J.: The Standard definition of creativity. *Creat. Res. J.* **24**(1), 92–96 (2012). <https://doi.org/10.1080/10400419.2012.650092>
2. Cropley, D.H.: Chapter 3 phases: creativity and the design process. In: Cropley, D.H. (ed.) *Creativity in Engineering Internet*, pp. 35–61. Academic (2015)
3. Simon, H.A.: *The Sciences of the Artificial*, 3rd edn. MIT Press (1996)
4. Gero, J.S.: Creativity, emergence and evolution in design. *Knowl.-Based Syst.* **9**(7), 438 (1996)
5. Dorst, K., Cross, N.: Creativity in the design process: co-evolution of problem–solution. *Des. Stud.* **22**(5), 425–437 (2001)
6. Wallas, G.: *The Art of Thought*. Harcourt, Brace and Company (1926)
7. Eindhoven, J.E., Vinacke, W.E.: Creative processes in painting. *J. Gen. Psychol.* **47**(2), 139–164 (1952)
8. Yang, J., Quan, H., Zeng, Y.: Knowledge: the good, the bad, and the ways for designer creativity. *J. Eng. Des.* **33**(12), 945–968 (2022)
9. Zangeneh Soroush, M., Zeng, Y.: EEG-Based exploration of nonlinear brain dynamics in design creativity during a modified TTCT-F Task. *J. Mech. Des.* **25**, 1–50 (2025)
10. Zangeneh Soroush, M., Zeng, Y.: EEG-based design creativity exploration through recurrence quantification analysis. *Design Sci.* **12**, e5 (2026)
11. Soroush, M.Z., Zeng, Y.: EEG-based study of design creativity: a review on research design, experiments, and analysis. *Front. Behav. Neurosci.* **18**, 18 (2024)
12. Sporns, O.: The complex brain: connectivity, dynamics, information. *Trends Cogn. Sci.* **26**(12), 1066–1067 (2022)
13. Nguyen, T.A., Zeng, Y.: A preliminary study of EEG spectrogram of a single subject performing a creativity test. In: *Proceedings of the 2014 International Conference on Innovative Design and Manufacturing (ICIDM)*, pp. 16–21 (2014)

14. Jia, W., Zeng, Y.: EEG signals respond differently to idea generation, idea evolution and evaluation in a loosely controlled creativity experiment. *Sci. Rep.* **11**(1), 2119 (2021)
15. Soroush, M.Z., Zhao, M., Jia, W., Zeng, Y.: Loosely controlled experimental EEG datasets for higher-order cognitions in design and creativity tasks. *Data Brief.* **52**, 109981 (2024)
16. Vinck, M., Oostenveld, R., van, W.M., Battaglia, F., Pennartz, C.M.A.: An improved in-dex of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. *NeuroImage.* **55**(4), 1548–1565 (2011)



RareGraph-AgenticAI: A Multimodal Knowledge Graph and Multi-agent LLM Framework for Rare Disease Evaluation and Gene Prioritization

Quan M. Nguyen^{1,2}  and Kai Wang^{1,3} (✉) 

¹ Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA, USA

wangk@chop.edu

² Bioengineering Graduate Group, University of Pennsylvania, Philadelphia, PA, USA

³ Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA

Abstract. Around the world, hundreds of millions of people live with rare diseases, but for many of them, a clear diagnosis remains out of reach due to phenotypic heterogeneity, fragmented knowledge, and inconsistent documentation. Existing AI approaches either rely on closed-source frontier models with external tool calls—limiting deployment in privacy-sensitive clinical settings—or operate on narrow structured databases that miss the rich clinical context embedded in disease narratives. We present RareGraph-AgenticAI, a two-stage framework that integrates: (1) a comprehensive rare-disease Knowledge Graph (KG) constructed via LLM-driven extraction from the full text of OMIM, Orphanet, and GeneReviews—encompassing 15,280 diseases, 391,237 phenotype annotations with importance, onset, polarity, severity, and frequency tiers, and 254,940 nodes across 17 semantic types; and (2) a multi-agent diagnostic pipeline powered entirely by the open-source Qwen3-8B model, which extracts multimodal patient evidence, applies composite deterministic scoring, and performs a novel LLM-based pairwise reranking step to resolve subtype ambiguity. Evaluated on four RareBench cohorts ($n = 1,122$), the system achieves Top-5 group accuracy up to 56% on LIRICAL and Top-5 up to 57% on MME, outperforming existing methods like DeepRare (with Qwen-3 8B backbone), Phen2Gene, and PhenoSS.

Keywords: LLM · Agent · OMIM · Orphanet · Mondo · GeneReviews · HPO · rare disease · knowledge graph

1 Introduction

Millions of families worldwide are affected by rare diseases, yet over half remain undiagnosed after odysseys exceeding five years [1]. Diagnosis is hampered by heterogeneous phenotypes, dispersed knowledge across OMIM [2], Orphanet [3], GeneReviews [4], MONDO [5], and HPO [6], and the difficulty of interpreting multimodal clinical data in

real time. Critical information—phenotype frequencies, variant mechanisms, differential rules—is often incomplete in structured databases or buried in unstructured narratives.

Recent LLM-based systems show promising results. Recent LLM-based systems show promising results. DeepRare [7] integrates over 40 specialized tools with GPT-4-class models to process heterogeneous clinical inputs and generate ranked diagnostic hypotheses. However, DeepRare requires closed-source APIs and internet-connected tool calls, making it unsuitable for privacy-sensitive clinical environments. Traditional tools—Phen2Gene [8] or PhenoSS [9]—operate deterministically on HPO annotations without leveraging richer clinical context in disease narratives.

We present RareGraph-AgenticAI, contributing: (a) training-free RareGraph, a KG constructed by Gemini 2.5 Pro from unstructured disease narratives, covering 15,280 diseases with 254,940 nodes across 17 types; and (b) AgenticAI, a training-free multi-agent pipeline using only Qwen3-8B [10], featuring information content-driven candidate narrowing that mimics expert reasoning, context-flag annotation for small-model accuracy, and novel LLM-based reranking that captures semantic relationships beyond rigid semantic matching.

2 Methods

2.1 RareGraph: Knowledge Graph Construction

Unlike MONDO [5] which integrates already-structured fields, RareGraph mines the full narrative text of OMIM [2], Orphanet [3], and GeneReviews [4] using Gemini 2.5 Pro and formulate them into KG. We extract: (i) phenotypes with frequency, importance tier (characteristic/supportive/incidental), polarity (present/absent), and onset; (ii) genes and variant mechanisms; (iii) inheritance; (iv) demographics; (v) differential diagnostic rules; (vi) treatments and contraindicated agents; (vii) pathways; (viii) diagnostic testing; (ix) surveillance; and (x) genetic counseling and other useful information. Entities are normalized to HPO [6], MONDO [5], and HGNC via cross-validation and reconciliation. To ensure KG quality, we filtered obsolete entities, rejected unsupported extractions, and deduplicated phenotypes by HPO. We prioritized high-confidence sources during merging, normalized identifiers post-extraction, and applied post-hoc validation to reduce noise, hallucinations, and inconsistent disease mappings.

2.2 AgenticAI: Multi-agent Diagnostic Reasoning Pipeline

The inference pipeline processes multimodal inputs through a coordinated multi-agent architecture powered entirely by Qwen3-8B [10] running locally via vLLM, proceeding through four stages.

Stage 1 — Multi-modal Extraction. Stage 1—Multimodal extraction. Five agents extract: (i) phenotypes with negation/onset/context; (ii) demographics; (iii) family history and inheritance; (iv) genetic testing; (v) gene mentions. A vision agent handles facial images; an optional genomics agent processes VCFs. Unlike HPO-only tools (Phen2Gene, PhenoSS), our pipeline can integrate multimodal evidence for richer patient representations.

Stage 2 — Normalization and Ontology Expansion. Extracted phenotypes are mapped to HPO identifiers via BioLORD-2023 [10] embedding similarity. To handle granularity mismatches, phenotype–disease retrieval optionally expands each HPO term to semantically related neighbors (synonyms, parents, children, siblings) within a configurable graph radius, enabling retrieval when exact matches are unavailable.

Stage 3—IC-driven Candidate Narrowing. Rather than treating all phenotypes equally, scoring mimics expert reasoning by prioritizing high-signal phenotypes. Candidates are scored by: (a) phenotype relevance weighted by HPO information content (IC)—rare phenotypes like ‘ectopia lentis’ contribute far more than common ones like ‘seizure’—combined with KG importance tiers and frequencies; (b) pathognomonic bonuses; (c) co-occurrence pair signals; (d) inheritance consistency; (e) demographic alignment; (f) gene evidence; and (g) exclusion phenotype penalties. This narrows from the most informative phenotypes first, paralleling how clinicians anchor on distinctive features. This produces the *Original Rank*.

Stage 4—LLM-Based Reranking (Novel). Deterministic scoring can miss semantic reasoning clinicians apply: (a) contextual interpretation—“recurrent infections” in Chronic Granulomatous Disease means bacterial/fungal in liver/lungs so respiratory-related diseases should be prioritized; (b) constellation reasoning—“short stature” + “joint hypermobility” + “easy bruising” forms a connective tissue gestalt for Ehlers-Danlos but is discordant for growth hormone deficiency; (c) implicit inference—“inability to extend fingers” signals sclerodactyly without HPO matches; (d) negation context—“no cardiac involvement” contradicts diseases with characteristic cardiomyopathy even when absent from exclusion lists.

Two reranking passes are applied. For subtype pairs sharing a MONDO parent, the LLM receives compact disease cards (hallmark phenotypes, genes, inheritance, exclusion features) and judges fit—producing the *Subtype Rank*. This addresses subtype scattering: diseases like CdLS types 1–5 distribute evidence across variants, each ranking lower than collective support warrants. Candidates are then grouped by MONDO parent and scored at group level—the *Reconciled Rank*.

3 Results

3.1 Knowledge Graph Validation

RareGraph spans 15,280 diseases and 254,940 nodes across 17 types (Fig. 1). Phenotype data covers 11,831 diseases (77.4%), with a median of 25 HPO annotations per disease. Of 391,237 total annotations, 122,288 (31.3%) are LLM-extracted entries absent from structured databases. Cross-source analysis reveals 17.4% of phenotype terms confirmed across all three sources, 16.8% by two, and 65.8% unique to one—demonstrating complementarity. The KG captured 7,046 negated phenotype annotations—features documented as absent—essential for differential diagnosis but rarely encoded in structured databases. Beyond phenotypes, Rare Graph captures clinical management context: inheritance for 9,869 diseases (64.6%), gene associations for 7,854 (21,029 links), demographics for 8,757, differentials for 5,890 (median 4 per disease), treatment for 6,742

across pharmacological, supportive, dietary, surgical, and gene therapy categories, and diagnostic testing for 8,126 diseases

3.2 Agentic AI Performance

We evaluated disease prioritization on the RareBench benchmark [11] using four public cohorts: MME (Matchmaker Exchange; 40 cases, 17 diseases), LIRICAL (370 cases, 252 diseases), HMS (Hannover Medical School; 88 cases, 39 diseases), and RAMEDIS (624 cases, 63 rare metabolic diseases)—totaling 1,122 cases. LLM-based reranking consistently improves Top-5 accuracy, with the most dramatic gain on MME (38% → 57%, +19 points), alongside improvements on LIRICAL (56% → 57%), RAMEDIS (35% → 42%), and HMS (11% → 19%). These gains confirm sub-type scattering and semantic ambiguity are genuine error sources mitigated by LLM reasoning.

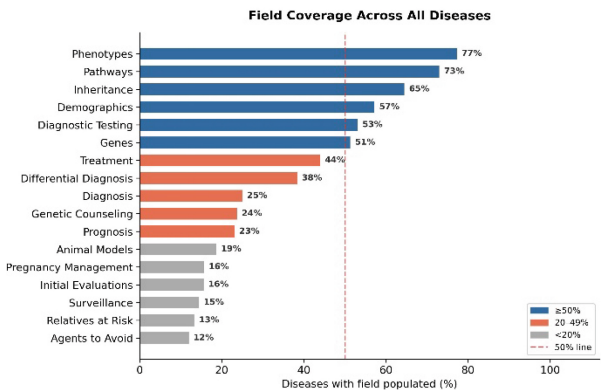


Fig. 1. Field coverage across 15,280 diseases in RareGraph. Six fields exceed 50% coverage; treatment, differentials, and management fields provide actionable context for thousands of diseases.

vs. Phen2Gene and PhenoSS. Both are training-free HPO-only tools. Our system outperforms Phen2Gene on LIRICAL (+14 Top-5), and MME (+43). PhenoSS achieves above 60% Top-5 on MME and LIRICAL but collapses on the largest-size dataset RAMEDIS (19%), revealing dataset-specific fragility. Our IC-driven scoring with multimodal evidence and KG reranking produces more balanced performance across all cohorts—the most important property for clinical deployment.

vs. DeepRare. Running DeepRare with Qwen3-8B (same model as ours), MME Top-5 drops from 90% (GPT-4) [7] to 2%, confirming published performance is driven by frontier model capacity, not entirely pipeline architecture. Under the same model, we outperform DeepRare on LIRICAL (+25) and MME (+53). DeepRare’s advantage on HMS and RAMEDIS with Qwen3-8B may partly reflect bias from Pheno Brain [12]—whose hyperparameters were tuned on 249 benchmark cases overlapping evaluation data.

Our system requires no training and achieves competitive results, showing structured KG grounding can substitute for model scale (Fig. 2).

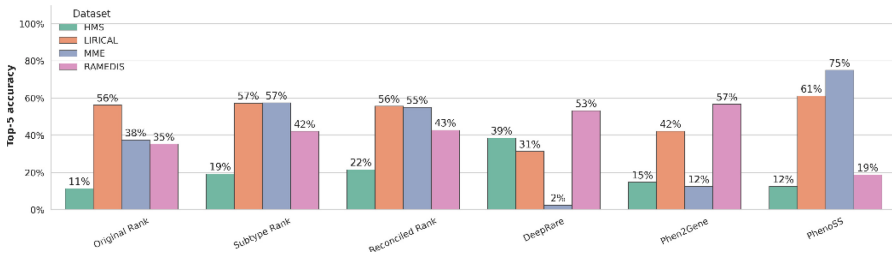


Fig. 2. Top-5@Acc Comparison across all models on RareBench datasets

4 Discussions

RareGraph-AgenticAI demonstrates that structured knowledge combined with a lightweight open-source LLM and targeted reranking delivers meaningful diagnostic prioritization without closed-source APIs, internet access, or training data. Several design choices distinguish our approach. IC-driven narrowing mimics expert reasoning by anchoring on rare, distinctive phenotypes rather than weighting all symptoms equally. Multimodal integration—phenotypes, demographics, family history, genetics, facial images—provides richer representations than HPO-only tools. LLM reranking captures contextual interpretation, constellation reasoning, implicit inference, and negation context that rigid matching cannot resolve. Exclusion phenotypes support differential diagnosis by capturing clinical differences that are rarely encoded in structured databases.

Limitations include: The gap with DeepRare reflects different deployment regimes (8B open-source vs. frontier closed-source with 40+ tools); and stronger open-source models may yield further gains. Future work will pursue prospective clinical validation and integration with exome sequencing pipelines.

Acknowledgments. This study was supported by NIH grants HG013031, OD037960, and the CHOP Research Institute.

Disclosure of Interests The authors have no competing interests.

References

1. The Lancet Global: The landscape for rare diseases in 2024. *The Lancet Glob. Health.* **12**(3), e341 (2024)
2. Hamosh, A., et al.: Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**(Database issue), D514–D517 (2005)

3. Ayme, S.: Orphanet, an information site on rare diseases. *Soins*. **672**, 46–47 (2003)
4. Adam, M.P., et al.: *GeneReviews*, Seattle (WA) (1993)
5. Vasilevsky, N.A., et al.: Mondo: Unifying diseases for the world, by the world. *medRxiv*, 2022.04.13.22273750 (2022)
6. Gargano, M.A., et al.: The Human phenotype ontology in 2024: phenotypes around the world. *Nucleic Acids Res.* **52**(D1), D1333–D1346 (2024)
7. Zhao, W., et al.: An agentic system for rare disease diagnosis with traceable reasoning. *Nature*. **651**(8106), 775–784 (2026)
8. Zhao, M., et al.: Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genom. Bioinform.* **2**(2), lqaa032 (2020)
9. Chen, S., et al.: PhenoSS: Phenotype semantic similarity-based approach for rare disease prediction and patient clustering. *medRxiv*, 2026.02.26.26347219 (2026)
10. Remy, F., Demuynck, K., Demeester, T.: BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *J. Am. Med. Inform. Assoc.* **31**(9), 1844–1855 (2024)
11. Chen, X., et al.: RareBench: Can LLMs Serve as Rare Diseases Specialists? *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2024)
12. Mao, X., et al.: A phenotype-based AI pipeline outperforms human experts in differentially diagnosing rare diseases using EHRs. *NPJ Digit. Med.* **8**(1), 68 (2025)

Author Index

A

Abbas, Ali Abou 7
Abbas-Aghababazadeh, Farnoosh 56
Abd-Alrazaq, Alaa 276
Abidi, Samina 40, 286
Abidi, Syed Sibte Raza 40, 286
Abou-Alwan, Saleh 389
Adhikari, Japp 96
Afrin, Tanzila 222
Ahmed, Arfan 276
Alfred, Myrte de 156
Ali, Dashti A. 167
Ali, Ghadir 101
Alkhateeb, Mais 276
Allen, Rebecca 13
Alliot, Jean-Marc 193
Amigó, Núria 106
Amitrano, Federica 410
Andreoni, Giuseppe 233
Andrić, Marina 204
Anwar, Lia 24
Apriceno, Gianluca 73, 79, 204
Armañanzas, Ruben 56
Arozarena Donelli, Pablo 254
Asaad, Aras T. 167
Asiri, Arwa 7
Atrsaedi, Arash 62
Azhir, Alaleh 415

B

Bacco, Luca 427, 433
Badr, Nabil Georges 1
Bae, Jaewon 383
Bailoni, Tania 73
Bainbridge, Shannon 389
Barbosa, Vanessa 383
Barnes, Lachlan D. 172

Bell, Morris 271
Bellard, Broderick 156
Bellazzi, Riccardo 198, 254, 410
Bermak, Amine 24
Bernard, Gordon R. 96
Bernardi, Alberto 399
Berry, Sameer 362
Bhargava, Vaibhav 383
Bhutani, Gautam 29
Bianchi, Anna 233
Bichindaritz, Isabelle 421
Biffi, Emilia 233
Black, Morgan 281
Blandeau, Mathias 111
Blaylock, Nate 121
Boddu, Vijaya Laxmi 127
Bolpagni, Marco 145
Bonfigli, Agnese 433
Boone, Ella 45
Boyd, Andrew D. 383
Boyd, Lauren 383
Brown, Ali 281
Brueckner, Raymond 121
Bulgakov, Vitaly 18
Buzzard, Paula 362

C

Cabitza, Federico 297
Cafazzo, Joseph 156
Cara, Irene 292
Casazza, Simone 145
Castiglia, Stefano Filippo 239, 244
Cecchini, Matthew 281
Chakraborty, Souvik 182
Chan, Adrian D. C. 389
Chan, Kinwei Arnold 341
Chang, Ming-Ching 324

Chaudhary, Akhil 127
 Cheng, Jingya 415
 Chierici, Marco 106
 Chinnapapannagari, Charitha 271
 Chojnicki, Michał 351
 Cichosz, Simon Lebech 198
 Clarahan, Brianna 383
 Coccia, Armando 410
 Cochez, Michael 399
 Coelho, Pedro 330
 Colelli Riano, Federico 410
 Colley, Helen 265
 Costabello, Luca 399
 Costantini, Simone 233
 Craven, Catherine K. 383
 Cruise, Danielle 35
 Curcin, Vasa 303
 Cvijanović, Dane 106

D

D'Addio, Giovanni 410
 Dagliati, Arianna 198
 Dahal, Rajashree 383
 Darvishvand, Fateme Golivand 377
 Davidović, Aleksandar 106
 Dawson, Rachel 383
 Daza, Daniel 399
 De Carli, Simone 145
 De Icco, Roberto 239, 244
 Deigmöller, Jörg 50
 Di Camillo, Barbara 106
 Di Eugenio, Barbara 383
 Di Giacinto, Alberto Maria 433
 Do, Richard K. G. 167
 Dragišić, Dalibor 106
 Dragoni, Mauro 73, 79, 133, 139, 145, 204, 297
 Dünninghaus, Florian 313

E

Eccher, Claudio 139, 145
 Eggert, Julian 50
 El Demellawy, Dina 389
 Ellis, Robert 362
 Elmi, Sayda 259, 271
 Engelen, Marc 399
 Erffmeier, Karla 313

Estiri, Hossein 415

F

Farhan, Hania 96
 Farina, Mirko 319
 Fee, Kevin 210
 Fernandes, Gabriel W. 172
 Franzin, Mattia 139, 145
 Furmidge, Rachel 265

G

Gan, Tao 116
 Garbe, Jakob 313
 García-Galindo, Alberto 56
 Gatta, Roberto 319
 Gazala, Varun 259
 Geleijnse, Gijs 292
 Gelmini, Kevin 204
 Giglione, Claudia 106
 Giordani, Flavia 433
 Glinkowski, Wojciech 351
 Godoy, Rafael Silva 346
 González-Gomariz, José 56
 Goodman, Michael 362
 Gosselin, Bernard 111
 Greenspan, Michael 222
 Gross, Ben J. 308
 Gross, Douglas 372
 Guadagno, Elena 45, 395
 Gueret, Christophe 399
 Guillamet, Guillem Hernández 216
 Guimarães, Amanda 330
 Gürses, Ömer Alperen 357

H

Haibe-Kains, Benjamin 56
 Haidar, Hazar 1
 Hallum, Luke E. 172
 Haque, Nihal 101
 Hasanuzzaman, Mohammed 210
 Hassan, Arshad 24
 Haughey, Kristi 383
 Hayeso, Helen Haile 116
 Heinrich, Steffen 50
 Hewryk, Oresta S. I. 68
 Hitzel, Danielle 383
 Holub, Olena 13

I

Ibáñez, Beatriz López 216
 Ibn Sultan, Rafi 335
 Iglesias-Grau, Josep 228

J

Jaana, Mirou 35
 Jacob, Justin 222
 Jager, Sven 62
 Jain, Suneil 210
 Jarnagin, William R. 167
 Jasionowska-Skop, Magdalena 351
 Jaspers, Yorrick 399
 Jehangir, Basra 187
 Jeong, Will Dowon 281
 Jiang, Keyuan 96
 Jurek-Loughrey, Anna 210
 Jurman, Giuseppe 106

K

Kafaie, Somayeh 127
 Kanaan, Dina 156
 Kantelhardt, Jan W. 313
 Kapan Tunçer, Nazife 357
 Karacan, Baris 383
 Kardassis, Dimitris 106
 Karstens, Amanda 383
 Katsimpokis, Dimitris 292
 Kemp, Stephan 399
 Kennedy, Duska 101
 Kersting, Kristian 377
 Khan, Arijit 198
 Khan, Faiq Ahmad 24
 Khetani, Mary 383
 Khodaei, Afsoon 389
 Kim, Edward 367
 Kim, Joshua 335
 Knake, Lindsey 383
 Kotula, Leszek 421
 Kreuzthaler, Markus 249
 Kugic, Amila 249
 Kumar, Vagish 182
 Kureshi, Nelofar 40
 Kwon, Namhee 121

L

Lai, Yong Jun 222
 Lang, Zi-Qiang 265

Latifi, Saif 303
 Leckman-Westin, Emily 324
 Lens, Clément 161
 Lerma, Bianca 79
 Li, Weizi Vicky 187
 Lin, Sherman 281
 Liu, Xinyi Celine 156
 Liut, Michael 101
 Loban, Katya 45
 Lohani, Raima 156
 Lonseko, Zenebe Markos 116
 Lopez, Karen Dunn 383
 López-De-Castro, Marcos 56

M

Madan, Anmol 362
 Magni, Paolo 106
 Mahmoudi, Ramzi 177
 Majed, Bilal 161
 Majouni, Sheida 286
 Malagola, Michele 319
 Malovini, Alberto 410
 Maltese, Girolamo 427
 Mancas, Matei 111
 Mannai, Zayneb 177
 Marinello, Elena 106
 Marquis, Pierre 161
 Massicotte, Daniel 439
 Matella, Malwina 265
 Mathur, Saurabh 377
 Mboowa, Gerald 85
 McCabe, Erin 372
 Mehdizavareh, Hadi 198
 Melo, Paulo Henrique Moreira 395
 Merone, Mario 427, 433
 Metcalf, Joshua 13
 Mgbow, Ujunwa 335
 Milanov, Marko 106
 Minville, Vincent 193
 Mohabir, Vina 156
 Mollanji, Roland 101
 Moraes, Matheus 330
 Moturu, Sai 362
 Mouysset, Sandrine 193
 Munjas, Jelena 106
 Muñoz Monjas, Aída 341
 Murdoch, Craig 265
 Murphy, Ross G. 210

N

Nadkarni, Abhijit 404
 Nakonechnyi, Alex 13
 Nascimento, Marcos 330
 Nataraj, Andrew 372
 Natarajan, Sriraam 377
 Nemati, Shamim 308
 Nguyen, Hong Duc 90
 Nguyen, Quan M. 445
 Nguyen, Quang Vu 313
 Ni, Kewei 56
 Nicora, Giovanna 198, 254
 Njoud, Haneen 62
 Noei, Shahryar 106
 Noleto, Diandra 330
 Noleto, Lariza 330
 Nyangoma, Dalson 85

O

O'Connell, Henry 121
 Ocker, Felix 50
 Olivato, Francesco 319
 Omheni, Nizar 177
 Özsoy, İsmail 357

P

Palchuk, Matvey B. 341
 Papalia, Giuseppe Francesco 427
 Papalia, Rocco 427
 Paravatham, Venkata Sai Pavan Raghavendra 271
 Parde, Natalie 383
 Pareek, Niketh 372
 Parimbelli, Enea 254
 Patel, M.D. Radha 308
 Paterson, Mary 265
 Pecchia, Leandro 427, 433
 Perez-Rey, David 341
 Perović-Blagojević, Iva 106
 Petitgout, Janine 383
 Petković, Aleksa 106
 Pham, Quynh 156
 Phan, Duc Tri 90
 Philip, Sherene 362
 Piperno, Ruben 427
 Pla Planas, Albert 62

Poenaru, Dan 45, 395
 Poette, Michael 193
 Poličar, Pavlin G. 151
 Portinale, Luigi 254
 Przelaskowski, Artur 351
 Puterman-Salzman, Lily 228

R

Rahimi, Samira Abbasgholizadeh 228
 Rajabi, Enayat 127
 Raman, Lakshmi 377
 Ramsis, M.D. Mattheus 308
 Ranavolo, Alberto 239, 244
 Rancati, Simone 254
 Rao, Nini 116
 Ratković, Tamara 106
 Raza, Shaina 156
 Reeder, Phillip 377
 Ribeiro, Carlos Henrique Costa 346
 Rivankar, Dinesh 404
 Rizvi, Syeda 276
 Rizzi, Silvia 145
 Romano, Joseph D. 68
 Rubio Ruiz, David 341
 Ruiz, Daniel 193
 Russo, Domenico 319
 Russo, Fabrizio 427

S

Saafi, Yasmin 177
 Saha, Shumit 156
 Samper Vila, Joan 216
 Sanna, Leonardo 133, 139, 145
 Santangelo, Gabriele 198
 Sarbhangia, Rohitaswa 29
 Schiegl, Adrian 249
 Schmid, Thomas 313
 Schulz, Stefan 249
 Schut, Martijn 399
 Seeliger, Katja 313
 Sejdic, Ervin 101
 Seper, Lukas 249
 Serrao, Mariano 239, 244
 Shah, Kalash 29
 Shin, Phil 101
 Shirvankar, Chetan 228

Shrestha, Manil 367
 Siavashpour, Mahvash 372
 Simpson, Amber L. 167
 Skhiri Gabbouj, Feten 111
 Skinner, Michael 377
 Smirnov, Pavel 50
 Snehana, J. 29
 Solinas, Erica 133
 Soni, Ameet 377
 Sopić, Miron 106
 Soroush, Morteza Zangeneh 439
 Spalińska, Antonina 351
 Spaliński, Krzysztof 351
 Spellman, Emily 383
 Stanimirović, Dalibor 151
 Storm, Fabio 233
 Subramanian, Vinod 121
 Supraja, S. 90
 Sweeney, Ellen 40

T

Tabassum, Aliya 276
 Tabia, Karim 161
 Tamim, Rayyan 35
 Tannetta, Dionne 187
 Tassorelli, Cristina 239, 244
 Tennankore, Karthik 286
 Thind, Kundan 335
 Tian, Jiazi 415
 Tiefenbacher, Alistair 249
 Tipton, Angie 383
 Tojal da Silva, Israel 346
 Torne, Pol 106
 Trabassi, Dante 239, 244
 Tsatsanis, Christos 106
 Tseng, Yu-Shan 383
 Turchin, Alexander 18

V

Vadalà, Gianluca 427
 van Laarhoven, Hanneke W. M. 292
 Vardon, Fanny 193
 Velmovitsky, Pedro Elkind 156
 Verhoeven, Rob H. A. 292
 Verlekar, Tanmay Tulsidas 404
 Vignato, Julie 383
 Vinson, Amanda 286
 Vladimirov-Sopić, Sandra 106
 von Bosse, Alexa 50
 Vukmirović, Luka 106

W

Walden, Andrew 187
 Walker, Dawn 265
 Wallard, Laura 111
 Wallon, Romain 161
 Wang, Kai 445
 Wang, Rui 324
 Wasserman, Jason K. 281
 Weis, Mackenzie 383
 Wiljer, David 156

Y

Yoosuf, Shehel 276

Z

Zaiane, Osmar 372
 Zhang, Jerry J. 172
 Zheng, Kai 308
 Zhu, Dongxiao 335
 Zhu, Wei-Ping 439
 Zomahoun, Hervé Tchala Vignon 228
 Zupan, Blaž 151