Hasso Plattner Institute, University of Potsdam
Digital Health & Machine Learning

Benchmark to Bedside: Building Modular Multimodal Machine Learning Infrastructure for Healthcare at Scale

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor rerum naturalium
(Dr. rer. nat.)

in Computer Science

to the Digital Engineering Faculty at the
Hasso Plattner Institute, University of Potsdam

by
Robin Philippus van de Water, M.Sc.

in Potsdam, Germany, on January 22, 2026.

License

This work is licensed under a Creative Commons License: Attribution-NonCommercial-ShareAlike 4.0 International. This does not apply to quoted content from other authors. To view this license, visit: https://creativecommons.org/licenses/by-nc-sa/4.0/

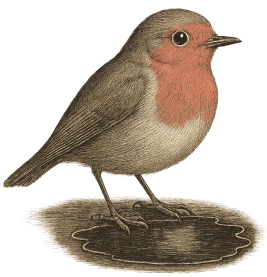
Reviewers

Prof. Dr. Christoph Lippert
Hasso Plattner Institute, University of Potsdam, Germany

Prof. Dr. Padhraic Smyth
University of California, Irvine, United States of America

Prof. Dr. Mykola Pechenizkiy
Eindhoven University of Technology, Netherlands

Ex Libris



Abstract

Healthcare systems generate vast amounts of comprehensive data daily; yet, translating this information into robust machine learning (ML) models that improve patient outcomes remains a challenge. The demand for scalable, practical, and problem-driven systems that use this heterogeneous data to its full capacity is growing as healthcare complexity and costs rise alongside the aging population.

The first contribution establishes reproducibility as a cornerstone of clinical ML research. Despite the increasing availability of open-access intensive care unit (ICU) datasets and publications, the field lacks standardized approaches to data preprocessing, task definition, and evaluation. Researchers often employ custom datasets, ambiguous definitions, and dataset-specific pipelines, making it impossible to compare models meaningfully or build upon previous work. We address this need with Yet Another ICU Benchmark (YAIB) — a modular, extensible framework designed to support reproducible clinical ML experiments across multiple datasets — and ReciPies — a concise, human-readable, and reproducible way to declare, execute, and share preprocessing pipelines. YAIB standardizes the entire modeling workflow from raw data harmonization through clinical concept definition to model fitting and evaluation, thereby reducing implementation overhead and enabling transparent, comparable research. Critically, the framework demonstrates that dataset selection, cohort definitions, and preprocessing choices often have a more substantial impact on prediction performance than the choice of model architecture itself.

The second contribution demonstrates the value of prospective clinical trials for specific healthcare Artificial Intelligence (AI) applications. As an interdisciplinary team of computer scientists and surgeons, we created a dataset that combines non-intrusive wearable vital signals with 1,285 patients, 270,603 hours of telemetric vital signs, and more than 8 million clinical measurements. We aimed to detect subtle endpoints in the general ward, which is often undermonitored and underrepresented in ML research. We developed a comprehensive machine-learning-driven early warning system that unifies preoperative parameters, intraoperative measurements, ICU data, clinical text, and continuous high-resolution vital signs from wearable devices deployed on surgical wards. The approach detects surgical site infections, pancreatic fistulas, hematomas, and bile leaks in the general ward. High-resolution sensor data provide improvements of up to 8% in the area

under the receiver operating curve (AUROC) and 109% in the area under the precision-recall curve (AUPRC), underscoring the importance of capturing dynamic physiological complexity as an early marker of clinical deterioration. Our findings suggest the need for hybrid wards with custom AI early warning systems that lower diagnostic costs and care burden.

The third contribution addresses the need for a standardized healthcare data infrastructure. To truly scale clinical AI across multiple institutions and datasets, the field requires tools that enable collaboration, reproducibility, and interoperability. This work contributes to the Medical Event Data Standard (MEDS) — a lightweight, ML-first data schema designed to standardize the representation of electronic health record (EHR) data for AI use. MEDS provides a flexible, low-level abstraction layer that integrates with existing EHR processing tools and modeling pipelines, reducing the friction in clinical machine learning research. In this framework, we have developed multiple interoperable tools for data transformation to MEDS, exploration, and harmonization. The open-source ecosystem dramatically improves reproducible research output. Moreover, this thesis includes an open-source benchmark of 11 clinical tasks across eight open-access datasets and a multicenter health system comprising 12 million patients, demonstrating the value of reproducible, multi-institutional research efforts. Moreover, we benchmark state-of-the-art EHR foundation models trained on 6.8 million patients in various data regimes. Low data regimes demonstrate promising results for these models. MEDS enables scalable, privacy-preserving model evaluation across datasets and health systems, which should drive compounding iterative improvements to health AI across institutional boundaries.

Throughout this work, a pattern emerges: successful clinical machine learning requires not only algorithmic innovation across the domain but also systematic attention to data quality, experimental methodology, and software infrastructure. The envisioned goal is to improve and prolong patient lives through trustworthy and equitable healthcare AI for the many.

Zusammenfassung (Abstract in German)

Gesundheitssysteme generieren täglich umfangreiche Datenmengen; dennoch bleibt die Umwandlung dieser Informationen in robuste Modelle des maschinellen Lernens (ML), die zu besseren Gesundheitsergebnissen führen, eine Herausforderung. Die Nachfrage nach skalierbaren, praktischen und problemgetriebenen Systemen, die diese heterogenen medizinischen Daten vollständig nutzen, wächst, da die Komplexität und die Kosten des Gesundheitswesens bei einer alternden Bevölkerung ansteigen.

Der erste Beitrag dieser Dissertation etabliert die Reproduzierbarkeit als Grundpfeiler der klinischen ML-Forschung. Trotz der zunehmenden Verfügbarkeit von frei zugänglichen Intensivstations-Datensätzen und Publikationen fehlt dem Feld ein standardisierter Ansatz für Datenvorverarbeitung, Aufgabendefinition und Evaluierung. Forscher verwenden häufig unterschiedliche benutzerdefinierte Datensätze, Definitionen und datensatzspezifische Verarbeitungspipelines, wodurch es unmöglich wird, Modelle aussagekräftig zu vergleichen oder auf früheren Arbeiten aufzubauen. Wir adressieren diesen Bedarf mit *Yet Another ICU Benchmark* (YAIB) — einem modularen, erweiterbaren Framework zur Unterstützung reproduzierbarer klinischer ML-Experimente über mehrere Datensätze hinweg — und *ReciPies* — einer menschenlesbaren und reproduzierbaren Möglichkeit, die Datenpipeline zu deklarieren, auszuführen und zu teilen. Wir standardisierten den gesamten Modellierungsworkflow von der Rohdatenharmonisierung über die Definition klinischer Konzepte bis hin zum Training und zur Evaluierung, wodurch der Implementierungsaufwand reduziert und transparente, vergleichbare Forschung ermöglicht wird. Essenziell ist, dass das Framework zeigt, dass die Datensatzauswahl, die Kohortendefinitionen und die Vorverarbeitungsentscheidungen oft einen wesentlich größeren Einfluss auf die Vorhersageleistung haben als die Wahl der Modellarchitektur selbst.

Der zweite Beitrag demonstriert den Wert prospektiver klinischer Studien für spezifische Anwendungen der künstlichen Intelligenz (KI) im Gesundheitswesen. Als interdisziplinäres Team von Informatikern und Chirurgen erstellten wir einen Datensatz, der nichtinvasive tragbare Vitalparameter mit 1.285 Patienten, 270.603 Stunden telemetrischer Vitalzeichen und mehr als 8 Millionen klinischen Messungen kombiniert. Unser Ziel war es, subtile klinische Endpunkte auf der Pflegestation zu erkennen, die häufig unterbeobachtet und in der ML-Forschung unterrepräsentiert sind. Wir entwickelten ein umfassendes, durch maschinelles Lernen gesteuertes Früh-

warnsystem, das präoperative Parameter, intraoperative Messungen, Intensivstationsdaten, klinische Texte sowie kontinuierliche hochauflösende Vitalzeichen von tragbaren Geräten auf chirurgischen Stationen vereint. Der Ansatz erkennt Wundinfektionen, Pankreasfissteln, Hämatome und Gallengangsaustritte in der allgemeinen Station. Hochauflösende Sensordaten führen zu Verbesserungen von bis zu 8 % in der *Area Under the Receiver Operating Characteristic* (AUROC) und 109 % in der *Area Under the Precision-Recall Curve* (AUPRC). Dies unterstreicht die Bedeutung der Erfassung dynamischer physiologischer Komplexität als frühes Zeichen klinischer Verschlechterung. Unsere Ergebnisse deuten auf die Notwendigkeit von Hybrid-Stationen mit maßgeschneiderten KI-Frühwarnsystemen hin; diese könnten zudem die diagnostischen Kosten und den Pflegeaufwand senken.

Der dritte Beitrag zeigt die Notwendigkeit einer standardisierten Gesundheitsdateninfrastruktur. Um klinische KI über mehrere Institutionen und Datensätze hinweg zu skalieren, benötigt das Feld Werkzeuge, die Zusammenarbeit, Reproduzierbarkeit und Interoperabilität ermöglichen. Diese Arbeit trägt zum *Medical Event Data Standard* (MEDS) bei — einem leichtgewichtigen, ML-zentrierten Datenschema zur Standardisierung der Darstellung elektronischer Patientenakten (ePA) für die KI-Nutzung. MEDS bietet eine flexible Low-Level-Abstraktionsschicht, die sich in bestehende ePA-tools und Modellierungspipelines integriert und die Komplexität in der klinischen ML-Forschung reduziert. In diesem *Framework* haben wir mehrere operative Werkzeuge für die Datentransformation, die Datenexploration sowie die Datenharmonisierung entwickelt. Das *Open-Source*-Ökosystem verbessert die Reproduzierbarkeit von Forschungsergebnissen erheblich. Darüber hinaus enthält diese Dissertation einen Open-Source-Benchmark mit elf klinischen Aufgaben, basierend auf acht frei zugänglichen Datensätzen und einem multizentrischen Versorgungssystem mit zwölf Millionen Patienten, der den Wert reproduzierbarer, institutionenübergreifender Forschungsbemühungen demonstriert. Darüber hinaus vergleichen wir hochmoderne ePA-Grundlagenmodelle, die auf 6,8 Millionen Patienten in verschiedenen Datenregimen trainiert wurden. Ein „Low-Data-Regime“ ergibt aussichtsvolle Ergebnisse für diese Modelle. MEDS ermöglicht skalierbare, datenschutzwahrende Modellbewertung über Datensätze und Gesundheitssysteme hinweg, was iterative Verbesserungen der Gesundheits-KI über Institutionengrenzen hinweg vorantreiben sollte.

Im Laufe dieser Arbeit ergibt sich ein Muster: Erfolgreiches maschinelles Lernen in der klinischen Praxis erfordert nicht nur algorithmische Innovationen in der Domäne, sondern auch systematische Aufmerksamkeit für Datenqualität, experimentelle Methodik und Softwareinfrastruktur. Wir hoffen, dass durch die ermöglichte vertrauenswürdige und gerechte Gesundheits-KI das Leben vieler Patienten verbessert und verlängert wird.

Samenvatting (Abstract in Dutch)

Zorgsystemen produceren dagelijks aanzienlijke hoeveelheden complex gegevens. De centrale uitdaging ligt echter in het omzetten van deze gegevens in effectieve *machine learning* (*ML*)-modellen die patientuitkomsten daadwerkelijk optimaliseren. Naarmate de complexiteit en kosten van de gezondheidszorg toenemen en de bevolking vergrijst, groeit de vraag naar schaalbare, praktische systemen die heterogene data optimal benutten. De eerste bijdrage beschouwt reproduceerbaarheid als een pijler van klinisch *ML* onderzoek. Ondanks de toenemende beschikbaarheid van vrij toegankelijke *intensive care unit* (*ICU*)-*datasets* en publicaties, ontbreken gestandaardiseerde oplossingen voor *dataprocessing*, taakdefinitie en evaluatie. Onderzoekers werken met aangepaste *datasets*, ambigue definities en *dataset*-specifieke *pipelines*, waardoor zinvolle modellenvergelijking en voortzetting op eerder werk onmogelijk wordt. Gedreven door dit problem introduceren we *Yet Another ICU Benchmark* (*YAIB*) — een modulair, uitbreidbaar *framework* dat is ontworpen om reproduceerbare klinische *ML* experimenten over meerdere *datasets* te ondersteunen — en *ReciPies* — een beknopte, leesbare en reproduceerbare manier om *preprocessing pipelines* te declareren, uit te voeren en te delen. *YAIB* standaardiseert de volledige modelleringsworkflow, van dataharmonisatie en klinische conceptdefinitie tot het trainen en evalueren van modellen. Dit verlaagt de implementatielast en maakt transparent, onderling vergelijkbaar onderzoek mogelijk. Een belangrijke bevinding: de keuze van dataset, cohortdefinities en *preprocessing* hebben vaak groter effect op voorspellende prestaties dan de keuze van modelarchitectuur zelf.

De tweede bijdrage toont de waarde aan van prospectieve klinische studies voor specifieke gezondheidszorg *Kunstmatige Intelligentie* (*KI*)-toepassingen. Als interdisciplinair team van computerwetenschappers en chirurgen creëerden wij een unieke *dataset* van 1.285 patiënten met 270.603 uur telemetrische vitale signalen en meer dan 8 miljoen klinische metingen, waarin niet-intrusieve draagbare vitale signalen gecombineerd zijn met ziekenhuisgegevens. Ons doel was het detecteren van subtiele klinische uitkomsten op de verpleegafdeling, – een setting waar weinig *ML* onderzoek naar is gedaan, mede dankzij het gebrek aan data. We ontwikkelden een veelomvattend *early warning* systeem met *ML* dat preoperatieve parameters, intraoperatieve metingen, *ICU* data, klinische tekst en continue hoogfrequente vitale signalen van draagbare apparaten integreert. Deze benadering detecteert *surgical site infections* (postoperatieve wondinfecties), pancreasfis-

tels, hematomen en gallekkages op de verpleegafdeling. Hoge resolutie sensordata leveren verbeteringen op tot 8% in de *area under the receiver operating curve* (*AUROC*) en 109% in de *area under the precision recall curve* (*AUPRC*), wat het belang van dynamische fysiologische complexiteit als vroeg signaal van klinische achteruitgang benadrukt. Deze bevindingen ondersteunden de ontwikkeling van hybride afdelingen met op maat gemaakte KI *early warning* systemen, met potentiële reductie van diagnostische kosten en zorgbelasting.

De derde bijdrage richt zich op schaalbare data-architectuur voor echte interoperabiliteit tussen instellingen. Echte opschaling van klinische KI over meerdere centra vereist tools die samenwerking, reproduceerbaarheid en interoperabiliteit mogelijk maken. Dit werk draagt bij aan de *Medical Event Data Standard* (*MEDS*) — een lichte, *ML*-eerste dataschema die EHR-*data* voor KI toepassingen standaardiseert. *MEDS* biedt een flexibele, laagniveau abstractielaag die integreert met bestaande *EHR processing tools* en *modeling pipelines*, waardoor frictie in klinisch *ML* onderzoek afneemt. Binnen dit *framework* ontwikkelden we interoperabele tools voor datatransformatie naar *MEDS*, data-exploratie en dataharmonisatie. Het *open source* ecosysteem verbetert de reproduceerbaarheid van onderzoeksresultaten aanzienlijk. Daarnaast omvat dit proefschrift een *open source benchmark* van 11 klinische voorspellingsproblemen over acht *open-access* datasets en een multicenter zorgsysteem met 12 miljoen patiënten, wat de waarde van reproduceerbaar, multi-institutioneel onderzoek aantoont. Verder trainden we *state of the art EHR foundation* modellen op 6,8 miljoen patiënten in verschillende *dataregimes*. *Low dataregimes* laten veelbelovende resultaten zien voor deze modellen. *MEDS* maakt schaalbare, privacybeschermende modelevaluatie over *datasets* en zorgsystemen heen mogelijk, wat iteratieve, cumulatieve verbeteringen in gezondheids-AI over institutionele grenzen heen stimuleert.

Door dit werk heen is een rode lijn te herkennen: succesvolle klinische *machine learning* vereist niet alleen algoritmische innovatie binnen het domein, maar ook systematische aandacht voor datakwaliteit, experimentele methodologie en software infrastructuur. Het doel is het verbeteren van de levenskwaliteit en het verlengen van levens door middel van betrouwbare en rechtvaardige KI in de gezondheidszorg voor iedereen.

Acknowledgments

Doing a PhD can feel great and meaningful or like pointless suffering depending on the moment; it is usually the people that make the difference. While never easy, I would like to express my gratitude to all those who supported me, including my family, friends, and colleagues; people who have made this possible.

Thank you to my supervisor Prof. Dr. Christoph Lippert for your support and nerdy conversations about running et al. Thank you to Prof. Dr. Bert Arnrich for the funding support. Thank you to Prof. Dr. Bernhard Renard and Prof. Dr. Tillman Rabl for the advice. Thank you to Prof. Dr. Padhraic Smyth for your stimulating feedback, conversations, and reviewing my thesis. Thank you to Prof. Dr. Mykola Pechenizkiy for reviewing my thesis. Thank you Prof. Dr. Dr. h. c. mult. Lothar H. Wieler, Prof. Dr. Felix Naumann, and Prof. Dr. Gerard de Melo for forming the rest of my committee.

Thank you to Stephan Mandt for the eye-opening research stay in Irvine, amongst whom: Sam, Justus, Ashwin, Sakshi, Tamanna, J.B, Catarina, Kushagra, Elliot, Markelle, Gavin, Harry, and Federica. I'll cherish the memories of exploring Yosemite and Laguna Beach, not to mention cycling like "crazy Europeans" on the beach highways with Thomas.

Thank you to Dr. Max, Dr. Axel (Maxel), and Daniela for the countless hours of meetings and discussions about everything and nothing in-between life-saving surgeries at the Charité. Thanks to Daniela, Youssef, and Hendrik for being great master thesis students. Thanks to all the MEDS folks, specifically Prof. Dr. Matthew McDermott for the countless stimulating conversations and insights in this great field. Thanks to the folks, scientific and supportive, at Mount Sinai in New York City, specifically Prof. Dr. Eugenia Alleva for your support and insights. Thanks to Dr. Harry Freitas da Cruz for the great advice. Thanks to Lena, Cornelia, Katharina, and Edit for all the administrative support throughout these years. Thank you to Dr. Patrick Rockenschaub for sharing your great ideas, frequent mentorship, and advice.

Thank you Monica and Stefan for your boundless support, friendship, and banter. Thank you to my great office mates over the years: Fabian, Pascal, and Dr. Nico for the productive procrastination. Thank you to Tamara and Florian for adopting me and providing me with another supportive, even if mostly empty, environment. Thanks to Dr. Orhan for emotional support and always agreeing with me no matter the subject. Thank you to all the

great people at HPI, of whom I hopefully didn't forget (m)any: Christoph, Berry, Charlotte, Kristina, Holly, Sidratul, Dr. Bjarne, Dr. Jonas, Dr. Justin, Dr. Arpita, Dr. Lin; Aadil, Dr. Elmien; Dr. Susanne, Liz, Francesca, J.P., Ferdous; Dr. Masoumeh, Dr. Rebecca, Benjamin, Dr. Eshant, Jana, Juliana, Matthias, Dr. Remo, Alex, Noel, Wei-Cheng (Wes), Hadya, Thomas, Hani, Anees, Hannah.

Thank you to Dr. Loïs for leading by example. Thank you to my mom and dad for your support and encouragement throughout my life.

Epigraph

All of old. Nothing else ever. Ever tried. Ever failed. No matter. Try again. Fail again. Fail better.
'Worstward Ho' by Samuel Beckett

Dedicated to Felixje (2007-2024)

Contents

Abstract in English **iii**

Abstract in German **vi**

Abstract in Dutch **viii**

Table of Contents **xi**

List of Publications **xiv**

List of Acronyms **xxiii**

List of Figures **xxiv**

List of Tables **xxviii**

List of Code Listings **xxxii**

1 Introduction **1**
1.1 Motivation 1
1.2 Thesis Outline 2
1.3 Research Contribution 3
1.4 Additional Contributions 10
1.5 Supervision and Teaching 12

2 Background **13**
2.1 A Terminology of Health Data 13
2.2 A Brief Overview of Temporal Modeling in Healthcare 16
2.3 Challenges in Machine Learning for Healthcare 23

3 Establishing Reproducibility in Open-access EHR Prediction Modeling **27**
3.1 Motivation 27
3.2 Related Work 28
3.3 Benchmark Design 35
3.4 Preprocessing and Feature Engineering with ReciPies 43
3.5 Experiments 46
3.6 Using YAIB as an Experimental ML Framework 49
3.7 Discussion 53
3.8 Conclusion 54

4 Creating a Continuous Multimodal AI with Wearable Vital Signs to Predict Postoperative Complications in the General Ward **55**
4.1 Motivation . . . 55
4.2 Related Work . . . 56
4.3 Retrospective Multicenter Prediction of Postoperative Complications . . 59
4.4 Prospective Clinical Study Design and Population . . . 69
4.5 Data Processing and ML Design . . . 76
4.6 Results . . . 84
4.7 Discussion . . . 93
4.8 Conclusion . . . 96

5 Building a Modular Tooling Ecosystem for High-capacity Collaborative Clinical ML **99**
5.1 Motivation . . . 100
5.2 Background and Related Work . . . 101
5.3 Data Conversion to MEDS . . . 109
5.4 Data Exploration with MEDS-Inspect . . . 112
5.5 Converting Clinical Coding Vocabularies with HarmonizEHR . . . 117
5.6 Scaling Clinical Model Benchmarking across Datasets and Tasks with MEDS-DEV . . . 119
5.7 Foundation Model Training and Evaluation on the Mount Sinai Health System . . . 127
5.8 Discussion . . . 141

6 Discussion and Outlook **147**
6.1 Ethical Considerations . . . 147
6.2 Future Work . . . 148
6.3 Recommendations . . . 150
6.4 Final Remarks . . . 151

Bibliography **153**

A Disclosure **183**
A.1 Funding and Affiliations . . . 183
A.2 Scientific Computing . . . 183
A.3 Use of Generative AI . . . 184

B Supplements Chapter 2 **185**
B.1 Literature Review . . . 185

C Supplements Chapter 3 **187**
C.1 Extended Results . . . 187
C.2 Datasets . . . 190
C.3 Outcome Definitions . . . 194
C.4 Usage of YAIB and Implementation . . . 198

C.5 Extending YAIB . . . 202
C.6 Experimental Setup and Reproducibility . . . 211
C.7 Hyperparameters . . . 214
C.8 Machine Learning Reproducibility Checklist . . . 223

D Supplements Chapter 4 **227**
D.1 Retrospective Multicenter Prediction of Postoperative Complications . . . 227
D.2 Detailed Inclusion Criteria . . . 233
D.3 Detailed Data Characteristics . . . 235
D.4 Model Evaluation . . . 237
D.5 Detailed Results . . . 240
D.6 TRIPOD+AI Statement for Clinical Prediction Models . . . 242

E Supplements Chapter 5 **267**
E.1 Data Description . . . 267
E.2 Task Definitions . . . 268
E.3 Hyperparameters . . . 272
E.4 Extended Results . . . 279
E.5 Dimensionality Reduction of Foundation Model Embeddings . . . 291

F German Layperson Summary **301**

List of Publications

Published

- **Robin P. van de Water**, Hendrik Schmidt, and Patrick Rockenschaub. "ReciPies: A Lightweight Data Transformation Pipeline for Reproducible ML". in: *Journal of Open Source Software* 11.117 (2026), p. 9261. doi: 10.21105/joss.09261 [VSR26]
- Matthias Kirchler, Matteo Ferro, Veronica Lorenzini, **Robin P. van de Water**, Christoph Lippert, and Andrea Ganna. "Large Language Models Improve Transferability of Electronic Health Record-Based Predictions across Countries and Coding Systems". In: *Accepted to npj Digital Medicine (2026)* [Kir+25b]
- Katharina Alefs, Susanne Ibing, Pia Francesca Rissom, Jan Carlo Schmid, Arkadiusz Kwasigroch, **Robin van de Water**, Bernhard Y. Renard, and Eugenia Alleva. "Towards Foundation Model-Based Propensity Score Matching from Electronic Health Records". In: Machine Learning for Health Symposium 2025. Sept. 2025. [Ale+25]
- Christoph Riepe, **Robin van de Water**, Axel Winter, Bjarne Pfitzner, Lara Faraj, Robert Ahlborn, Maximilian Schulze, Daniela Zuluaga, Christian Schineis, Katharina Beyer, Johann Pratschke, Bert Arnrich, Igor M. Sauer, and Max M. Maurer. "90-Day Mortality Prediction in Elective Visceral Surgery Using Machine Learning: A Retrospective Multicenter Development, Validation, and Comparison Study". In: *International Journal of Surgery* 111.6 (June 2025), p. 3742. doi: 10.1097/JS9.0000000000002372. [Rie+25b]
- Bjarne Pfitzner, Max M. Maurer, Axel Winter, Christoph Riepe, Igor M. Sauer, **Robin van de Water**, Christian Denecke, Johann Pratschke, and Bert Arnrich. "Differentially-Private Federated Learning with Non-IID Data For Surgical Risk Prediction." In: *International Journal of Semantic Computing 19.3 (2025)*. [Pfi+25]
- Axel Winter, Bjarne Pfitzner, **Robin P. van de Water**, Lara Faraj, Christoph Riepe, Wolf-Heinrich Hahn, Felix Krenzien, Christian Schineis, Thomas Malinka, and Wenzel Schöning. "Overcoming the Data Barrier: Transfer Learning for 90-Day Mortality Prediction in General Surgery–a Retrospective Multicenter Development and Comparison Study". In: *International Journal of Surgery* (2025), pp. 10–1097. [Win+25]
- **Robin Philippus van de Water**, Ethan Steinberg, Michael Wornow, Patrick Rockenschaub, and Matthew McDermott. MIMIC-IV Demo Data in the Medical Event Data Standard (MEDS). 2025. doi: 10.13026/T2Y8-EA41. [Van+25d]

- **Robin P. Van de Water**. "Scaling Up Clinical ML from Datasets to Entire Health Systems through the MEDS Ecosystem". In: *SBHD 2025 International Conference on Systems Biology of Human Diseases*. Berlin, June 2025 [Van25c]

- **Robin van de Water**, Hendrik Nils Aurel Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. "Yet Another ICU Benchmark: A Flexible Multi-Center Framework for Clinical ML". in: *Proceedings of The Twelfth International Conference on Learning Representations*. May 2024. doi: 10.48550/arXiv.2306.05109. [Van+24a]

- **MEDS Working Group**: (alphabetized) Bert Arnrich, Edward Choi, Jason A. Fries, Matthew B. A. McDermott, Jungwoo Oh, Tom J. Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, **Robin van de Water**. "Medical Event Data Standard (MEDS): Facilitating Machine Learning for Health". In: *ICLR 2024 Workshop on Learning from Time Series* for Health. 2024. [Gro24]

- Matthew B. A. McDermott, Aleksia Kolo, Chao Pang, Edward Choi, Ethan Steinberg, Hyewon Jeong, Jack Gallifant, Jason Alan Fries, Jeffrey N. Chiang, Jungwoo Oh, Justin Xu, Kamilė Stankevičiūtė, Kiril Vadimovic Klein, Mikkel Fruelund Odgaard, Nassim Oufattole, Patrick Rockenschaub, Pawel Renc, **Robin van de Water**, Shalmali Joshi, Simon Austin Lee, Teya Bergamaschi, Tom Pollard, Vincent Jeanselme, Nigam Shah, Michael Wornow, Aparajita Kashyap, Xinzhou Jiang, Yanwei Li, Young Sang Choi, Yuta Kobayashi, and Ryan King. "MEDS Decentralized, Extensible Validation (MEDS-DEV) Benchmark: Establishing Reproducibility and Comparability in ML for Health". In: *Machine Learning for Health (ML4H)*, 15-16 December 2024, Vancouver, Canada (Nov. 2024). [McD+24]

- Matthew B. A. McDermott, Justin Xu, Teya S. Bergamaschi, Hyewon Jeong, Simon A. Lee, Nassim Oufattole, Patrick Rockenschaub, Kamilė Stankevičiūtė, Ethan Steinberg, Jimeng Sun, Robin P. Van De Water, Michael Wornow, John Wu, and Zhenbang Wu. "MEDS: Building Models and Tools in a Reproducible Health AI Ecosystem". In: *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining* V.2. Toronto ON Canada: ACM, Aug. 2025, pp. 6243–6244. isbn: 979-8-4007-1454-2. doi: 10 . 1145/ 3711896 . 3737608. [McD+25d]

- Max M. Maurer, Bjarne Pfitzner, **Robin P. van de Water**, Lara Faraj, Christoph Riepe, Daniela Zuluaga, Felix Krenzien, Nathanael Raschzok, Robert Siegel, Christian Schineis, Bert Arnrich, Katharina Beyer, Johann Pratschke, Igor M. Sauer, and Axel Winter. "Privacy Preserving Federated Learning for 90-Day Mortality Prediction in Colorectal Surgery: A Multicenter Retrospective Development and Comparison Study". In: *International Journal of Surgery* (London, England) (Aug. 2025). issn: 1743-9159. doi: 10.1097/JS9.0000000000003084 [Mau+25]

- **Robin van de Water**, Axel Winter, Max M. Maurer, Felix August Treykorn, Daniela Zuluaga, Bjarne Pfitzner, Igor M. Sauer, and Bert Arnrich. "Combining Hospital-grade Clinical Data and Wearable Vital Sign Monitoring to Predict Surgical Complications". In: *ICLR 2024 Workshop on Learning from Time Series For Health*. Mar. 2024. [Van+24c]

- **Robin van de Water**, Axel Winter, Max M. Maurer, Felix August Treykorn, Bjarne Pfitzner, Igor M. Sauer, and Bert Arnrich. "Combining Time Series Modali- ties to Create Endpoint-driven Patient Records". In: *ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR): Harnessing Momentum for Science*. Mar. 2024. [Van+24b]

- Axel Winter, **Robin P. van de Water**, Bjarne Pfitzner, Marius Ibach, Christoph Riepe, Robert Ahlborn, Lara Faraj, Felix Krenzien, Eva M. Dobrindt, and Jonas Raakow. "Enhancing Preoperative Outcome Prediction: A Comparative Retrospective Case–Control Study on Machine Learning versus the International Esodata Study Group Risk Model for Predicting 90-Day Mortality in Oncologic Esophagectomy". In: *Cancers 16.17* (2024), p. 3000. [Win+24]

- Bjarne Pfitzner, Max M. Maurer, Axel Winter, Christoph Riepe, Igor M. Sauer, **Robin van de Water**, and Bert Arnrich. "Differentially-Private Federated Learning with Non-IID Data for Surgical Risk Prediction". In: 2024 *IEEE First International Conference on Artificial Intelligence for Medicine, Health and Care (AIMHC)*. IEEE, 2024, pp. 120–129. [Pfi+24]

- Orhan Konak, **Robin van de Water**, Valentin Döring, Tobias Fiedler, Lucas Liebe, Leander Masopust, Kirill Postnov, Franz Sauerwald, Felix Treykorn, and Alexander Wischmann. "HARE: Unifying the Human Activity Recognition Engineering Workflow". In: *Sensors 23.23 (2023)*, p. 9571.

- Orhan Konak, Alexander Wischmann, **Robin van de Water**, and Bert Arnrich. "A Real-time Human Pose Estimation Approach for Optimal Sensor Placement in Sensor-based Human Activity Recognition". In: *Proceedings of the 8th International Workshop on Sensor-Based Activity Recognition and Artificial Intelligence*. Lübeck Germany: ACM, Sept. 2023, pp. 1–6. isbn: 979-8-4007-0816-9. doi: 10.1145/3615834.3615848.

- **Robin van de Water**, Francesco Ventura, Zoi Kaoudi, Jorge-Arnulfo Quiane-Ruiz, and Volker Markl. "Farming Your ML-based Query Optimizer's Food". In: 2022 IEEE 38th International Conference on Data Engineering (ICDE). Kuala Lumpur, Malaysia: IEEE, May 2022, pp. 3186–3189. isbn: 978-1-6654-0883-7. doi: 10.1109 / ICDE53745 . 2022 . 00294. Best demonstration paper award

- **Robin van de Water**, Francesco Ventura, Z Kaoudi, J Quiane-Ruiz, and Volker Markl. "Farm Your ML-based Query Optimizer's Food!–Human-Guided Training Data Generation–". In: *Conference on Innovative Data Systems Research (CIDR). 2022*

- Lientje Maas, Mathan Geurtsen, Florian Nouwt, Stefan F. Schouten, **Robin van de Water**, Sandra Van Dulmen, Fabiano Dalpiaz, Kees Van Deemter, and Sjaak Brinkkemper. "The Care2Report System: Automated Medical Reporting as an Integrated Solution to Reduce Administrative Burden in Healthcare." In: *Hawaii International Conference on System Sciences*. 2020, pp. 1–10.

Under submission

- Matthew McDermott, Ethan Steinberg, Jason Fries, **Robin P. van de Water**, Chao Pang, Patrick Rockenschaub, Pawel Renc, Jungwoo Oh, Kamilė Stankevičiūtė, Justin Xu, Tom Joseph Pollard, Nassim Oufattole, Michael Wornow, Teya Bergamaschi, Hyewon Jeong, Simon Lee, Vincent Jeanselme, Kiril Klein, Mikkel Odgaard, Maria Elkjær Montgomery, Arkadiusz Sitek, Mads Nielsen, Jeffrey Chiang, Noa Dagan, Isaac Kohane, Shalmali Joshi, Edward Choi, and Nigam Shah. "MEDS: A Simple, Interoperable Data Standard and Ecosystem for Health AI Research". In: *Under review at New England Journal Of Medicine (NEJM) AI (2026)* [McD+25c]

- Jan Carlo Schmid, Susanne Ibing, Stefan Kalabakov, Katharina Alefs, **Robin P. van de Water**, Arkadiusz Kwasigroch, Eugenia Alleva, Bert Arnrich, Bernhard Y. Renard, and Maia Kayal. "Expert-Defined, Chart-Reviewed Features Outperform EHR Foundation Models for Predicting Crohn's-like Disease of the Pouch". In: *Under review at The American Journal of Gastroenterology (2026)* [Sch+25b]

Preprints (preparing for submission)

- **Robin P. van de Water**, Axel Winter, Daniela Zuluaga Lotero, Bjarne Pfitzner, Lara Faraj, Bert Arnrich, Patrick Rockenschaub, Wenzel Schoning, Thomas Malinka, Christian Denecke, Johann Pratschke, Igor M. Sauer, and Max M. Maurer. Continuous Multimodal AI with Wearable Vital Signs Predicts Postoperative Complications in the General Ward. Nov. 2025. doi: 10.1101/2025.11.25.25340950.

- Alisher Turubayev, Anna Shopova, Fabian Lange, Mahmut Kamalak, Paul Mattes, Victoria Ayvasky, Bert Arnrich, Bjarne Pfitzner, and **Robin P. van de Water**. Closing Gaps: An Imputation Analysis of ICU Vital Signs. Oct. 2025. doi: 10.48550/arXiv.2510.24217. arXiv: 2510.24217 [cs].

Manuscripts in preparation

- **Robin P. van de Water**, Katharina Marie Alefs, Susanne Ibing, Stefan Kalabakov, Aadil Rasheed, Pia Francesca Rissom, Jan Carlo Schmid, Arkadiusz Kwasigroch, Eugenia Alleva, and Christoph Lippert. "FLAMES: EHR Foundation Models as a Resource for Health Systems". 2025

- **Robin P. van de Water**, Christoph Riepe, Bjarne Pfitzner, Lara Faraj, Daniela Zuluaga, Christian Schineis, Katharina Beyer, Johann Pratschke, Igor Sauer, Axel Winter, and Max Maurer. "Machine Learning for 30-Day Mortality Prediction for High-Risk General Emergency Surgery". In: In preparation (2025)

- Wouter van Amsterdam, Michael Kamp, Rajesh Ranganath, **Robin P. van de Water**, Florian Markowetz, Evangelia Christodoulou, Yamuna Krishnamurthy, Christoph Lippert, Jeff Clark, Julia E. Vogt, Raul Santos-Rodriguez, Thomas Gärtner, Gilbert Koch, and Brett Beaulieu-Jones. "Missing Incentives, Missing Oversight: The Challenges of AI Monitoring in Clinical Practice". 2025

Seminars

- Raul Santos-Rodriguez, Kacper Sokol, Julia E. Vogt, and Sven Wellmann and **all authors of the abstracts in this report**. "From Research to Certification with Data-Driven Medical Decision Support Systems (Dagstuhl Seminar 25052)". In: Dagstuhl Reports 15.1 (2025). Ed. by Raul Santos- Rodriguez, Kacper Sokol, Julia E. Vogt, and Sven Wellmann, pp. 201–220. issn: 2192-5283. doi: 10.4230/DagRep.15.1.201. [San+25b]
- Digital Health Partnership: Poster presentation in October 2024, May 2025, and November 2025.

List of Acronyms

ACES Automated Cohort Extraction System for Event Streams. 114, 121, 123, 127–129

CASHEWS Clinical ASsist Heuristic Early Warning System. 84–89, 91, 237

MEDS-DEV MEDS Decentralized External Validation. xxv, 121, 123–126, 129, 147, 151, 268, 271, 279, 281, 282

YAIB *Yet Another ICU Benchmark*. xxiv, 2, 4, 8, 28–43, 46, 51, 53, 54, 84, 100, 102, 147, 149, 151, 198, 199, 201, 202, 204–208, 210, 211, 214, 224

AI artificial intelligence. 1, 13, 56, 147

AKI Acute Kidney Injury. 28, 30, 31, 45, 46, 48

ARIMA Autoregressive Integrated Moving Average. 19

ASA American Society of Anesthesiologists. 5, 67

ATT Artificial Time Tokens. 22, 106

AUMCdb Amsterdam University Medical Center Database. 16, 31, 33–36, 47–50, 52, 53, 103, 121–123, 126, 187, 188, 191, 197, 211

AUPRC Area Under the Precision Recall Curve. 41, 48, 50–52, 61, 67, 85, 124, 125, 132, 133, 187, 232, 279, 281, 282

AUROC Area Under the Receiver Operating Characteristic. 41, 48–52, 61, 67, 85, 93, 124–126, 132, 133, 187, 232, 279, 281, 282

BERT Bidirectional Encoder Representations from Transformers. 105, 107, 118, 125

CASSANDRA Clinical ASSist AND aleRt Algorithms. 2, 59, 69, 93, 149

CCI Charlson Comorbidity Index. 5, 61–68

CD Clavien-Dindo classification. 71, 72, 262

CDM Common Data Model. 8, 101, 111, 112, 117, 127

CPU central processing unit. 77, 142

CSDI Conditional Score-based Diffusion models for Imputation. 207, 208

CT Computed Tomography. 72, 95

CUMC Columbia University Medical Center. 100, 105, 110, 111, 119, 123, 127, 130

CVSM continuous vital sign monitoring. 59

CW Corsano CardioWatch. 73, 75, 78

DL deep learning. 21, 48, 53, 206, 209

EBCL Event-based Contrastive Learning. 22, 106, 107

ECG electrocardiogram. 73, 77

EHR Electronic Health Record. 3, 6, 14–17, 22, 28, 57, 100, 102–104, 106, 110–112, 127, 129–131, 138, 140–144, 147, 152

eICU eICU Collaborative Research Database. 16, 29, 31, 33–36, 47–49, 52, 102, 104, 122, 123, 126, 187–189, 191, 196, 197, 211

EN elastic net. 47, 211, 222

ETL Extract, Transform, Load. 7, 9, 100, 103, 104, 109, 111, 119, 123, 127, 128, 147

EWS Early Warning System. 2, 56, 58, 59, 80, 96

FHIR Fast Healthcare Interoperability Resources. 9, 100, 101, 104, 108, 109, 113

FM foundation model. 22, 100, 101, 104–108, 117, 125, 127, 130, 132–138, 140, 152

FOSS Free and Open Source Software. 7, 113, 142, 148

GBT Gradient Boosted Trees. 94

GDPR General Data Protection Regulation. 24, 82, 126, 147

GLM Generalized Linear Model. 20

GPT Generative Pretrained Transformer. 105, 107

GPU graphics processing unit. 21, 94, 95

GRU Gated Recurrent Unit. 31, 47, 50–52, 82, 211, 218

GTC GreenTEG CORE. 73, 75, 78

GW general ward. 56, 69

HIPAA Health Insurance Portability and Accountability Act. 24, 126, 147, 183

HiRID High Time Resolution ICU Dataset. 16, 31, 33–36, 47–49, 52, 104, 109, 121–123, 125, 126, 187, 188, 191, 196, 197

HitL Human-in-the-Loop. 11, 119

HRV heart rate variability. 78, 91

i.i.d. independent and identically distributed. 23

I2B2 Informatics for Integrating Biology and the Bedside. 101, 108, 109

ICD International Classification of Diseases. 102, 118

ICU intensive care unit. 3, 4, 15, 27, 28, 31, 33, 56, 57, 85, 88, 99, 100, 102, 123

INSPIRE INformative Surgical Patient dataset for Innovative Research Environment. 16, 104, 109, 121–123, 126

IQR Interquartile Range. 35, 48, 78, 92

JSON JavaScript Object Notation. 14, 40, 41, 43

k-NN k-Nearest Neighbors. 20, 61

KDIGO Kidney Disease Improving Global Outcomes. 30, 35, 195, 205

KF Kidney Function. 28, 31, 45, 46, 48, 189, 190, 222

LGBM Light Gradient Boosting Machine. 46, 47, 50, 205, 211, 223

LLM Large Language Model. 1, 20, 21, 79, 107, 129, 138, 140, 148, 152

LOINC Logical Observation Identifiers Names and Codes. 103, 111, 118, 119

LoS Length of Stay. 28, 31, 45, 46, 48, 188, 189

LR logistic regression. 46, 47, 62, 211, 222

LSTM Long Short-Term Memory. 21, 31, 47, 199, 200, 211, 219

MAE Mean Absolute Error. 48, 52, 189

MCC Matthews correlation coefficient. 61, 232

MEDS Medical Event Data Standard. 2, 6, 8, 9, 99–104, 107–109, 111, 142–147

MIMIC Medical Information Mart for Intensive Care. 31, 33

MIMIC-III Medical Information Mart for Intensive Care III. 16, 29–31, 33, 34, 47, 102, 188, 189

MIMIC-IV Medical Information Mart for Intensive Care IV. 16, 29, 31, 35, 36, 47–52, 102, 105, 113, 122, 123, 126, 187, 188, 191, 196, 197

ML machine learning. 1, 13, 15, 27–29, 34, 46, 48, 50, 53, 56, 57, 59, 69, 102, 108, 147, 187, 205, 209

ML4H machine learning for healthcare. 1, 13, 14, 24, 100, 146

MLM Masked language modeling. 22, 106, 107

MOTOR Many Outcome Time Oriented Representations. 107, 127, 133, 134, 136, 137

MRI Magnetic Resonance Imaging. 72, 73

MSDW Mount Sinai Data Warehouse. 110, 267

MSHS Mount Sinai Health System. 7, 109–117, 119, 121–123, 126, 127, 129, 130, 132–135, 138–141, 143, 146, 147, 149

NEWS National Early Warning Score. 58

NLP natural language processing. 21, 86, 100, 105, 107, 138

NTP Next token prediction. 22, 106, 107, 140

NWICU Northwestern ICU Database. 16, 101, 105, 116, 122, 123, 126

OHDSI Observational Health Data Sciences and Informatics. 101, 113

OMOP Observational Medical Outcomes Partnership. 8, 9, 100, 101, 103, 104, 108–112, 117, 119, 127, 128, 267

PCA Principal Component Analysis. 137, 291, 293–299

PCORnet Patient-Centered Outcomes Research Network. 101, 109

PDMS Patient Data Management System. 73, 103, 129

POPF Postoperative Pancreatic Fistula. 72

PPG photoplethysmography. 8, 73, 77, 78, 84, 86

RCT Randomized Controlled Trial. 15, 20

REWS Remote Early Warning Score. 85, 87, 89, 91, 92, 95

RNN Recurrent Neural Network. 21, 31, 211

RoPE Rotary Position Embedding. 22, 106, 107

RWD Real-world Data. 14, 101, 149

RxNorm Normalized names for clinical drugs. 103, 111

SHAP SHapley Additive exPlanations. 42, 82, 83, 89, 91

SICdb Salzburg Intensive Care Database. 16, 105, 119, 122, 123, 126, 144, 202–204

SM Stanford Medicine. 110, 111, 119, 123, 130, 138–140

SNOMED CT Systematized Nomenclature of Medicine – Clinical Terms. 102, 111, 118

SOP Sentence order prediction. 22, 106

SSI-III deep organ space surgical site infections. 5, 56, 71, 72, 150

SSL Self-supervised learning. 22, 106

SVM Support Vector Machine. 20

t-SNE t-distributed Stochastic Neighbor Embedding. 79, 80, 83, 91, 118, 137, 291, 293–299

TCN Temporal Convolutional Network. 47, 82, 212, 220

TD Temporal Denoising Objective. 22, 106

TF-IDF Term Frequency-Inverse Document Frequency. 118

TRIPOD Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis. 84

TTE Time to event. 22, 106, 107

UMAP Uniform Manifold Approximation and Projection. 79, 80, 137, 291, 293–299

UMLS Unified Medical Language System. 103, 118, 119

WandB Weights and Biases. 41, 202

YAML Yet Another Markup Language. 43, 111, 121, 123

List of Figures

2.1 A timeline of a patient's electronic health record (EHR) events. 17

3.1 Exclusion criteria applied to the base cohort for all datasets. 36
3.2 Schematic overview of the Yet Another ICU Benchmark (YAIB) pipeline. . 37
3.3 A schematic overview of RECIPIES' architecture. 45
3.4 A code snippet demonstrating RECIPIES' usage. 45
3.5 Performance comparison of different sepsis definitions on MIMIC-IV. . . 51
3.6 Performance in AUPRC and AUROC of GRU on ICU mortality when trained on one dataset (row) and evaluated on all others (columns). . . . 52
3.7 Performance in MAE of Transformer on Length of Stay when trained on one dataset (row) and evaluated on all others (columns). 52
3.8 Fine-tuning an eICU model for ICU mortality prediction on HiRID. . . . 53

4.1 Schematic illustration of the study concept divided into four different phases. 60
4.2 Performance of the XGBoost model for 90-day mortality prediction. . . . 64
4.3 Feature importance analysis for 90-day mortality prediction using SHAP values. 66
4.4 Exclusion criteria for the CASSANDRA study 70
4.5 Data collection and processing pipeline 73
4.6 Distribution of subjects by surgical procedures over the study period . . 74
4.7 Age distribution of the study cohort subjects 76
4.8 Comparison of the dynamics of ECG and PPG signals. 77
4.9 Clinical notes embedded and visualized by their proximity to the endpoint 79
4.10 Medication texts embedded and visualized by their proximity to the endpoint after applying dimensionality reduction techniques. 80
4.11 Our data segmentation and aggregation pipeline 86
4.12 Experiments to determine model architecture and the amount of selected features 87
4.13 Model performance comparison 88

4.14 Modality ablation and prediction performance curves 89
4.15 Explaining model feature contributions 90
4.16 Implementation of *CASHEWS+* 91
4.17 Ablation experiments for segment length 92

5.1 OMOP CDM v5.4 data schema 112
5.2 MEDS-Inspect main interface overview 114
5.3 Distribution of the number of unique codes per subject in the MSHS MEDS dataset 115
5.4 The most frequent code view in the MSHS MEDS dataset 115
5.5 MEDS-Inspect subject timeline view for NWICU 116
5.6 Histogram of the value of diastolic blood pressure measurements in the MSHS MEDS dataset 116
5.7 Searching for sepsis associated codes in the NWICU MEDS dataset 116
5.8 The frequency of codes grouped by coding vocabularies for the MSHS MEDS dataset 117
5.9 A t-SNE visualization of the unique codes in datasets converted to MEDS 118
5.10 Data characteristics across MEDS datasets. 120
5.11 Performance metrics for the samples weighted benchmarking MEDS-DEV compatible models 124
5.12 AUROC and AUPRC aggregated by dataset performance metrics for the samples weighted benchmarking *MEDS-DEV* compatible models 124
5.13 AUROC and AUPRC aggregated by task performance metrics for the samples weighted benchmarking *MEDS-DEV* compatible models 125
5.14 Aggregated by task performance metrics for the samples weighted benchmarking *MEDS-DEV* compatible models 125
5.15 Foundation model training in the context of the Mount Sinai Health System128
5.16 Unique code count distribution for the MEDS MSHS dataset 129
5.17 Comparing FMs at multiple centers based on the same tokenization procedure for the same tokenizer size. 131
5.18 Tokenized dataset specifics for the MSHS dataset. 132
5.19 Linear probing results on the MEDS MSHS dataset using FM embeddings 133
5.20 Linear probing results on the MEDS MSHS dataset using FM embeddings (continued) 134
5.21 Aggregated results for linear probing with FMs 135
5.22 t-SNE plots of FM embeddings for the Leukocytosis task 136
5.23 t-SNE plots of FM embeddings for the Hyperkalemia task 137
5.24 Event irregularity distribution for the SM and MSHS datasets 139

5.25 N-gram repetition rates for EHR data compared to high-quality natural language text . . . 140
5.26 Zipf's law comparison of medical codes and natural language corpora. . . 141
5.27 Architecture for creating a low-barrier Mount Sinai AI ecosystem. 145

C.1 Performance in Mean Absolute Error (MAE) of the Transformer model on *Length of Stay (LoS)*. . . . 189
C.2 AUPRC for fine-tuning an eICU GRU model for ICU mortality prediction on HiRID. . . . 189
C.3 Exclusion criteria applied to the base cohort. . . . 191
C.4 Additional exclusion criteria applied for the classification tasks. 193
C.5 Additional exclusion criteria applied for the regression tasks. 194
C.6 Experiment definition schematic for YAIB. . . . 199
C.7 The attention mechanism of Conditional Score-based Diffusion models for Imputation (CSDI) adapted from [Tas+21]. . . . 208
C.8 Schematic overview of 5 times repeated 5-fold cross-validation. 212

D.1 Schematic representation of the inclusion and exclusion criteria. 228

E.1 Performance metrics for the subjects weighted benchmarking *MEDS-DEV* compatible models . . . 279
E.2 Performance metrics for the subjects weighted benchmarking MEDS-DEV compatible models. . . . 280
E.3 Performance metrics for the subjects weighted benchmarking MEDS-DEV compatible models. . . . 281
E.4 Aggregated performance metrics for the samples weighted benchmarking *MEDS-DEV* compatible models . . . 281
E.5 Aggregated by task performance metrics for the subjects weighted benchmarking *MEDS-DEV* compatible models . . . 282
E.6 Binary accuracy and calibration error aggregated by task for the subjects weighted benchmarking *MEDS-DEV* compatible models . . . 282
E.7 Binary accuracy and calibration error aggregated by task for the samples weighted benchmarking *MEDS-DEV* compatible models . . . 282
E.8 Linear probing results on the MSHS dataset using foundation model embeddings. . . . 283
E.9 Linear probing results on the MSHS dataset using foundation model embeddings (continued). . . . 284
E.10 Aggregated results for linear probing with foundation models. 285
E.11 Linear probing results in binary accuracy and calibration error on the MSHS dataset using foundation model embeddings . . . 286

E.12 Linear probing results in binary accuracy and calibration error on the MSHS dataset using foundation model embeddings (continued) 287
E.13 Linear probing calibration error and brier score 288
E.14 Linear probing binary accuracy and calibration error (continued) 289
E.15 Aggregated linear probing results by task with foundation models 290
E.16 Dimensionality reduction for hypoglycemia 292
E.17 Dimensionality reduction for elevated creatinine 293
E.18 Dimensionality reduction for hyperkalemia 294
E.19 Dimensionality reduction for metabolic acidosis 295
E.20 Dimensionality reduction for anemia 296
E.21 Dimensionality reduction for leukocytosis 297
E.22 Dimensionality reduction for thrombocytopenia 298
E.23 Dimensionality reduction for hypotension 299

List of Tables

1.1 Overview of datasets with Medical Event Data Standard (MEDS) Extract, Transform, Load (ETL) support 9

2.1 Healthcare data categories, their contents, the applications of machine learning (ML) within each category, and example open-access datasets. . 15
2.2 Overview of open-access medical datasets composed primarily of EHR data. 16
2.3 Overview of data modalities available in these popular open-access medical datasets composed primarily of EHR data. 17
2.4 Overview of foundation models for EHR data. 22

3.1 Comparison of existing benchmarks and *YAIB* on ICU data, ordered by publication date. 31
3.2 Key characteristics of openly accessible ICU datasets. 34
3.3 Characteristics of 1) the included datasets and 2) the task cohorts. 35
3.4 Prediction task overview. 46
3.5 Baseline performance on the classification tasks. 48
3.6 Baseline performance on the regression tasks. 48
3.7 ICU mortality prediction on HiRID with ($> 24h$) and without ($> 30h$) possibility of causal leakage. 49
3.8 AUROC performance comparison of including static data for ICU mortality prediction. 50
3.9 Comparing MIMIC-IV sepsis cohorts according to three different definitions. 51
3.10 Sepsis prediction on MIMIC-IV for different definitions of sepsis. 51

4.2 Model performance for the multicenter retrospective analysis for both endpoints following external validation. 64
4.3 Organ-specific subgroup analysis for 90-day and 30-day mortality prediction. 65
4.4 Patient journey statistics 73

4.5 Overview of aggregated features and sampling frequency for wearable devices . . . 75
4.6 Description of the collected wearable data . . . 78
4.8 The number of features in each of the most granular modalities. . . . 81
4.9 The modified Remote Early Warning Score (REWS) . . . 82
4.10 Description of the collected data . . . 84

5.1 Overview of known datasets with MEDS ETL support . . . 104
5.2 Overview of foundation models and FM benchmarks for EHR data. . . . 106
5.3 Demographic distribution of the Mount Sinai Health System (MSHS), the Columbia Irving University Medical Center (CUMC), EHRSHOT, and the Stanford Medicine (SM) data itself. . . . 111
5.4 Demographic distribution of MSHS data across Train, Tuning, Held Out splits . . . 113
5.5 Tasks defined across datasets over each split of each dataset. . . . 122
5.6 Fit statistics for Zipf's law analysis in Figure 5.26. . . . 139

C.1 Comparing the use of dynamic feature generation (FG) to the baseline of ICU mortality prediction, AUROC (↑). . . . 187
C.2 Comparing the use of dynamic feature generation (FG) to the baseline of ICU mortality prediction, AUPRC (↑). . . . 187
C.3 Baseline AUPRC (↑) performance comparison of including static data for ICU Mortality Prediction. . . . 188
C.4 Baseline AUROC (↑), AUPRC (↑) performance of the included ML algorithms on the demo cohorts. . . . 188
C.5 Baseline performance on the regression tasks. . . . 189
C.6 Clinical concepts used as input to the prediction models. . . . 191
C.6 Clinical concepts used as input to the prediction models (continued) . . . 192
C.7 Staging of AKI according to Kidney Disease Improving Global Outcomes (KDIGO). . . . 195
C.8 Options for the `train` command. . . . 201
C.9 Options for the `evaluate` command. . . . 202
C.10 General arguments for the use of *YAIB*. . . . 202
C.11 Average estimated duration of training tasks with YAIB. . . . 211
C.12 Model hyperparameters for deep learning models. . . . 215
C.13 Model hyperparameters for machine learning models. . . . 216
C.14 Chosen hyperparameters for Gated Recurrent Unit (GRU). . . . 218
C.15 Chosen hyperparameters for Long Short-Term Memory (LSTM). . . . 219
C.16 Chosen hyperparameters for Temporal Convolutional Network (TCN). . . . 220

C.17 Chosen hyperparameters for Transformer. 221
C.18 Chosen hyperparameters for logistic regression (LR). 222
C.19 Chosen hyperparameters for elastic net (EN). 222
C.20 Chosen hyperparameters for Light Gradient Boosting Machine (LGBM). . 223

D.1 Eliminated features during preprocessing due to completeness below 50% 229
D.2 Statistical analysis of the 8 patient characteristics 229
D.3 Statistical analysis of the 37 pre-existing conditions 230
D.4 Statistical analysis of the 17 intervention characteristics and statistical comparison of these between the two included surgical centers. 230
D.5 Statistical analysis of the 16 laboratory values and statistical comparison of these between the two included surgical centers. 232
D.8 Patient inclusion flowchart. 233
D.9 Patient complication groups. 233
D.10 Major organ system groups included in the study. 235
D.11 Numeric target statistics for the prospective cohort 235
D.14 Lab values and their counts 235
D.15 ICU observations and their amounts 237
D.16 ICU scores and their amounts 237
D.12 Binary target statistics for the prospective cohort 238
D.13 Categorical target statistics for the prospective cohort 239
D.17 The p-values for the AUROC of leaving out specific data modalities by comparing eight model configurations using the Friedman test. 241
D.18 Horizon ablation results. 241
D.19 Main results. 241
D.20 Model variation results. 241
D.21 Feature ablation results. 241
D.22 Wearable feature ablation results. 241
D.23 Patient journey described over stay intervals. 253

E.1 Description of the MSHS dataset 267
E.2 Hyperparameters for MEDS-Tab configuration. 273
E.3 Hyperparameters for GenHPF model configuration. 274
E.4 Hyperparameters for CEHR-BERT pretraining. 275
E.5 Hyperparameters for CEHR-BERT finetuning. 276
E.6 Hyperparameters for CEHR-XGPT model pretraining. 277
E.7 Hyperparameters for MOTOR (FEMR) model configuration. 278

List of Code Listings

5.1 **The core MEDS schema as of version 0.4.0** . . . 109
C.1 **Example preprocessing pipeline structure.** . . . 198
C.2 **Example configuration for a task definition configuration.** We identify 4 main categories: the prediction type (classification or regression), the deep learning loss, the dataset variables, and preprocessing settings, and the cross-validation settings. Note that the dataset is fully configurable here. . . . 200
C.3 **Example configuration for hyperparameters of LSTM.** Tunable hyperparameter ranges can be floating points, integers, and categorical values. 201
C.4 **Running YAIB.** We train the LSTM model on the MIMIC-III demo dataset for the Mortality24 task. . . . 201
C.5 **ID and table configuration for the Salzburg Intensive Care Database Database in JSON.** . . . 203
C.6 **Helper functions that calculate the origin and temporal relation between events in the Salzburg Intensive Care Database.** . . . 205
C.7 **JSON definition for rate of potassium chloride across included datasets.** 206
C.8 **Example ML model definition.** We create a Random Forest classifier that implements the default *YAIB* prediction model interface . . . 206
C.9 **Example DL model definition.** We use the inbuilt Torch RNN layers to built a Recurrent Neural Network. . . . 207
C.10 **Temporal Fusion Transformer model definition.** Note that this implementation is similar to existing code and can use existing methods and the original dataloader of *YAIB* which avoids code duplication. . . . 208
C.11 **Temporal Fusion Transformer parameter configuration.** . . . 209
C.12 **Implementing the CSDI architecture in YAIB**. Note that our implementation is very similar to the original github repository, which demonstrates the flexibility of implementing new models in YAIB. . . . 209
C.13 **Implementing JSD using SciPy in YAIB.** In this case we used the Ignite interface, but users can also choose to extend from the TorchMetric or SK-Learn interface. . . . 210
C.14 **Adding a wrapper for Group fairness metric.** . . . 210
C.15 **Metrics recorded for binary classification.** Adding the metrics to this dictionary results in automatic logging. See the whole file for more details. 210
D.1 **Implementation of the Nadeau and Bengio correction of dependent Student's t-test [NB03]**. Inspired by an open-source implementation [De 21]. . . . 240

E.1 **AUMCdb MEDS configuration predicates defined in the syntax of ACES for 8 endpoints.** Defined for the extended fork of *MEDS-DEV* at `https://github.com/rvandewater/MEDS-DEV` 268
E.2 **Hyperkalemia prediction task configuration in the ACES syntax applicable to 8 different datasets.** Defined for *MEDS-DEV* at `https://github.com/rvandewater/MEDS-DEV` 271

1

Introduction

Health has always been a central aspect of human life, and the development of medicine has significantly improved our ability to treat diseases and improve the quality of life. More humans lead substantially longer and healthier lives than in the past, thanks to advances in science and technology. However, as we may never fully understand complex and emergent biological systems, we increasingly face the challenge of understanding the complex interactions between human functions, diseases, and treatments. Moreover, the increasing complexity of modern healthcare systems, with their vast amounts of data and diverse patient populations, necessitates new approaches to analyze and interpret this information effectively. We are also facing a global shortage of healthcare professionals [Agy+23], which is worsened by the increasing demand for healthcare services due to an aging population and a rise in chronic diseases [Nep+25].

Fortunately, improvements in data availability, computing power, and model architectures have led to significant advances in machine learning (ML) and artificial intelligence (AI) in a broader sense, which have the potential to significantly impact daily practice in several ways [Zha+22]. These technologies can help us effectively analyze large datasets, identify patterns, and make predictions that can improve patient outcomes and reduce costs; for example, invasive and costly diagnostic tools would only be used when the prediction systems deem it necessary. In a world where emergent capabilities of Large Language Models (LLMs) have challenged the limits of what we thought was possible with ML [Bro+20], this seems par for the course. Suddenly, AI has entered the public consciousness and promises to change many aspects of society [Bom+21]. However, the use of machine learning for healthcare (ML4H) is still in its early stages, and much research remains to be conducted to integrate it effectively, safely, and efficiently into daily practice.

1.1 Motivation

The motivation for this thesis is to address several key challenges in the methodology of developing and deploying machine learning for healthcare (ML4H) systems for clinical settings. We achieve this by addressing three major challenges in the areas of reproducibility, clinical ML modeling, and data tooling. These mirror the contributions presented in each chapter.

Comparing modeling, preprocessing, and evaluation approaches is often challenging; many works claim to have an effective system for a particular task, but the results are often not reproducible due to various reasons. In cases that they are reproducible, there is no way of knowing if they provide the best possible performance. Open datasets are increasingly available but used in manifold ways, and code, when shared, is often not well-documented, outdated, or performs poorly.

Moreover, these systems rely on large amounts of sparse, heterogeneous, and noisy data, making it challenging to develop accurate and reliable models. Deploying parties must ensure that models are robust to shifts in data distributions and hospital workflows and, importantly, feasible for a particular task. An important factor in the feasibility is having the correct data available at the right time. The only way to ensure this is often by performing a prospective study with clearly defined data collection protocols, clinical outcome definitions, and modelling approaches.

Finally, to effectively scale up these systems, we need tooling that allows us to stop reinventing the wheel for every new dataset or task. To get a running start, we need tools that address the community's needs and allow us to build on each other's work. Examples are standardized data formats, data exploration tools, preprocessing pipelines, and evaluation frameworks that can be easily adapted to different datasets and tasks.

1.2 Thesis Outline

This thesis is structured as follows:

Chapter 1 introduces the motivation, the research contributions of my work, and outline of this thesis.

Chapter 2 provides preliminaries, background, and related work on the topics covered in this thesis. It serves as a primer for research on applying machine learning in healthcare.

Chapter 3 introduces the *Yet Another ICU Benchmark* (*YAIB*) framework, which we designed to support reproducible clinical ML experiments with open-access ICU datasets. It also demonstrates the effectiveness of several model architectures for clinically relevant tasks, providing researchers with a foundation for future work.

Chapter 4 presents the work I did in the framework of the retrospective and prospective Clinical ASSist AND aleRt Algorithms (CASSANDRA) study[1], in close collaboration with clinicians and experimental surgical researchers at a major visceral surgery center. We have employed the retrospective datasets to predict severe postoperative complications. We have developed a multimodal early warning system for postoperative complications within the prospective study. Moreover, this chapter discusses the challenges of developing and deploying an Early Warning System (EWS) in contexts where data is scarce or of insufficient quality.

Chapter 5 introduces the collaborative effort of Medical Event Data Standard (MEDS), a flexible, low-level ML-first format that integrates with EHR processing and modeling

[1]Grant: https://innovationsfonds.g-ba.de/projekte/versorgungsforschung/cassandra.395

tools. It accelerates the training of predictive models on clinical tasks as an event-based standard with an expanding array of tools. We describe the development of several tools and employ them on real-world data to demonstrate their effectiveness.

Chapter 6 concludes this thesis by taking a broader look at where the field is headed, discussing the implications of this work, and outlining future research directions.

1.3 Research Contribution

Scientific advancement hinges on collaborative dialogue, shared expertise, and diverse perspectives. This thesis emerged from multiple collaborative studies where I was the primary driver. Throughout this work, I use "we" to reflect the collective nature of these research endeavors. In the following sections, I recognize the valuable contributions of colleagues and co-authors while delineating my role in each work.

Chapter 2

This chapter outlines the preliminaries of this work and discusses the clinical machine learning field with a focus on Electronic Health Record (EHR) data. It is mainly based on my own literature review at the beginning of my PhD and has been amended with additional work that I found during the course of this research. This remains unpublished at the time of writing.

Chapter 3

Van de Water et al. [Van+24a]: This chapter is primarily based on this work, which was presented at the 2024 International Conference on Learning Representations, where we introduced a novel benchmarking framework for reproducing clinical machine learning experiments on intensive care unit (ICU) and compatible with EHR data more broadly. Additionally, we demonstrated that the choice of dataset, cohort definition, and preprocessing have a major impact on prediction performance, often more so than the model class.

Contributions: R.P.v.d.W. performed the conceptualization, method development, main writing, and experiments. P.R. conceptualized, assisted with writing, and ensured the soundness of the experiments. The system implementation was carried out by R.P.v.d.W., H.S., and P.R. P.T. and P.E. verified the clinical soundness of the cohorts. B.A. provided funding and proofread this work.

Van de Water et al. [VSR26]: ReciPies addresses the common problem of scattered and opaque data preprocessing steps in Python machine learning workflows by providing a human-readable, Configuration as Code approach that allows users to declare transformations as structured recipes with ordered steps applied to semantically-defined variables. This declarative framework enhances reproducibility, reduces the cognitive load of feature engineering, and makes methodological choices explicit, while supporting multiple backends and enabling recipes to be prepared once and applied repeatedly to both training and new data.

Contributions: R.P.v.d.W., H.S., and P.R., where we all contributed to the design and implementation of the framework. R.P.v.d.W. improved the codebase, conceptualized this work, performed the experiments, and wrote the paper. P.R. supervised and assisted with writing and framing.

Turubayev et al. [Tur+25]: This work is the result of the continuation of a master project I supervised on evaluating imputation methods for ICU data. We extend an extensible and reusable benchmark with currently 15 imputation and 4 amputation (data removal) methods, created for benchmarking on major ICU datasets. It is currently available as preprint. The chapter only tangentially covers this work as a discussion on the extension of the *YAIB* framework.

Contributions: R.P.v.d.W. conceived the presented idea and supervised the project. A.T., A.S., F.L., M.K., P.M., and V.A. planned, executed, and reported on experiments. R.P.v.d.W. carried out the re-run of the experiments for this paper and prepared the original draft; A.T., A.S., F.L., M.K., P.M., and V.A. reviewed and revised the draft.

The following abbreviations were used: R.P.v.d.W. (Robin P. van de Water), H.S. (Hendrik Schmidt), P.T. (Patrick Thoral), P.E. (Paul Elbers), B.A. (Bert Arnrich), and P.R. (Patrick Rockenschaub), A.T. (Alisher Turubayev), A.S. (Anna Shopova), F.L. (Fabian Lange), M.K. (Mahmut Kamalak), P.M. (Paul Mattes), and V.A. (Victoria Ayvasky).

Chapter 4

Van de Water et al. [Van+25c] The main contribution from the prospective clinical study involves the development of a comprehensive Machine-Learning-Driven Early Warning System for detecting postoperative complications, as detailed in Van de Water et al. [Van+25c] (preparing for submission to Nature Medicine). This study addresses the critical gap in postoperative monitoring by developing a comprehensive Early Warning System that combines preoperative, intraoperative, ICU, and general ward parameters with high-resolution continuous vital sign measurements from wearable devices deployed on nursing wards, where visceral surgery complications frequently occur despite the absence of real-time monitoring. The integrated approach demonstrated improved early warning prediction for multiple complications, including surgical site infections, postoperative pancreatic fistulas, hematomas, and bile leaks, highlighting the need for hybrid monitoring systems that extend beyond ICU stays to enhance the clinical usability of machine learning in healthcare settings.

A.W. and M.M.M. conceptualized and conducted the prospective trial. R.P.v.d.W, A.W., and M.M.M. conceptualized this work. A.W. supervised the study and data collection process; R.P.v.d.W designed and supervised the data integration, modality fusion, preprocessing, and aggregation; R.P.v.d.W created the experimentation framework and modality data integration. R.P.v.d.W. conceptualized the evaluation approach and created Figures 1-5 and all figures in the methods section and appendix. B.P. analyzed the wearable data and created the preprocessing pipelines. D.Z.L. and R.P.v.d.W. created the alarm system, as visualized in Figure 5, and corresponding statistics, with supervision. L.F. and M.M.M. provided feedback on the figures. P.R. provided supervision for the ML approaches. F.K., W.S., T.M., and C.D. contributed to the verification of patient labels. R.P.v.d.W, A.W., D.Z.L., P.R., and M.M.M. wrote the manuscript with the assistance and feedback of all the other co-authors.

Van de Water et al. [Van+24c] and Van de Water et al. [Van+24b]: Two workshop papers accepted at the 2024 International Conference on Learning Representations, Van de Water et al. [Van+24c] and Van de Water et al. [Van+24b], were based on intermediate data collected during the prospective study. They assessed the critical challenge of incomplete healthcare datasets and data shortage after ICU discharge by developing a comprehensive collection framework that combines preoperative, intraoperative, ICU, and general patient parameters with high-resolution continuous vital sign measurements from wearable devices deployed on nursing wards. The integration of this wearable data demonstrated improved prediction accuracy for deep organ space surgical site infections (SSI-III), suggesting that hybrid monitoring systems extending beyond ICU stays are essential for effective machine learning modeling in clinical settings.

Contributions R.P.v.d.W's involvement encompassed conceptualization, development, experimental design, and writing, while A.W. and D.Z. handled clinical data collection. A.W. and M.M.M. contributed to the clinical conceptualization, F.T. implemented portions of the data processing pipeline, and B.P. provided experimental feedback. D.Z. assisted with related work research, and the projects received financial supervision from I.S. and B.A.

Riepe et al. [Rie+25b]: The machine learning benchmarking study described in Riepe et al. [Rie+25b], published in the International Journal of Surgery, is a retrospective study of 7,711 patients undergoing major elective visceral surgery across two tertiary centers, compared machine learning models against conventional scoring systems, American Society of Anesthesiologists (ASA) score and revised Charlson Comorbidity Index (CCI), for preoperative 90-day mortality prediction, where one model significantly outperformed both conventional scores. Moreover, a multi-organ training approach proved superior to organ-specific models, suggesting substantial potential for improving surgical outcome prediction beyond traditional scoring methods:

Contributions Conceptualization involved C.R., A.W., R.P.v.d.W., and M.M., while data curation was handled by C.R., A.W., R.A., and M.M. Formal analysis was conducted by C.R. and R.P.v.d.W., with methodology development involving C.R., R.P.v.d.W., A.W., M.S., and M.M. Project administration was managed by A.W., B.A., I.S., and M.M., while software development involved C.R., R.P.v.d.W., and B.P. Supervision was provided by C.S., K.B., J.P., B.A., I.S., and M.M., with validation performed by R.P.v.d.W. Visualization was handled by C.R. and R.P.v.d.W., with C.R. writing the original draft and receiving review and editing input from R.P.v.d.W., B.P., M.S., L.F., D.Z., I.S., and M.M.M.

The following abbreviations were used: R.P.v.d.W. (Robin P. van de Water), A.W. (Axel Winter), D.Z. (Daniela Zuluaga), B.P. (Bjarne Pfitzner), L.F. (Lara Faraj), P.R. (Patrick Rockenschaub), B.A. (Bert Arnrich), I.S. (Igor M. Sauer), M.M.M. (Max M. Maurer), F.T. (Felix Treykorn), M.I. (Marius Ibach), C.R. (Christoph Riepe), R.A. (Robert Ahlborn), F.K. (Felix Krenzien), E.M.D. (Eva M. Dobrindt), J.R. (Jonas Raakow), K.B. (Katharina Beyer), C.D. (Christian Denecke), J.P. (Johann Pratschke), M.S. (Maximilian Schulze), and C.S. (Christian Schineis).

Chapter 5

This chapter is based on the collaborative effort of Medical Event Data Standard (MEDS) and involves several publications in which I was involved in conceptualization, writing, and experiments. All authors who contributed are listed at the end of this subsection.

Group [Gro24]: This work was accepted at the Workshop Learning From Time-series For Health at the 2024 International Conference on Learning Representations. The Medical Event Data Standard (MEDS) is a lightweight schema that enables machine learning over EHR data by providing maximum interoperability across datasets, existing tools, and model architectures. By creating a simple standardization layer between datasets and model-specific code, MEDS facilitates more reproducible, robust, computationally performant, and collaborative machine learning research in healthcare.

Contributions: All primary authors contributed to conceptualizing, designing, and implementing the MEDS framework. I contributed to the design, paper writing, and related work research. The MEDS working group was made up of (alphabetically ordered): Edward Choi, Jason A. Fries, Matthew B. A. McDermott, Jungwoo Oh, Tom J. Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, Robin P. van de Water. Bert Arnrich provided financial support.

McDermott et al. [McD+24] This work was accepted as a demo track submission at the Machine Learning for Health (ML4H) symposium colocated with Conference on Neural Information Processing Systems (NeurIPS) 2024. MEDS-DEV (Decentralized, Extensible Validation) represents a first step towards a standardized benchmarking infrastructure for healthcare research, addressing key limitations in data standardization, task definition consistency, and multi-institutional participation by enabling the evaluation of four state-of-the-art EHR foundation models across multiple institutions for the first time. As health systems increasingly deploy models clinically, MEDS-DEV will play a crucial role in model validation and accelerating ML development for EHR data, with the community already growing to several dozen members and expanding to the broader ML4H audience.

Contributions: This effort is a collaboration of many groups and individuals. I was involved in the core MEDS-DEV working group and the AUMCdb dataset and conceptualized and co-wrote the paper. The other authors are Matthew McDermott, who mainly conceptualized and drove this work, and, in alphabetical order: Teya S. Bergamaschi, Jeffrey N. Chiang, Edward Choi, Young Sang Choi, Jason A. Fries, Jack Gallifant, Raffaele Giancotti, Xinzhuo Jiang, Hyewon Jeong, Vincent Jeanselme, Shalmali Joshi, Alistair Johnson, Apara Kashyap, Kiril V. Klein, Aleksia Kolo, Yuta Kobayashi, Ryan C. King, Simon A. Lee, Yanwei Li, Matthew B. A. McDermott, Maria E. Montgomery, Mikkel Odgaard, Jungwoo Oh, Nassim Oufattole, Chao Pang, Tom J. Pollard, Pawel Renc, Patrick Rockenschaub, Nigam H. Shah, Martin Sillesen, Ethan Steinberg, Kamilė Stankevičiūtė, Michael Wornow, Justin Xu, Mads Nielsen.

McDermott et al. [McD+25d]: This work is a tutorial that was accepted at the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2025). Health AI suffers from a systemic reproducibility crisis that hinders research across academia and industry, prompting the proposal of MEDS and its open-source ecosystem to empower researchers by simplifying the construction of AI models over longitu-

dinal EHR datasets and enabling straightforward benchmarking. MEDS has achieved widespread adoption with usage at over 15 institutions across 8 countries, featuring 7+ open-source tools, supporting 10+ published models, and providing Extract, Transform, Load (ETL) pipelines for major public datasets. The tutorial aims to equip researchers with the resources to get started with MEDS.

Contributions: I was involved in writing and designing the paper, as well as creating the tutorial texts and datasets. The primary author is Matthew B. A. McDermott; the other authors are alphabetically ordered: Justin Xu, Teya S. Bergamaschi, Hyewon Jeong, Simon A. Lee, Nassim Oufattole, Patrick Rockenschaub, Kamilė Stankevičiūtė, Ethan Steinberg, Jimeng Sun, Robin P. van de Water, Michael Wornow, John Wu, Zhenbang Wu.

Van de Water [Van25c]: This work was presented at the Mount SinAI retreat (2025) and the SBHD 2025 International Conference on Systems Biology of Human Diseases as an abstract. Here, I demonstrate the progress of deploying MEDS infrastructure on a database of 6.8 million patients. Several building blocks are necessary to successfully implement clinical AI in hospital systems, such as the Mount Sinai hospital system, including integrated health information systems, robust computational infrastructure, accessible software layers for data scientists and clinicians, active clinician involvement in deployment, and effective model evaluation and maintenance processes.

Contributions: I was involved in conceptualizing, designing, and writing the software and abstract.

Van de Water et al. [Van+25a]: This work is in early stages and consists of benchmarking and applying EHR foundation models on the Mount Sinai Health System (MSHS) data. I am the primary driver of this work with senior authors Christoph Lippert and Eugenia Alleva. A large number of other authors involved in the MSHS will make contributions to this work in terms of cohort creation and task definitions; these are currently listed as co-authors but will be finalized as the work progresses.

Contributions: I have been involved in the conceptualization, method development, foundation model training experiments, evaluation, and writing of this work. Christoph Lippert and Eugenia Alleva have provided advice on conceptualization, supervision, and feedback. Detailed contributions are still being defined as this work is ongoing.

Chapter 6

Chapter 6 is a discussion chapter that reflects on the implications of this work and outlines future research directions. It is based on my own reflections and shortly discusses societal implications and future work. This remains unpublished at the time of writing.

1.3.1 Software and Tooling

I have driven the development of several Free and Open Source Software (FOSS) projects that are related to the topics covered in this thesis and are targeted primarily at researchers. These projects have been used in the works presented in this thesis and are publicly available for use by the broader research community:

- **GUIDataFarm:** GUIDataFarm is a human-in-the-loop framework for efficiently generating and labeling large query workloads. It follows a data-driven white-box approach to learn from pre-existing small workload patterns, input data, and computational resources. Repository: `https://github.com/rvandewater/GUIDataFarm`. Papers: [Van+22b; Van+22a]
- **YAIB:** *Yet Another ICU Benchmark* (*YAIB*) is a modular framework designed to support reproducible clinical ML experiments with open-access ICU datasets. It provides an end-to-end solution for model evaluation, including predefined tasks like mortality and sepsis. It highlights the critical role of dataset selection, cohort definition, and preprocessing in model performance. Repository: `https://github.com/rvandewater/YAIB`. Paper: [Van+24a]
- **ReciPies:** The ReciPies package is a preprocessing framework operating on Polars and Pandas dataframes. It provides a set of tools for preprocessing clinical data, including time series data, and we designed it to be a standalone extension of the YAIB framework which developed into its' own work at a later point in time. Repository: `https://github.com/rvandewater/ReciPies`. Paper: [VSR26]
- **CASS-PROPEL:** developed together with Bjarne Pfitzner, CASS-PROPEL is a framework for developing static prediction models. Repository: `https://github.com/rvandewater/CASS-PROPEL`. Papers: [Van+24c; Van+24b; Win+24; Rie+25a]
- **harmonizEHR:** harmonizEHR is a tool for translating free text to codes and between coding vocabularies. It is designed to work with the Medical Event Data Standard (MEDS) and supports the translation of clinical data into standardized formats. Repository: `https://github.com/rvandewater/harmonizehr`.
- **Alarmist:** Alarmist is a tool for evaluating and visualizing early warning systems (EWS) for clinical prediction tasks. It provides a framework for building and evaluating traditional EWS models. Repository: `https://git.bihealth.org/vandewrp/alarm`
- **CASHEWS:** An extensive suite of segmentation algorithms for the diverse CASSANDRA modalities, including wearable photoplethysmography (PPG) data, clinical notes, lab values, and ICU data. Repository: `https://github.com/rvandewater/CASHEWS`
- **OMOP_schema:** a Python package designed to read, manage, and convert Observational Medical Outcomes Partnership (OMOP) data into the correct schema. It provides tools to handle OMOP Common Data Model (CDM) tables, convert schemas between different formats (e.g., PyArrow, Polars, Pandas), and load datasets efficiently. Repository: `https://github.com/rvandewater/OMOP_schema`

The Medical Event Data Standard (MEDS)

The Medical Event Data Standard (MEDS) is a flexible, low-level ML-first data schema that integrates with EHR processing and modeling tools. The larger working group intends to make reproducible clinical ML research easier. I contributed to the design and implementation of the MEDS schema and tooling ecosystem:

Table 1.1: Overview of datasets with Medical Event Data Standard (MEDS) Extract, Transform, Load (ETL) support

Dataset	Stays	Version	Frequency	Origin	Year	License	MEDS ETL	My Role	Full Dataset Name
AUMCdb [Tho+21]	23,000	v1.0.2	up to 1 min	Netherlands	2019	Not specified	GitHub [VRR25]	Main Dev.	Amsterdam University Medical Center Database
eICU [Pol+18]	201,000	v2.0	5 minutes	USA	2017	PhysioNet	GitHub [VZM25]	Contributor	eICU Collaborative Research Database
EHRSHOT [Wor+23a]	6,739	v3.0	heterogeneous	USA	2023	Not specified	Pre-converted and OMOP [van25d]	-	EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models; a subset of the Stanford Medicine health system dataset.
HiRID [Hyl20]	34,000	v1.1.1	2/5 minutes	Switzerland	2020	PhysioNet	GitHub [Van25d]	Main Dev.	High-Resolution ICU Dataset
INSPIRE [Lim+24]	130,000	v1.2	Not specified	South Korea	2024	Korea Health Data	GitHub [van25b]	Main Dev.	INformative Surgical Patient dataset for Innovative Research Environment
MIMIC-IV [Joh+22]	73,000	v3.1	~1 hour	USA	2020	PhysioNet	GitHub	Contributor	Medical Information Mart for Intensive Care IV
NWICU [Mou+24]	25,000	v0.1.0	Not specified	USA	2023	PhysioNet	GitHub [van25c]	Main Dev.	Northwestern ICU Database
SICdb [Rod+23a]	27,350	v1.0.8	1 minute	Austria	2024	PhysioNet	GitHub [van25e]	Main Dev.	Salzburg Intensive Care Database
MSHS*	6.8M	OMOP 5.3	heterogeneous	USA	2025	Not specified	GitHub [van25d]	Main Dev.	Mount Sinai Health System
CUMC *	6.7M	OMOP 5.4	heterogeneous	USA	2025	Not specified	OMOP ETLs	-	Columbia University Medical Center
SM*	2.57M	OMOP 5.4	heterogeneous	USA	2023	Not specified	OMOP ETLs	-	Stanford Medicine
OMOP† [Vos+15]	-	5.x	heterogeneous	various	various	Various	GitHub [van25d]	Main Dev.	Observational Medical Outcomes Partnership (OMOP) datasets
FHIR† [BS13]	-	v4/v5	heterogeneous	-	-	Not specified	GitHub [van25a]	Main Dev.	Fast Healthcare Interoperability Resources (FHIR) datasets

* These datasets are not publicly available but are internally used with MEDS; my MEDS ETL pipeline [van25d] should facilitate their conversion as well. The Mount Sinai dataset updates continuously from ETL pipelines that feed into it from the Mount Sinai Health System. Metadata timestamps are provided. The ETL pipeline is designed to work with any Observational Medical Outcomes Partnership (OMOP)[Vos+15] 5.x dataset.

† These are generic ETL pipelines that can convert any dataset in the respective format to MEDS format. The OMOP pipeline supports any OMOP 5.x dataset, while the FHIR pipeline supports datasets in FHIR v4 and v5 formats.

- MEDS core: The core framework for the Medical Event Data Standard, providing a unified approach to representing and sharing medical event data. Repository: `https://github.com/Medical-Event-Data-Standard/meds`
- MEDS-DEV (MEDS Decentralized, Extensible Validation): A decentralized and extensible validation framework for MEDS, enabling flexible and scalable validation of medical event data. Paper: [McD+24]. Repository: `https://github.com/Medical-Event-Data-Standard/MEDS-DEV`
- MEDS-Inspect (main developer): An interactive tool for exploring and visualizing MEDS-compliant data, facilitating data understanding and model development. Repository: `https://github.com/rvandewater/MEDS-Inspect`
- MIMIC-IV demo data in the Medical Event Data Standard (MEDS) [Van+25d]: A demo dataset in MEDS format based on MIMIC-IV, showcasing the capabilities of the MEDS framework for clinical machine learning research. Repository: `https://physionet.org/content/mimic-iv-demo-meds/`

For details on the ETLs I (co-) developed see Table 1.1.

1.4 Additional Contributions

In the publications below, I am included as a co-author. I was involved in implementation, ideation, and/or writing. This work is out of scope for this thesis as it does not directly relate to the challenges addressed in this work; however it might be referenced and cited in certain parts.

Published

- Matthias Kirchler, Matteo Ferro, Veronica Lorenzini, Robin van de Water, Christoph Lippert, and Andrea Ganna. "Large Language Models Improve Transferability of Electronic Health Record-Based Predictions across Countries and Coding Systems". In: *Accepted to npj Digital Medicine* (2025)
- Bjarne Pfitzner, Max M. Maurer, Axel Winter, Christoph Riepe, Igor M. Sauer, Robin van de Water, Christian Denecke, Johann Pratschke, and Bert Arnrich. "Differentially-Private Federated Learning with Non-IID Data For Surgical Risk Prediction." In: *International Journal of Semantic Computing* 19.3 (2025). (Visited on 2025-11-23)
- Axel Winter, Robin P. van de Water, Bjarne Pfitzner, Marius Ibach, Christoph Riepe, Robert Ahlborn, Lara Faraj, Felix Krenzien, Eva M. Dobrindt, and Jonas Raakow. "Enhancing Preoperative Outcome Prediction: A Comparative Retrospective Case–Control Study on Machine Learning versus the International Esodata Study Group Risk Model for Predicting 90-Day Mortality in Oncologic Esophagectomy". In: *Cancers* 16.17 (2024), p. 3000. (Visited on 2024-10-23)
- Bjarne Pfitzner, Max M. Maurer, Axel Winter, Christoph Riepe, Igor M. Sauer, Robin van de Water, and Bert Arnrich. "Differentially-Private Federated Learning with Non-IID Data for Surgical Risk Prediction". In: *2024 IEEE First International Conference on Artificial Intelligence for Medicine, Health and Care (AIMHC)*. IEEE, 2024, pp. 120–129. (Visited on 2024-10-23)
- Orhan Konak, Robin van de Water, Valentin Döring, Tobias Fiedler, Lucas Liebe, Leander Masopust, Kirill Postnov, Franz Sauerwald, Felix Treykorn, and Alexander Wischmann. "HARE: Unifying the Human Activity Recognition Engineering Workflow". In: *Sensors* 23.23 (2023), p. 9571. (Visited on 2024-10-23)
- Orhan Konak, Alexander Wischmann, Robin van de Water, and Bert Arnrich. "A Real-time Human Pose Estimation Approach for Optimal Sensor Placement in Sensor-based Human Activity Recognition". In: *Proceedings of the 8th International Workshop on Sensor-Based Activity Recognition and Artificial Intelligence*. Lübeck Germany: ACM, 2023-09, pp. 1–6. ISBN: 979-8-4007-0816-9. DOI: 10.1145/3615834.3615848. (Visited on 2024-10-23)
- Lientje Maas, Mathan Geurtsen, Florian Nouwt, Stefan F. Schouten, Robin van de Water, Sandra Van Dulmen, Fabiano Dalpiaz, Kees Van Deemter, and Sjaak Brinkkemper. "The Care2Report System: Automated Medical Reporting as an Integrated Solution to Reduce Administrative Burden in Healthcare." In: *HICSS*. 2020, pp. 1–10. (Visited on 2024-10-23)

Under submission

- Jan Carlo Schmid, Susanne Ibing, Stefan Kalabakov, Katharina Alefs, Robin P. van de Water, Arkadiusz Kwasigroch, Eugenia Alleva, Bert Arnrich, Bernhard Y. Renard, and Maia Kayal. "Expert-Defined, Chart-Reviewed Features Outperform EHR Foundation Models for Predicting Crohn's-like Disease of the Pouch". In: *Under review at The American Journal of Gastroenterology* (2025)

Preprints

- Alisher Turubayev, Anna Shopova, Fabian Lange, Mahmut Kamalak, Paul Mattes, Victoria Ayvasky, Bert Arnrich, Bjarne Pfitzner, and Robin P. van de Water. *Closing Gaps: An Imputation Analysis of ICU Vital Signs*. 2025-10. DOI: 10.48550/arXiv.2510.24217. arXiv: 2510.24217 [cs]. (Visited on 2025-10-29)

Manuscripts in preparation

- Wouter van Amsterdam, Michael Kamp, Rajesh Ranganath, Robin P. van de Water, Florian Markowetz, Evangelia Christodoulou, Yamuna Krishnamurthy, Christoph Lippert, Jeff Clark, Julia E. Vogt, Raul Santos-Rodriguez, Thomas Gärtner, Gilbert Koch, and Brett Beaulieu-Jones. "Missing Incentives, Missing Oversight: The Challenges of AI Monitoring in Clinical Practice". 2025

Seminars

- Raul Santos-Rodriguez, Kacper Sokol, Julia E. Vogt, and Sven Wellmann. "From Research to Certification with Data-Driven Medical Decision Support Systems (Dagstuhl Seminar 25052)". In: *Dagstuhl Reports* 15.1 (2025). Ed. by Raul Santos-Rodriguez, Kacper Sokol, Julia E. Vogt, and Sven Wellmann, pp. 201–220. ISSN: 2192-5283. DOI: 10.4230/DagRep.15.1.201. (Visited on 2025-11-26)

1.4.1 Science of Data Systems Contributions

The master's thesis at the Technical University of Berlin (TU Berlin) resulted in the following publications in the field of Human-in-the-Loop (HitL)-driven data generation, which we contributed to the broader scientific community.

The work Van de Water et al. [Van+22a] was published as an abstract at the Conference on Innovative Data Systems Research (CIDR) 2022. Van de Water et al. [Van+22b] was published at the IEEE 38th International Conference on Data Engineering (ICDE) 2022 and received the *Best demonstration paper award*. The motivation is that the collection of labeled training data for ML-based query optimizers is tedious and time-consuming, requiring extensive job development and execution to obtain ground-truth labels. DataFarm addresses this by providing a framework that efficiently generates tailored training data through active learning, user feedback via a Graphical User Interface (GUI), and learning from existing workloads to produce benchmark jobs and labeled datasets.

1.5 Supervision and Teaching

Below is a list of the supervision activities in which I was involved, including conceptualization, design, partial implementation, and writing. I was also the main supervisor for all master's theses and the master's project.

- Master Thesis: Daniela Zuluaga Lotero, *Evaluation of Early Warning Systems for postoperative complications from continuous vital sign monitoring in surgical wards* (2025)
- Master Thesis: Youssef Tharwat Zaky Aly Mecky, *Comparative analysis of explainable AI methods for ICU risk prediction models* (2024)
- Master Thesis: Hendrik Schmidt, Master Thesis, *Benchmarking Model-agnostic Multi-source Supervised Domain Adaptation for Clinical Prediction on ICU Data* (2023)
- Master Project: Alisher Turubayev, Anna Shopova, Fabian Lange, Mahmut Kamalak, Paul Mattes, Victoria Ayvasky, *Deep Learning Data Generation for Medical Prediction Systems* (2023)
- Course grading and project supervision for *Statistics in Connected Healthcare* (2022, 2023)

2

Background

The use of artificial intelligence (AI), data science, or machine learning (ML) with health data has been marked by a gradual shift from expert systems and traditional statistical methods and models to data-driven approaches [Kul19], culminating in unsupervised and self-supervised models trained on large datasets [Bom+21; Ste24; Wor+23b]. Important work in health informatics preceded this shift, such as: the digitalization of patient records [WF89], the development of standardized coding systems and ontologies [RNC94; LHM93], data formats [Mur+09], and the creation of large-scale health repositories and databases [Gol+00; MM96]. These advancements have enabled the collection and analysis of vast amounts of diverse health data, which is essential for training machine learning models. Moreover, different approaches to applying computer science and/or AI to health care have come (long) before, such as logical inference systems [Pop73]. Currently, multiple computer-aided systems are implemented in clinical workflows [Lia+25], though a minority use the promising ML approaches developed in the past 20 years. Adoption is expected to increase given current investments and progress, though there are still major hurdles that will be discussed in this work.

Machine learning for healthcare (ML4H) is a widening field of research where existing methods are applied and, often, improved to work better with the characteristics of the data, practice, and clinical goals. Many challenges remain [Gha+20; Adi+25; Che+24], even in a post-ChatGPT world [Bro+20]. Different developments in the broader field have been applied and refined within healthcare contexts. For example, advancements in natural language processing allow for a more effective analysis of clinical notes and other unstructured text data [Li+20].

This chapter covers the preliminaries of health data and explores ML4H. We aim to emphasize the challenges addressed in this work, including reproducibility, modeling a range of data modalities, and building robust health modeling infrastructure.

2.1 A Terminology of Health Data

Unsurprisingly, health data, in all its forms, is the foundation of ML4H. Health data can be categorized into several (partially overlapping) types [Cas25], each with its own

characteristics and applications. The following section gives an overview of the data types commonly used in ML4H.Cascini [Cas25] distinguishes data in terms of their source. Their terminology will be explained and used in this work. A short reference is given in Table 2.1.

Real-world Data (RWD) encompasses patient-related health information collected from routine healthcare sources like electronic health records and insurance claims, which is then analyzed to generate real-world evidence about medical products' safety and effectiveness. *Administrative health data* consists of billing records generated when patients interact with healthcare systems, primarily collected for payment purposes. *Biobank* data comprises extensive collections of biological specimens linked to personal and health information, organized explicitly for medical research purposes. *Clinical trial* data refers to structured information collected during controlled research studies to answer specific research questions and evaluate medical interventions; note that this can be seen as the opposite of real-world data since it is collected in a controlled environment. *Electronic Health Records EHRs* are digital patient charts containing comprehensive health information, including medical histories, medications, treatment plans, and test results. It can be part of either retrospective real-world data or collected as part of a clinical trial. *Genomic and proteomic data (omics)* provide molecular-level insights from systematic studies of genes, proteins, and their interactions to understand biological functions and disease mechanisms. *National health registries* collect and analyze data on specific diseases across populations for surveillance and research purposes and provide this data to the research community. *Person-generated health data* includes health-related information created by patients using wearable devices, mobile apps, and home monitoring systems outside traditional clinical settings. Finally, *social and behavioral determinants of health* encompass environmental and social factors like education, addresses, and income that influence health outcomes and quality of life.

We further categorize the data into *structured* (often tabular), *unstructured* (often text and images), and *semi-structured* data (often in formats such as JavaScript Object Notation (JSON), Extensible Markup Language (XML), or standardized clinical reports). Then, we also make the distinction between *temporal (dynamic)* and *non-temporal (static)* data; the distinction is quite nuanced, as some perceived non-temporal data might change slowly over time; e.g., the age group of a person. Below, we outline some of the types of health data that are particularly relevant to this work.

2.1.1 Electronic Health Records (EHR)

An Electronic Health Record (EHR) is a digital version of traditional patients' paper charts. They could collect comprehensive patient data, such as demographics, medical history, medications, allergies, immunization status, laboratory test results, radiology images, vital signs, and billing information [BMV16]. The data in EHR systems is typically semi-structured and may be stored across multiple systems developed by different manufacturers. These systems are often primarily designed for billing and other administrative purposes. However, they can also serve a secondary purpose, such as providing clinical decision support, facilitating research, and supporting quality improvement initiatives. Figure 2.1 illustrates a sample timeline of a patient's electronic

Table 2.1: Healthcare data categories, their contents, the applications of machine learning (ML) within each category, and example open-access datasets.

Category	Typical contents & structure	Core ML applications	Datasets
Electronic Health Records (EHRs)	Structured tables (demographics, ICD codes, labs, meds, procedures); unstructured notes; irregular longitudinal sequences	Phenotyping, outcome prediction, decision support, risk scoring, representation learning	MIMIC-IV [Joh+23c], eICU-CRD [Pol+18]
Medical Imaging	Radiology (CT, MRI, X-ray, US), pathology whole-slide images, ophthalmic OCT/fundus, dermatology photos; pixel-dense arrays	Detection/segmentation, diagnosis, prognosis, reconstruction, triage	ChestX-ray8[Wan+17], BraTS [Men+15], UK Biobank MRI
Omics	Genomics (WGS/WES), transcriptomics (RNA-seq), epigenomics, proteomics, metabolomics	Variant calling, gene–disease association, drug-target interaction, prognosis	1000 Genomes [Aut+15], TCGA [Aut+15], GTEx [Lon+13], UK Biobank exome [Bac+21]
Physiologic Signals	Continuous ECG/EEG/PPG/ABP, ICU bedside vitals, polysomnography; high-frequency waveforms	Arrhythmia detection, seizure forecasting, sleep staging, sepsis prediction	MIT-BIH [MM01], MIMIC-WFDB [Moo+22], Sleep-EDF [Kem+00]
Wearable Sensors	Accelerometry, gyroscope, temperature, SpO_2, GPS; multivariate time-series	Activity recognition, fall detection, digital biomarkers, rehabilitation adherence	UCI-HAR [Jor13], MobiAct [Vav+16], UK Biobank accelerometry [Doh+17]
Administrative & Claims Data	Billing/procedure codes, prescription claims, cost totals, provider identifiers; long tabular format	Fraud/waste detection, cost prediction, risk adjustment, cohort identification	US Veterans Affairs data [Jus+06] Merative MarketScan [18]

health record events. Data quality and completeness are a major challenge with *retrospective* EHR data. This type of data is also referred to as observational [Jea24].

2.1.2 Prospective Studies

Prospective studies collect data over a specified time period, usually with a specific research question in mind; this allows researchers to observe changes in health status, behaviors, and other factors. In healthcare, prospective studies are frequently employed in clinical trials and observational studies to evaluate the effectiveness of treatments or interventions. Importantly, this data can be higher quality than retrospectively retrieved EHR data. For example, defining accurate cohorts and labels for specific subpopulations can be a considerable effort. A Randomized Controlled Trial (RCT) is a prospective study that randomly assigns participants to different treatment groups to compare outcomes. This design helps minimize bias and confounding factors and is often considered the gold standard in evaluating ML based interventions [Pla+22]. This type of data is also known as experimental data [Jea24] (c.f. observational).

2.1.3 Open-access Medical Datasets

Medical datasets are crucial for the development and validation of ML-powered systems. As EHRs have become more common, publicly available datasets have emerged. Because intensive care unit (ICU) data is readily available, this primarily comprises ICU data [Sau+22b]. See Table 2.2 for an overview. Most datasets are available after human subjects research approval and de-identification on PhysioNet [Gol+00]. The

Table 2.2: Overview of open-access medical datasets composed primarily of EHR data.

Dataset	Stays	Version	Frequency	Origin	Year	Full Dataset Name
ALOTT [LRH25]	15,000	1.0	<240 Hz	USA	2025	ALarms, Outcomes Telemetry with Timing
AUMCdb [Tho+21]	23,000	v1.0.2	>1 minute	Netherlands	2019	Amsterdam University Medical Center Database
CDSL [Rit+24]	4,479	1.0	-	Spain	2024	COVID Data for Shared Learning
EHRSHOT [Wor+23a]	6,739	3.1	-	USA	2023	A Benchmark Dataset for Few-Shot Learning in EHRs
eICU [Pol+18]	201,000	v2.0	5 minutes	USA	2017	eICU Collaborative Research Database
HiRID [Hyl20]	34,000	v1.1.1	2 / 5 minutes	Switzerland	2020	High Time Resolution ICU Dataset
INSPIRE [Lim+24]	130,000	1.2	-	South Korea	2024	INformative Surgical Patient dataset for Innovative Research Environment
MIMIC-III [JPM19]	60,000	v1.4	1 hour	USA	2016	Medical Information Mart for Intensive Care III
MIMIC-IV [Joh+23c]	73,000	v3.1	1 hour	USA	2020	Medical Information Mart for Intensive Care IV
NWICU [Mou+24]	25,000	v0.1.0	-	USA	2023	Northwestern ICU Database
PicDB [LZY20]	12,881	v1.1.0	-	China	2020	Paediatric Intensive Care Database
SICdb [Rod+23a]	27,350	v1.0.8	1 minute	Austria	2024	Salzburg Intensive Care Database
VitalDB [Lee+22a]	6,388	v1.0.0	1-7 seconds	South Korea	2018	Vital Signs DataBase
Zigong [Xu+22]	2,790	v1.0	-	China	2022	Critical care database comprising patients with infection at Zigong Fourth People's Hospital

datasets are often used for research to develop machine learning models that predict patient outcomes, identify risk factors, and improve clinical decision-making. Data quality for these datasets is often reasonable, as they are collected and preprocessed rigorously. However, the datasets may still be affected by missing data, selection bias, and confounding factors.

2.2 A Brief Overview of Temporal Modeling in Healthcare

When applying ML methodology to healthcare data, a crucial aspect is temporality; *temporal leakage*, defined as using data from the future at any current timestep, can render our approaches ineffective and our experimental results misleading. If we formalize the data as a set of observations, we can represent the data as a set of tuples, where each tuple consists of a feature vector $\mathbf{x}_i$ and a corresponding time t_i:

$$\mathcal{T} = \{(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2), \ldots, (\mathbf{x}_n, t_n)\}$$

where $\mathbf{x}_i \in \mathbb{R}^d$ represents the feature vector at time t_i, and $t_1 < t_2 < \cdots < t_n$ ensures temporal ordering; the latter is often hard to guarantee as the records may be updated without clear data provenance [Joh+23a]. We assume here that all features can be represented as real-valued vectors, which is often the case in practice, given modern embedding techniques for (unstructured) image and text data.

Table 2.3: Overview of data modalities available in these popular open-access medical datasets composed primarily of EHR data. Datasets marked with * are not primarily ICU datasets, but they do contain ICU stays. Full dataset names are found in Table 2.2.

Dataset	Medications	Vital Signs	Labs	Procedures	Clinical Notes	Waveform	Imaging	ICU
ALOTT	✓	✓	✓	✗	✓	✓	✗	✓
AUMCdb	✓	✓	✓	✓	✓	✓	✗	✓
CDSL*	✗	✓	✓	✗	✗	✓	✓	✓
EHRSHOT*	✓	✓	✓	✓	✓	✗	✗	✓
eICU	✓	✓	✓	✓	✓	✗	✗	✓
HiRID	✓	✓	✓	✓	✗	~	✗	✓
INSPIRE*	✓	✓	✓	✓	✗	○	✗	✓
MIMIC-III	✓	✓	✓	✓	✓	○	✗	✓
MIMIC-IV	✓	✓	✓	✓	✓	○	○	✓
NWICU	✓	✓	✓	✓	✓	✗	✗	✓
PicDB*	✓	✓	✓	✓	✗	✓	✗	✓
SICdb	✓	✓	✓	✓	✓	~	✗	✓
VitalDB	✗	✓	✓	✗	✗	✓	✗	✓
Zigong*	✓	✓	✓	✓	✓	~	✗	✓

Legend: ✓Available; ✗Not available; ~Partially available; ○Available in a separate database with a foreign key.

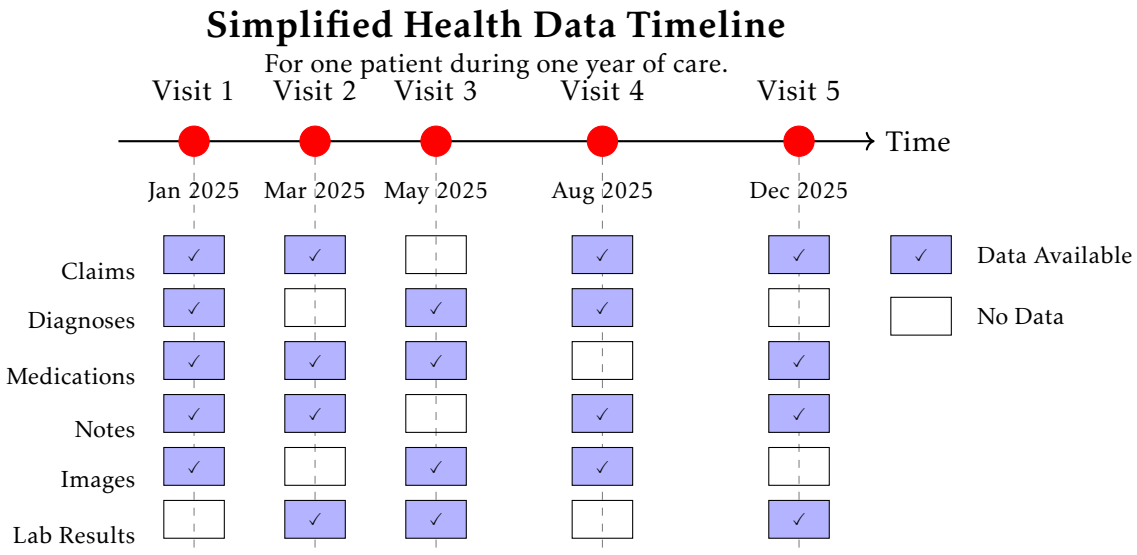


Figure 2.1: A timeline of a patient's electronic health record (EHR) events. Each event is represented by a red circle, and the data sources are indicated by boxes. The checkmarks indicate the presence of data for each event, while empty boxes indicate missing data; this can be due to a number of reasons. For example, 1) the modality was not recorded because it did not have a clinical need, 2) the data was not captured due to technical issues, 3) the patient opted out of storing certain data, 4) the patient visited a different healthcare provider, 5) hospital procedures changed.

2.2.1 Sequence Modelling Tasks

We can define different sequence modeling tasks based on the desired output and the nature of the input data. Temporal modeling is a special case of sequence modeling where there is a clear temporal ordering of data points. In some cases, the interval between time points may be simplified to be constant; however, this is often not the case or only partially so in healthcare data. We distinguish one-to-one, one-to-many, many-

to-one (sequence-to-point) and many-to-many (sequence-to-sequence) mappings. In this work, we mostly discuss many-to-one and many-to-many mappings.

One-to-one Maps a single input to a single output, but could maintain an internal temporal state:

$$f : \mathbf{x} \mapsto \mathbf{y}$$

For example, predicting the risk of sepsis based on a single snapshot of patient data at time t.

One-to-many Maps a single input to a sequence of outputs:

$$f : \mathbf{x} \mapsto \mathcal{T}_{\text{output}} = \{(\mathbf{y}_1, t_1), (\mathbf{y}_2, t_2), \ldots, (\mathbf{y}_m, t_m)\}$$

For example, generating natural language from a single medical image.

Many-to-one (sequence-to-point prediction) Maps an input sequence to a single output, often used for outcome prediction:

$$f : \mathcal{T}_{\text{input}} = \{(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2), \ldots, (\mathbf{x}_m, t_m)\} \mapsto \mathbf{y}$$

For example, our objective can be to predict ICU mortality based on the first 24 hours of patient data.

Many-to-many (sequence-to-sequence mapping) Maps input sequences to output sequences for forecasting tasks. The most straightforward manner is to predict one value per timestep:

$$f : \mathcal{T}_{\text{input}} = \{(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2), \ldots, (\mathbf{x}_m, t_m)\} \mapsto \mathcal{T}_{\text{output}} = \{(\mathbf{y}_1, t_1), (\mathbf{y}_2, t_2), \ldots, (\mathbf{y}_m, t_m)\}$$

This can help predict future vital sign trajectories or disease progression. The output can also be windowed (discretized) into discrete time steps:

$$\mathcal{T}_{\text{output}} = \bigcup_{s=1}^{S} (\mathbf{y}_s, t_{s,1}), (\mathbf{y}_{s,2}, t_{s,2}), \ldots, (\mathbf{y}_{s,n_s}, t_{s,n_s})$$

where S is the total number of discrete timesteps, $t_{s,i} \in [T_{s-1}, T_s)$ where T_s defines the boundary of discrete timestep s.

2.2.2 Supervised Learning

In supervised learning, the goal is to learn a mapping from input data $\mathbf{x}$ to output labels y using a labeled dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$. The model is trained to minimize a loss function $\mathcal{L}(f(\mathbf{x}_i), y_i)$ that quantifies the difference between predicted outputs $f(\mathbf{x}_i)$ and true labels y_i. Supervised learning can be further divided into classification and regression.

Classification

In classification tasks, the output labels y belong to a finite set of classes $\{C_1, C_2, \ldots, C_k\}$. The model learns to assign input data $\mathbf{x}$ to one of these classes. Often, classification is used for diagnosing diseases, predicting patient outcomes, or identifying risk factors. For binary classification, the model outputs a probability $\hat{y} = f(\mathbf{x})$ that the input belongs to class C_1, with the complementary probability $1 - \hat{y}$ for class C_2. The loss function commonly used is the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c})$$

Often, we simplify problems to the binary classification setting. For example, *Anomaly detection*, where we aim to identify rare events or outliers of interest, can also be modeled as a classification task, where the model learns to distinguish between normal and abnormal patterns in the data.

Regression

In regression tasks, the output labels y are continuous values. The model learns to predict a real-valued output based on the input data $\mathbf{x}$. Regression is often used for predicting continuous outcomes, such as length of stay or lab values. Moreover, regression models can also be used for time-series forecasting, where the goal is to predict future values of a variable based on its past observations.

Autoregressive Models Autoregressive models are a type of regression model. They have different connotations in statistics, although their definitions can overlap. An example is the Autoregressive Integrated Moving Average (ARIMA) model [Box+15], a traditional time-series forecasting method. Within the ML domain, we generally refer to models that predict future values based on a specific number of past observations, often referred to as the context length. Output at time t depends on (several) previous observations and predictions:

$$\hat{\mathbf{y}}_t = f(\mathbf{x}_t, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_{t-p})$$

2.2.3 Unsupervised Learning

In unsupervised learning, the goal is to find patterns or structure in the data $\mathbf{x}$ without labeled outputs y. It is also used to learn representations of the data, such as embeddings, which can then be used for downstream tasks, including classification or regression. An embedding is a learned mapping from a discrete input space to a continuous vector space. For categorical variables (such as medical codes, drug names, or procedure identifiers), we define an embedding function:

$$\mathbf{e} : \mathcal{V} \to \mathbb{R}^{d_e}$$

where $\mathcal{V} = \{v_1, v_2, \ldots, v_{|\mathcal{V}|}\}$ is the vocabulary of discrete tokens (e.g., ICD codes, medication names, but also clinical note terms), and d_e is the embedding dimension. Each token $v_i \in \mathcal{V}$ is mapped to a dense vector $\mathbf{e}(v_i) \in \mathbb{R}^{d_e}$. The training objective is often Masked Language Modeling [Dev+19], next token prediction [Ste+23], and/ or Contrastive Learning. Embeddings are used as input to a downstream model, such as a classifier or regressor, which one trains on labeled data. Outputs of this task, such as loss (e.g., cross-entropy loss or mean squared error), can then be used as feedback to retrain or fine-tune the embedding model. Moreover, Reinforcement Learning from Human Feedback (RLHF) [Chr+17] can be employed to further improve the quality of the embeddings as it is done in popular LLM-based chatbots [Cha+25].

2.2.4 Semi-supervised Learning

Semi-supervised learning combines both labeled and unlabeled data to create prediction models. For example, pretraining an embedding model on a large corpus of unlabeled clinical notes, followed by fine-tuning on a smaller labeled dataset for a specific task, such as disease classification or outcome prediction.

2.2.5 Model Families

A wide variety of models exist and there are multiple ways to categorize them; here, we provide a brief overview of three broad categories: statistical models, traditional machine learning, and neural networks/deep learning. Moreover, we separate the category of foundation models, which are large-scale pre-trained deep learning models that can be fine-tuned for specific tasks.

Statistical Models

Statistics has a long application history in healthcare from modeling costs [Mih+11] to predicting treatment effects [Imb04]. These models are often interpretable and can provide insights into the mathematical relationships between variables. Approaches include the Generalized Linear Model (GLM), Logistic Regression, and cluster analysis. However, they are often limited to a set number of input variables and may struggle with complex temporal dependencies [Raj+20]. Moreover, retrospective data analysis often does not satisfy the constraints; an Randomized Controlled Trial (RCT) or other clinical trial with a clear causal question is often required to construct these models. This makes them costly.

Traditional Machine Learning

Traditional machine learning methods partially overlap with statistical models. This category encompasses linear regression, logistic regression, decision trees, Support Vector Machine (SVM), and k-Nearest Neighbors (k-NN). These methods often rely on handcrafted features and require domain expertise to select relevant features from the data. Features could include vital signs, lab results, and demographic information in a clinical setting. These model architectures are often not inherently temporal, but can

be adapted to work with temporal data by aggregating features over time windows or using lagged variables. However, these models may struggle with complex temporal dependencies and high-dimensional (unstructured) data, which are common in healthcare applications.

Neural Networks and Deep Learning

Neural networks are a class of machine learning models inspired by the structure and function of the human brain. They consist of layers of interconnected nodes (neurons). Each neuron applies a nonlinear activation function to a weighted sum of its inputs, allowing it to “learn" complex patterns in the data. Neural networks with high dimensions are referred to as deep learning (DL). However, this usually comes at the cost of requiring large amounts of data and more computational resources. This is partially mitigated by the progress in graphics processing unit (GPU) technology, which can massively parallelize the training of deep learning models. Moreover, it is often difficult to interpret and analyze their learned mappings [Rud19]. Common temporal architectures include Recurrent Neural Network (RNN) [She20], Long Short-Term Memory (LSTM) [HS97], and transformers [Vas+17].

Large Language Models and Foundation Models Transformer architectures have become popular for natural language processing (NLP) and computer vision in recent years. Transformers use self-attention to capture long-range dependencies in data in a self-supervised manner, enabling them to process sequences of varying lengths efficiently. this, therefore, is a continuation of neural networks and deep learning. An Large Language Model (LLM), initially developed from transformer architecture, such as BERT [Dev+19] and GPT-3 [Bro+20], have demonstrated remarkable capabilities in generating human-like text, summarizing large amounts of information, and performing more complex tasks. Moreover, they have found numerous applications, including those in healthcare. Foundation models [Bom+21] are large unsupervised models that are also starting to emerge in the medical domain to learn representations of patients [Ste24; Wor+23a]. Challenges include bias in training data, the need for substantial computational resources, and the risk of overfitting to specific datasets [Wor+23b]. For a survey of current Foundation Models, see: Table 2.4. Due to the apparent drawbacks of these models, even in raw performance [Che+24; Bro+25; Pan+25a], we advocate for a comprehensive evaluation to determine if a tool is right for the job.

2.2.6 Miscellaneous terminology

Multi-modal models

Healthcare data is often multimodal, combining various data types such as structured data (e.g., lab results), unstructured text (e.g., clinical notes), images (e.g., X-rays), and time-series data (e.g., vital signs).

Multi-modal models use models, ensembles of models, and data fusion (preprocessing) to integrate this information. One can classify these techniques in the segment of the modelling pipeline where they are applied: early, intermediate, and late

Table 2.4: Foundation models for EHR data. We summarize the datasets that were used for evaluation. How they utilized tokenization and what type of pretraining loss they employed to train the models. Adapted from Pang et al. [Pan+25a], used under CC BY 4.0 license, and expanded to include more characteristics and models. This table is further expanded upon in Chapter 5.

Model	Dataset	Temporal Repr.	Tokenization	Pretrain Loss
BEHRT [Li+20]	Clinical Practice Research Datalink [Her+15]	PE & Age Embedding	Code-based	MLM
Med-BERT [Ras+21a]	Cerner Health Facts and Truven Health MarketScan	PE & Age Embedding	Code-based	MLM
CLMBR [Ste+21]	Stanford Medicine [Cal+23]	RoPE	Code-based	NTP
MedGPT [Kra+21]	London Hospital + MIMIC-III	RoPE	Text & Code based	NTP
CEHR-BERT [Pan+21]	CUMC-NYP	Time2vec +ATT	Code-based	MLM
GatorTron [Yan+22]	UF Health, Pubmed, Wikipedia + MIMIC III	None	Text-based	MLM & SOP
GenHPF [Hur+24]	MIMIC-III, IV + eICU	ALiBi	Text-based	SSL
Foresight [Kra+24]	London Hospital + MIMIC-III	Sinusoidal	Text & Code based	NTP
MOTOR [Ste+23]	Stanford Medicine + MERATIVE[18]	RoPE	Code-based	TTE
TransformEHR [Yan+23b]	VA Bedford	Sinusoidal	Code-based	Visit MLM
EBCL [Ouf+24b]	MIMIC-IV [Joh+22]	Custom	Code-based	EBCL
CEHR-GPT [Pan+24]	CUMC-NYP	ATT	Code-based	NTP
CORE-BEHRT [Odg+24]	Capital Region of Denmark	RoPE	Code-based	MLM
EHRMAMBA [Fal+24]	MIMIC-IV	Time2vec	Code-based	NTP
ETHOS [Ren+24]	MIMIC-IV	Age quantiles + ATT	Code-based	NTP
CEHR-XGPT [Pan+25b]	CUMC-NYP	ATT	Code-based	TD + TTE
Benchmark	**Dataset**	**Temporal Repr.**	**Tokenization**	**Pretrain Loss**
Long-context LLAMA, MAMBA [Wor+24]	Stanford Medicine	RoPE	Code-based	NTP
MEDS-DEV [McD+24]	Various*	Various*	Various*	Various*
FoMoH [Pan+25a]	CUMC-NYP	Various*	Code-based	Various*

MLM: Masked language modeling, NTP: Next token prediction, SOP: Sentence order prediction, SSL: Self-supervised learning, TTE: Time to event, EBCL: Event-based Contrastive Learning, ATT: Artificial Time Tokens, TD: Temporal Denoising Objective, RoPE: Rotary Position Embedding
*Benchmarks for foundation models for EHR data have various temporal representations and pretraining objectives.

fusion [ZZG24]. *Early fusion* combines raw data from different modalities before feeding it into a model. *Intermediate fusion* combines features extracted from each modality using separate models before making predictions. *Late fusion* combines predictions from separate models trained on each modality: an ensemble method. Moreover, various models have been proposed to be able to model multimodal data more effectively. For example, Sutter et al. [Sut+24] proposes a multi-model variational auto-encoder to learn a shared latent representation across modalities. Moreover, models built on Contrastive Language-Image Pretraining (CLIP) [Rad+21] have been widely adopted in integrating text and image data for medical applications [Zha+25b].

Causality

In healthcare applications, understanding causal relationships is important for effective clinical decision-making; this applies to any considered clinical interventions. Hospital processes severely influence the *data generating process* and can lead to implementing flawed systems [Jos+25]. While traditional machine learning focuses on prediction through statistical associations, causal inference seeks to identify relationships where interventions on one variable directly influence outcomes in another. This distinction becomes critical in healthcare, where correlation-based predictions may lead to spurious treatment recommendations or biased clinical policies.

Alarm Systems

Alarm systems are an application of temporal modeling in healthcare, where the goal is to predict critical events or conditions (anomalies) based on real-time data streams. These systems often use machine learning to identify patterns indicative of impending events, such as patient deterioration or equipment failure. By continuously monitoring patient data, alarm systems can provide clinicians with early warnings, enabling timely interventions and potentially improving patient outcomes. We typically use an 'online' approach where the system updates every timestep t. The model predicts the probability of an event occurring in the future based on the data observed up to time t_n:

$$\hat{p}_{t_n} = f_\theta\left(\mathcal{T}_{\leq t_n}\right) = \mathbb{P}(\text{event in } [t_n, t_{n+1}, \ldots] \mid \mathcal{T}_{\leq t_n})$$

Where f_θ is the machine learning based system (or outputs of the model) parameterized by θ, $\mathcal{T}_{\leq t_n} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_n, t_n)\}$ is the sequence of all observed feature vectors up to and including time t_n. An alarm is triggered if $\hat{p}_{t_n}$ exceeds a threshold τ:

$$\text{Alarm at } t_n \iff \hat{p}_{t_n} \geq \tau$$

Then we define a binary alarm indicator a_{t_n} to create the outputs for the alarm system:

$$a_{t_n} = \begin{cases} 1, & \text{if } \hat{p}_{t_n} \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

Where a_{t_n} is the binary alarm indicator at time t_n, $\hat{p}_{t_n}$ is the predicted probability of an event in the prediction window, and τ is the alarm threshold. Note that we can make τ dependent on a function of time, the previous predictions of the model, and the specific patient, which allows for dynamic thresholding based on the context of the prediction. This technique is particularly beneficial in healthcare, where patient conditions and risk factors can vary significantly, and alarms must be tailored accordingly to enhance usability. Moreover, we can report additional metrics, such as confidence scores, uncertainty estimates, and summary statistics, to inform intervention decisions.

2.3 Challenges in Machine Learning for Healthcare

While traditional machine learning often assumes independent and independent and identically distributed (i.i.d.) data, healthcare applications frequently involve a variety

of challenges. The following subsections outline these categories, which were derived from a systematic meta-review; Appendix B.1 provides details on the search methodology. In this work, we address a combination of these challenges in each chapter; however, the challenges outlined here remain active areas of research in ML4H and specific engineering efforts within health systems are required to resolve them for every implemented system (see, for example, subsection 2.3.4).

2.3.1 Dataset Quality and Availability

Access to sizeable, high-quality datasets is a major challenge for current medical prediction systems. This is mentioned in several systematic reviews [Si+21; Med+19; Sta+22; Elf+21; XCS18; Meh+17]. Multiple specific problems are present in this field. Clinical observations are often recorded at *irregular intervals*, with vital signs measured more frequently for critically ill patients and less often for stable ones, resulting in irregular sampling. Data is collected across *multiple time scales*, from high-frequency physiological signals to long-term disease progression, and often comes from various modalities such as structured lab results, unstructured clinical notes, and time-series vital signs. Integration of these data modalities is often complex. Healthcare practices, patient populations, and measurement technologies evolve over time, leading to *concept drift* [Rot+23] that can degrade model performance. Moreover, ensuring the *privacy and security* of sensitive patient data is critical, with regulations such as Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR) necessitating techniques like federated learning [Pfi+24]. Another issue is the use of small sample sizes or rare events, such as rare diseases or adverse drug reactions, which challenge model development and evaluation [Med+19; Elf+21; Sta+22].

2.3.2 Fairness and Generalization

Bias in healthcare data can result in unfair or inaccurate predictions, particularly for underrepresented populations, making *fairness* a key consideration [Jea24; Do+23]. Addressing bias and ensuring fairness in machine learning models minimizes the risk of perpetuating health disparities, which are present in much of the data used to train models. Moreover, models trained on specific datasets may not generalize well to other populations or healthcare settings, even within the same institute, highlighting the importance of (external) *validation* [Roc+25] with unseen datasets. Domain adaptation and domain generalization [Zho+23] are other possible solutions to this problem. True generalizability may not be achievable due to inherent differences in healthcare operating procedures, model goals, and patient populations [Fut+20].

2.3.3 Reproducibility and Comparability

As in many areas of science, reproducibility is a pressing matter [AJ25]. Computer Science, especially the area of Machine Learning, is facing similar challenges [Sem+25]. A strict methodological approach to studies is often essential for bringing theoretical innovations into practice in the medical domain. Several works in our literature review have mentioned the problem of the lack of reproducibility [Sye+21a; Moo+21b;

Med+19; Kun+21; KVM17]. As mentioned in [Sye+21a], clinical variables, data preprocessing, and model specifications are often not reported. Additionally, [Moo+21b] mentions that meta-analyses are impossible because of the lack of details provided in the current work. [Med+19] Points towards the fact that many studies only cover datasets from individual centers, limiting comparability between approaches. Moreover, Kunze et al. [Kun+21] describes the lack of sharing code or complete models as a barrier to further development. Finally, Kamio et al. [KVM17] and Muralitharan et al. [Mur+21] mention that studies should use a standard set of performance metrics that show realistic results for datasets with a low prevalence of "positive" cases (e.g., the endpoint "death after surgeries" commonly does not exceed 5% of all recorded cases [Mur+21]).

2.3.4 Clinical Validation

Validation of completed systems is a critical step for their effective use in practice. Machine learning models trained on specific datasets may not generalize to other populations or healthcare settings. External validation and domain adaptation techniques could improve performance across diverse environments. Several literature reviews have pointed out the necessity of prospective validation in addition to the often-performed retrospective validation [Sye+21b; Mur+21; XCS18; Meh+17; Sta+22; KVM17]. All works mention that, for the practical application of predictive modelling in critical care, prospective studies need to be conducted. Additionally, Mehta et al. [Meh+17] notes that health datasets are often incomplete because data collection has not yet been completed; this could result in the model performing considerably worse in prospective studies. Muralitharan et al. [Mur+21] adds that the contribution to "alarm-fatigue" should be carefully studied in a practical setting when implementing alert systems. Alarm or alert fatigue describes the phenomenon in which medical personnel are overwhelmed by alerts generated by hospital equipment; these alerts are often of questionable necessity for the patient's situation. Furthermore, Xiao et al. [XCS18] poses that the automation of training and deployment within a privacy-ensured medical practice might hinder validation in practice. Finally, we note that deploying machine learning models in real-world clinical settings is challenging due to the healthcare workflow's (undocumented) complexity. Models must be integrated into existing workflows and provide actionable insights to clinicians [Boa+24].

2.3.5 Interpretability and Transparency

To effectively integrate artificial intelligence into daily practice, it is essential to prioritize the interpretability of proposed actions and patient diagnoses. In recent years, numerous studies have specifically addressed the "black box" issue in machine learning and deep learning, where these problems are exacerbated. The critique of many predictive systems is that they might predict specific outcomes without indicating why these outcomes were predicted. Ideally, Clinical decision-making requires an understanding not just of what the model predicts, but also why and when these predictions are made, particularly for models with multiple modalities, samples, timesteps, and features. Here, we encounter a paradox in which we want to use complex, non-linear interactions to predict outcomes, while keeping the predictions as simple as possible

such that they can be “explained". We recognize that this topic is also discussed in several of the analyzed works [Sye+21a; Xie+22; Si+21; Mur+21; XCS18; Sta+22]. Syed et al. [Sye+21a] Poses that only two of the analyzed 61 studies addressed this concern by providing visualizations for medical professionals. Furthermore, Xie et al. [Xie+22] points out the difference between post-hoc and inherently interpretable models: the latter are considered more desirable, as they do not leave the explainable aspect for developer interpretation. Xie et al. [Xie+22] mentions mimic learning and attention interpretation mechanisms as solutions to interpretability. The former "simplifies" a deep learning model by training a simpler and inherently more interpretable model (e.g., a decision tree). The concept of attention is used to show the relevance of specific data sequences for their predictions. Xiao et al. [XCS18] mentions that Deep Learning models often do not capture uncertainties and describes the concept of mimic learning (seen in Xie et al. [Xie+22]) as knowledge distillation. Lastly, Stam et al. [Sta+22] describes that explainability is important in gaining trust for application in practice.

3

Establishing Reproducibility in Open-access EHR Prediction Modeling

This section largely mirrors the work Van de Water et al. [Van+24a] published at the International Conference on Learning Representations 2024. Additional experiments, related work, and descriptions have been added that were formerly only in the Appendix. We also outline the new contributions that have been added since publishing in subsection 3.3.10. It also contains contributions outlined in Van de Water et al. [VSR26], published at the Journal of Open Source Software (JOSS) in 2026, and has been extended with additional context and explanations. Moreover, the extensions from Turubayev et al. [Tur+25] are partially covered, though, the results of this work are not included in this chapter.

As indicated in chapter 2, intensive care unit (ICU) has long been a focus for research into data-driven decision support, owing to the impact of medical decisions as well as the breadth and depth of data collected in this setting [JPM17]. The COVID-19 pandemic confirmed the need for reliable machine learning (ML)-based clinical decision support that can alert healthcare professionals to worsening patient states, help them make a clinical diagnosis, or recommend treatment [Med+19].

3.1 Motivation

Despite a steep increase in the number of published ICU prediction models [Shi+19a], hardly any have made their way into clinical practice [Ein+22; Fle+20b]. A major obstacle to translation is an ongoing lack of comparability and reproducibility [JPM17]. By using custom datasets and definitions, preprocessing pipelines, and evaluation schemes, the benefits of novel models are conflated with differences between patient case mix, task definitions, and cohort selection [Sar+23; Kel+19a]. Reviewing models for early prediction of sepsis, for example, Moor et al. [Moo+21b] found that the definition of sepsis, the time of prediction, and the available features differed substantially between

the 22 included studies; similar results were found in an earlier review [Fle+20a]. Even among studies from the same research group [Hyl20; Yèc+22], cohort definitions may vary substantially, precluding a meaningful comparison. Inconsistencies in imputation and feature extraction further complicate an objective evaluation of research progress.

The increasing availability of open-access EHR datasets — of which ICU data represent the most comprehensive category — is a first, important step towards urgently needed model comparability [Sau+22a]. However, models derived from the same dataset may still vary considerably in their analytical setup. Earlier work has therefore created benchmarks that establish a single pipeline for preprocessing and modeling [Yèc+22; Har+19; SBO20]. These benchmarks are hard-coded for a given dataset, following proprietary formats and supporting a limited, fixed set of tasks. Extending an existing benchmark to include new datasets or tasks requires changes to the benchmark's — often lightly documented — source code. Despite the existence of multiple benchmarks, new models are therefore rarely evaluated on more than one dataset or do not use *any* benchmark [Shi+19b].

We address this gap by providing *Yet Another ICU Benchmark* (*YAIB*) as a modular multi-dataset framework specifically designed for extensibility. Building on recent work to harmonize ICU data [Ben+23] (i.e., match time-scale, clinical definitions, and units across datasets), we standardize the entire modeling workflow from the definition of clinical concepts (a medical abstraction to facilitate patient care) and data extraction to model fitting and evaluation across several established open-source ICU datasets [Sau+22a]. We provide a predefined set of common prediction tasks, developed in collaboration with clinical intensivists, that can be easily extended to fit user needs. Our benchmark, by default, provides endpoint prediction for ICU mortality, sepsis [Sin+16], Acute Kidney Injury (AKI) [KDI12], Kidney Function (KF), and Length of Stay (LoS). With this work, we aim to **(1)** dramatically reduce the overhead of developing new ICU prediction methods, **(2)** provide a transparent, open-source, and reproducible definition of experiments, and **(3)** unify ML workflows for ICU prediction modeling.

3.2 Related Work

Our work builds upon several previous efforts to harmonize the definition, development, and evaluation of ICU prediction models. *YAIB* combines these existing works in a novel, end-to-end fashion to enable quick, reproducible, and comparable model development.

3.2.1 Benchmarks

To improve comparability between models trained on these ICU datasets, several benchmarks or benchmark-like applications have been developed (Table 3.1). These solutions mainly differ in the tasks and models they support. Notably, existing benchmarks heavily focus on benchmarking results, often hardcoding key steps like data extraction, task definition, preprocessing, feature generation, and sometimes model training. While they may reduce implementation overhead when evaluating new ML approaches, the present benchmarks are difficult to adapt to user requirements. Core code base changes

are often necessary if the users' problems do not fit into the provided task definitions. Even advanced modeling frameworks such as [Jar+21] and [Sv23] share this weakness, as they do not support reproducible data extraction or task definitions; thus, they do not provide an end-to-end solution like *YAIB*.

3.2.2 Multi-dataset Support

Due to considerable heterogeneity in data structure, existing benchmarks tend to focus on a single dataset, most frequently MIMIC-III [JPM17; Pur+18; Har+19; Bar+20; Wan+20b; Jar+21]. As MIMIC-III also has a large existing user base [Sye+21b], it thus often becomes the default choice [Shi+19a]. This has potentially resulted in a self-enforcing bias towards the MIMIC-III datasets, which represent a single-center US population. Even frameworks that work with its successor MIMIC-IV lack backward compatibility [Man+21; Gup+22]. Among the few multi-dataset solutions, Tang et al. [Tan+20] operates on both eICU and MIMIC-III, but lacks many of the model architectures found in other works and no longer appears to be in active development. Oliver et al. [Oli+23] provides a hardcoded pipeline to combine several datasets without providing cohort definitions, benchmarking, or an end-to-end pipeline. Finally, Yang et al. [Yan+23a] recently proposed PyHealth as a comprehensive deep learning toolkit for both ML researchers and healthcare practitioners; it is perhaps most closely related to our work. Unfortunately, PyHealth only supports subsets of the full datasets, and tasks must be defined anew for each dataset. It also does not currently include time series or ways to deal with missing data, limiting its use for novel clinical or ML developments.

3.2.3 Data Harmonization

Our benchmark uses the `ricu` [Ben+23] R package as a basis for the harmonization of the different ICU databases, to which we add a Python interfacing capability and cohort-building functionality. Other data harmonization attempts preceded `ricu` [Ben+23]. Wang et al. [Wan+20b] and Adibuzzaman et al. [Adi+16] developed solutions that solely work for MIMIC-III and do not natively support dataset interoperability. Similarly, Gupta et al. [Gup+22] and Mandyam et al. [Man+21] exclusively operate on MIMIC-IV, lacking backward compatibility of developed methods. Note that we have excluded Mandyam et al. [Man+21], from Table 3.1 as it does not provide any ML-benchmarking capabilities. Several further tools have been developed for processing EHR data [Jar+21; Tan+20]. In Wang et al. [Wan+20b], the authors introduce a preprocessing pipeline for the MIMIC-III dataset. This work is limited to use with just MIMIC-III, even though MIMIC-IV has been published [Joh+22]. Tang et al. [Tan+20] operates on both eICU and MIMIC-III but does not include many of the model architectures found in our work. Another package that has been in development since 2021, focused especially on model availability, is PyHealth [Zha+21]; it does not provide dataset interoperability, disregards a large part of the datasets, provides a limited amount of hardcoded tasks, no imputation methods, and a very limited, non-customizable preprocessing pipeline.

Finally, Jarrett et al. [Jar+21] provides a comprehensive pipeline for the use of AutoML. Still, it does not provide a harmonized end-to-end — raw dataset to model performance — benchmarking pipeline for testing new models, tasks, and datasets. Ad-

ditionally, it does not include built-in extensible support for established open-access datasets (excluding MIMIC-III and closed-source datasets).

Although we provide the definition of a number of tasks, we encourage and facilitate the creation of new tasks and appropriate cohorts. With *YAIB* we can train several common models (see Table 3.1 and Section 3.5.2) to perform these prediction tasks, of which implementation details (e.g., Temporality and Type) have been aligned with conventions described in earlier work. These benchmark models are defined for a particular prediction task and a preprocessed cohort. A cohort defines the subset of patients selected according to the available data points. For our cohorts, we attempt to predict a patient's mortality from the first 24 hours of a stay. Additionally, we excluded patients who died or were discharged within 30 hours of ICU admission to prevent data leakage. Our Acute Kidney Injury (AKI) cohort was based on the definition defined by Khwaja [KDI12], where the Kidney Disease Improving Global Outcomes (KDIGO) stage was ≥ 1. Sepsis-3 is defined according to Singer et al. [Sin+16]. Organ dysfunction is defined as an increase of the Sequential Organ Failure Assessment (SOFA) of ≥ 2 points. The patient's risk of sepsis onset in the next 6 hours is predicted every hour of the ICU stay.

ICU Mortality One of the most common tasks for ICU datasets is the prediction of mortality within the ICU. The MIMIC-III dataset [Joh+16] has been used in over 20 of the total 61 works (34%), selected through a systematic review [Sye+21b].

Acute Kidney Injury (AKI) This is a common prediction task performed in combination with ICU EHR data [Mey+18; Koy+18a].

Sepsis A common cause of mortality is the phenomenon medically defined as Sepsis. It is commonly defined as an organ dysfunction caused by an infection, the most recent definition is sepsis-3 [Sin+16]. It is of vital importance that antimicrobial treatment is applied at an early stage.

Circulatory Failure Circulatory failure prediction has been explored in earlier works, specifically on the HiRID dataset [Hyl20]. As circulatory failure is a common event in several critical illnesses, it is a relevant prediction endpoint for our benchmark. It also provides a basis for universal evaluation of models applied in the work of Hyland et al.[Hyl20].

Length of Stay (LoS) In a public health crisis, but also during regular operation, keeping ICU stations in operation is costly and takes a considerable toll on medical personnel. Additionally, efficient ICU throughput and planning are vital to the operation of a hospital. We provide a length-of-stay task to predict the remaining hours spent in the ICU.

3.2.4 Comparison with Existing Frameworks

Several frameworks have been proposed before that only partially address the challenges of reproducibility and transparency in clinical ML.

Table 3.1: Comparison of existing benchmarks and *YAIB* on ICU data, ordered by publication date.

			Johnson et al. [JPM17]	Purushotham et al. [Pur+18]	Harutyunyan et al. [Har+19]	Barbieri et al. [Bar+20]	Wang et al. [Wan+20b]	Jarrett et al. [Jar+21]	Sheikhalishahi et al. [SBO20]	Tang et al. [Tan+20]	Yèche et al. [Yèc+22]	Mandyam et al. [Man+21]	Gupta et al. [Gup+22]	Yang et al. [Yan+23a]	Saveliev et al. [Sv23]	Oliver et al. [Oli+23]	**Van de Water et al. [Van+24a]**
Datasets	MIMIC-III		✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✓	✗	✓	✓
	MIMIC-IV		✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓	✓
	eICU		✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✓	✗	✓	✓
	HiRID		✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓
	AUMCdb		✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓
Prediction tasks	Mortality risk		✓	✓	✓	✗	✓	✗	✓	✓	✓	✓	✗	✓	✗	✗	✓
	Circulatory failure		✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	✗	✗	*
	Kidney Function (KF)		✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	✗	✗	✓
	Respiratory failure		✗	✗	✗	✗	✗	✓	✗	✓	✓	✗	✓	✗	✗	✗	*
	Sepsis		✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
	Acute Kidney Injury		✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
	Phenotyping§		✗	✓	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗	*
	Interventions		✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	*
	Length of Stay (LoS)		✗	✓	✓	✗	✓	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓
	Readmission§		✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	*
Preproc.	Feature engineering		✗	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✗	✓	✗	✓
	Temporal imputation		✗	✗	✗	✗	✓	✓	✗	✓	✓	✓	✓	✗	✓	✗	✓
	Temporal resampling		✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✓
	Modular pipeline		✗	✗	✗	✗	✗	✓	✗	✓	✗	✗	✓	✗	✓	✗	✓
Model architectures	ML	LR	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗	✓	✗	✗	✗	✓
		Random forest	✗	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗	✓
		Gradient boost	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗	✓
	DL	RNN	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗	✓	✓	✗	✓
		LSTM	✗	✗	✓	✗	✗	✓	✓	✓	✓	✗	✓	✓	✓	✗	✓
		GRU	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	✗	✓	✓	✗	✓
		Temporal CNN	✗	✗	✗	✗	✗	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓
		Transformer	✗	✗	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✓
Code available			✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Extensible†			✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✓
Dataset interoperability‡			✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓

*: These tasks are not included by default but may be easily added through our cohort definition pipeline.
§: Due to lack of recorded database information, these tasks can only be defined for MIMIC III and IV.
†: Interface and extensive instructions to add interoperable modules following a provided abstraction (datasets, prediction tasks, models) and adjust existing modules without extensive rewriting or refactoring.
‡: Provides an uncoupled interoperable dataset definition, allowing a.o. transfer learning and domain adaption.

Clairvoyance

Clairvoyance [Jar+21] is "a Unified, End-to-End AutoML Pipeline for Medical Time Series". As such, it does not focus on ICUs or benchmarking but instead standardizes model learning (imputation and training), model evaluation, and model selection, focusing on the computational aspects of developing a model. Clairvoyance comes with some code to define a task for estimation of treatment effects on MIMIC III data. However, this task is hard-coded and lightly documented, primarily as a Clairvoyance demo. The exemplary nature of this task is further illustrated by the fact that, at no point the authors mention possible confounders/colliders of the treatment effect and whether they are conceivably adjusted for by the covariates, rendering any causal interpretation

moot. It is unclear how this task can be easily adapted or extended to other databases without significant amounts of custom code.

The Advantages of *YAIB* Compared to Clairvoyance are that *YAIB* puts ICU data and tasks front and center. *YAIB* supports the whole workflow, from raw data to clinical concepts to well-defined cohorts. This approach greatly facilitates the transparent and reproducible preprocessing of (often messy) ICU data, which Clairvoyance does not cover. We strongly believe that unless tasks can be adapted easily and reproducibly, it will lead to inevitable ad-hoc adaptations of the task that often end up irreproducible. *YAIB*, therefore, improves on existing modeling frameworks by putting an equal emphasis on standardized data processing for meaningful model development.

TemporAI

TemporAI [Sv23] is a package that is currently in early development without a peer-reviewed publication associated with it. While it promises to provide: "prediction, causal inference, and time-to-event analysis, as well as common preprocessing utilities and model interpretability methods," it is unclear from current documentation how to use established datasets with this package or how to use relevant medical prediction tasks.

The Advantages of *YAIB* Compared to TemporAI are similar to those between *YAIB* and Clairvoyance: *YAIB* puts ICU data and tasks front and center for both ML scientists and clinicians. *YAIB* supports the whole workflow, from raw data to clinical concepts to well-defined cohorts. This approach greatly facilitates the transparent and reproducible preprocessing of (often messy) ICU data, which TemporAI, similarly to Clairvoyance, does not cover. However, we would like to note that using TemporAI (or Clairvoyance) with the *YAIB* pipeline to create a different end-to-end pipeline is possible, as it allows for "swapping out" components. We provide the functionality in our *YAIB*-`cohorts` repository to convert any cohort to a format compatible with Clairvoyance and TemporAI.

PyHealth

PyHealth [Yan+23a] is "a comprehensive deep learning toolkit designed for both ML researchers and healthcare practitioners." PyHealth aims to support all EHR databases. It is thus similar in scope to our proposed framework. Unfortunately, upon closer inspection, PyHealth only supports a small subset of the information in MIMIC and eICU. While diagnoses and prescriptions are, in theory, included, they are processed as a simple bag of diagnosis codes or drug codes without information on strength/duration or semantic interpretation of what they represent (e.g., what is a vasopressor needed in calculating the SOFA score). Vital signs are not supported at all, presumably because PyHealth reads information from raw .csv files and may struggle to process large quantities of vital sign data. PyHealth further states that the datasets are independent of task definitions. This, unfortunately, appears to mean that they have to be implemented anew for each database, with custom dataset-specific code for the same task.

Furthermore, all currently available ICU tasks in PyHealth use static data only and do not include any time series.

Advantages of *YAIB* compared to PyHealth: *YAIB* supports all databases within a common, principled interface (see the response on data harmonization above). Moreover, *YAIB* enables a single task definition that one can directly use for any included dataset. As far as they can work with time series data, *YAIB* can incorporate any model defined in PyHealth.

3.2.5 Publicly Available ICU Datasets

Our benchmark currently supports four established ICU datasets [Sau+22b]: the Medical Information Mart for Intensive Care (MIMIC) version III [Joh+16] and IV [Joh+23c], the eICU Collaborative Research Database (eICU) [Pol+18], the High Time Resolution ICU Dataset (HiRID) [Hyl20], and the Amsterdam University Medical Center Database (AUMCdb) [Tho+21]. These datasets contain similar data items but differ in size and scope (Table 3.2). Together, they cover 334,812 ICU stays— from the USA and Europe. We plan to integrate three recently released ICU datasets in the future [Rod+23a; Jin+23].

3.2.6 Dataset Characteristics

The Medical Information Mart for Intensive Care (MIMIC)-III dataset is the most commonly used for ML in ICU settings; [Sye+21b] found 61 eligible studies that used a form of the MIMIC dataset. It was collected in the USA at the Beth Israel Deaconess Medical Center [Joh+16]. The newer MIMIC-IV includes several improvements, among which are newer patient records and a revised structure including regular hospital information [Joh+23c].The eICU Collaborative Research Database (eICU) [Pol+18] is an effort to collect the first sizable (200,000 admissions) multicenter dataset. It was collected using Philips ICU monitoring systems in the USA at 208 participating hospitals. The High Time Resolution ICU Dataset (HiRID) dataset was collected at Bern University Hospital, Switzerland, and has incorporated more observations than the aforementioned datasets [Hyl20]. The Amsterdam University Medical Center Database (AUMCdb) is the most recently released ICU dataset [Tho+21]. Collected in the Netherlands, it has a temporal resolution of up to 1 minute and has prioritized patient de-identification. Note that there is no benchmark software for this dataset yet. Each dataset we use has undergone de-identification procedures, and we have not tried to re-identify the people involved, as per the user agreement for each dataset. Table 3.2 shows some key characteristics of each dataset. A more comprehensive overview of ICU datasets can be found in the work of [Sau+22b].

The authors of MIMIC-III and eICU have made small selected datasets available for the purpose of experimentation. These datasets are also publicly available on PhysioNet. We support the publicly accessible "demo" datasets provided for eICU[1] and

[1] https://physionet.org/content/eicu-crd-demo

Table 3.2: Key characteristics of openly accessible ICU datasets. Note that accessing each dataset requires completing a credentialing procedure.

Dataset	MIMIC-III / IV	eICU	HiRID	AUMCdb
Stays*	40k (0.1k)** / 73k	201k (2k)	34k	23k
Version	v1.4 / v2.2	v2.0	v1.1.1	v1.0.2
Frequency (time-series)	1 hour	5 minutes	2 / 5 minutes	up to 1 minute
Origin	USA	USA	Switzerland	Netherlands
Originally published	2015 [Joh+16] / 2020 [Joh+23c]	2017 [Pol+18]	2020 [Hyl20]	2019 [Tho+21]
License	A (C) / A	A (C)	A	B
Repository link	ρ (ρ)/ ρ	ρ (ρ)	ρ	α

Note that accessing each full dataset requires completing a credentialing procedure.
*: Stays were taken and rounded from the latest available versions of the databases as of the time of writing.
**: The brackets () indicate characteristics of the demo (freely accessible) version of the dataset
A: PhysioNet Contributor Review Health Data License 1.5.0
B: Access Request Form and End User License Agreement for AmsterdamUMCdb 1.6
C: Open Data Commons Open Database License v1.0
ρ: PhysioNet
α: Amsterdam Medical Data Science

MIMIC-III[2]. Per the demo dataset license (Open Data Commons Open Database License v1.0, see Table 3.2, License C), it is permitted to adapt and share the data. Still, we recommend the user complete a human subject research training to ensure the dataset's usage does not violate the usage proposal. They contain respectively 2,500 (eICU) and 100 stays (MIMIC-III) before exclusion. For the purposes of testing and validating *YAIB*, we have created demo-cohorts, *extracted solely from these datasets*, for each of our supported tasks. Usage of the task cohorts and dataset is only permitted in accordance with the above license.

3.2.7 Exclusion Criteria

We included all available ICU stays of adult patients in our analysis. For each stay, we applied the following exclusion criteria to ensure sufficient data volume and quality: remove any stays with **1)** an invalid admission or discharge time defined as a missing value or negative calculated length of stay, **2)** less than six hours spent in the ICU, **3)** less than four separate hours across the entire stay where at least one feature was measured, **4)** any time interval of ≥12 consecutive hours throughout the stay during which no feature was measured. Figure 3.1 details the number of stays overall and by dataset excluded this way. The task specific cohorts can be found in section C.2.

[2] https://physionet.org/content/mimiciii-demo

Table 3.3: Characteristics of 1) the included datasets (above) and 2) the task cohorts (below).

General characteristics	AUMCdb	HiRID	eICU	MIMIC-IV
Version	1.02	1.1.1	2.0	2.0
Number of patients	19,790	-[1]	160,816	53,090
Number of ICU stays	22,636	32,338	182,774	75,652
Age at admission (years)	65 [55, 75][2]	65 [55, 75]	65 [53, 76]	65 [53, 76]
Female	7,699 (35)	11,542 (36)	83,940 (46)	33,499 (44)
Race				
Asian	-	-	3,008 (3)	2,225 (3)
Black	-	-	19,867 (11)	8,223 (12)
White	-	-	140,938 (78)	51,575 (76)
Other	-	-	16,978 (9)	5,514 (8)
Unknown			1,983	8,115
Admission type				
Medical	4,131 (21)	-	134,532 (79)	49,217 (65)
Surgical	14,007 (72)	-	31,909 (19)	25,674 (34)
Other	1,225 (6)	-	4,702 (3)	761 (1)
Unknown	1,069	-	11,631	0
Hospital length of stay (days)	-	-	6 [3, 10]	7 [4, 13]
Task cohorts	**AUMCdb**	**HiRID**	**eICU**	**MIMIC-IV**
ICU mortality				
Number of included stays	10,535	12,859	113,382	52,045
Died	1,660 (15.8)	1,097 (8.2)	6,253 (5.5)	3,779 (7.3)
Onset of acute kidney injury				
Number of included stays	20,290	31,772	164,791	66,032
KDIGO* ≥ 1	3,776 (18.6)	7,383 (23.2)	62,535 (37.9)	27,509 (41.7)
Onset of Sepsis				
Number of included stays	18,184	29,894	123,864	67,056
Sepsis-3 criteria	764 (4.2)	1,986 (6.6)	5,835 (4.7)	3,730 (5.6)
Kidney function (creatinine)				
Number of included stays	8,003	7,499	69,117	35,657
Creatinine value	0.97 [0.70, 1.61]	0.92 [0.67, 1.50]	1.00 [0.71, 1.68]	1.00 [0.70, 1.60]
ICU remaining length of stay				
Number of included stays	22,636	32,338	182,774	75,652
ICU length of stay (hours)	24 [19, 77]	24 [19, 50]	42 [23, 76]	48 [26, 89]

[1] HiRID only provides stay-level identifiers.
[2] Since AUMCdb only includes age groups, we calculated the median of the group midpoints.
* KDIGO, Kidney Disease Improving Global Outcomes [KDI12].
Numeric variables are summarized by *median [Interquartile Range (IQR)]*.
Categorical variables are summarized by *incidence (%)*.

3.3 Benchmark Design

YAIB addresses the issues identified above and provides a unified interface to develop clinical prediction models for the ICU. An experiment in *YAIB* consists of four steps: **1)** define clinical concepts from the raw data; **2)** extract the patient cohort and specify the prediction task; **3)** preprocess the data and generate features; and **4)** train and evaluate the ML models (Figure 3.2).

3.3.1 Design Philosophy

We strongly believe that medical research is inherently complex and that — rather than providing a rigid benchmark — there lies most value in providing a modular setup where the user can exchange any part with something that better suits their needs and, importantly, do so reproducibly. For example, users frequently want to highlight a par-

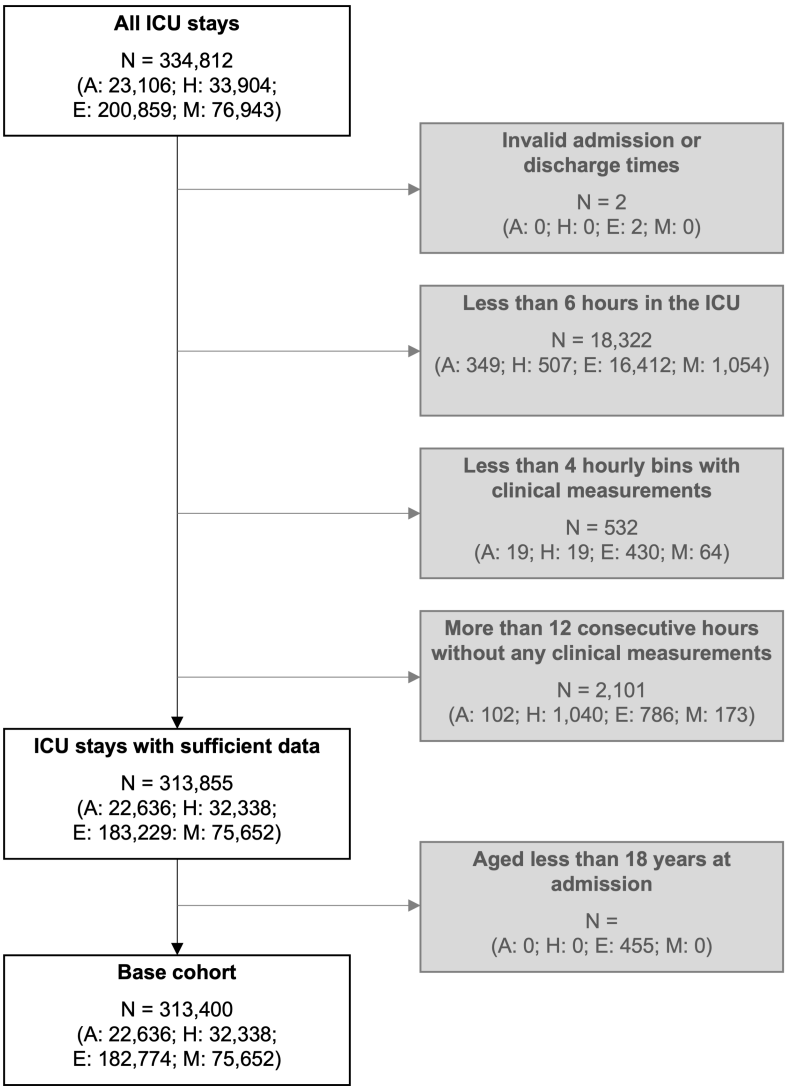


Figure 3.1: Exclusion criteria applied to the base cohort for all datasets. **N**: Total amount of cases. **A**: AUMCdb, **H**: HiRID, **E**: eICU, **M**: MIMIC-IV

ticular aspect of their model, prompting them to adapt the default tasks. Changes, however minor, can render results incomparable. We, therefore, prioritized extensibility across the entire experiment lifecycle. This high level of extensibility may increase the complexity of our benchmark. We mitigate this by providing a range of default experiments for users with limited access to medical expertise or content with a fixed set of medical tasks. The experiments were designed to be directly comparable and provide a common benchmark. This allows for a standardized evaluation of models similar to existing benchmarks but still benefits from out-of-the-box support for multiple datasets and easy adaptability if need be. While we did our best to ensure extensibility, *YAIB* cannot currently support all possible use cases. Specialized use cases like federated learning or reinforcement learning currently require custom code. However, we keep

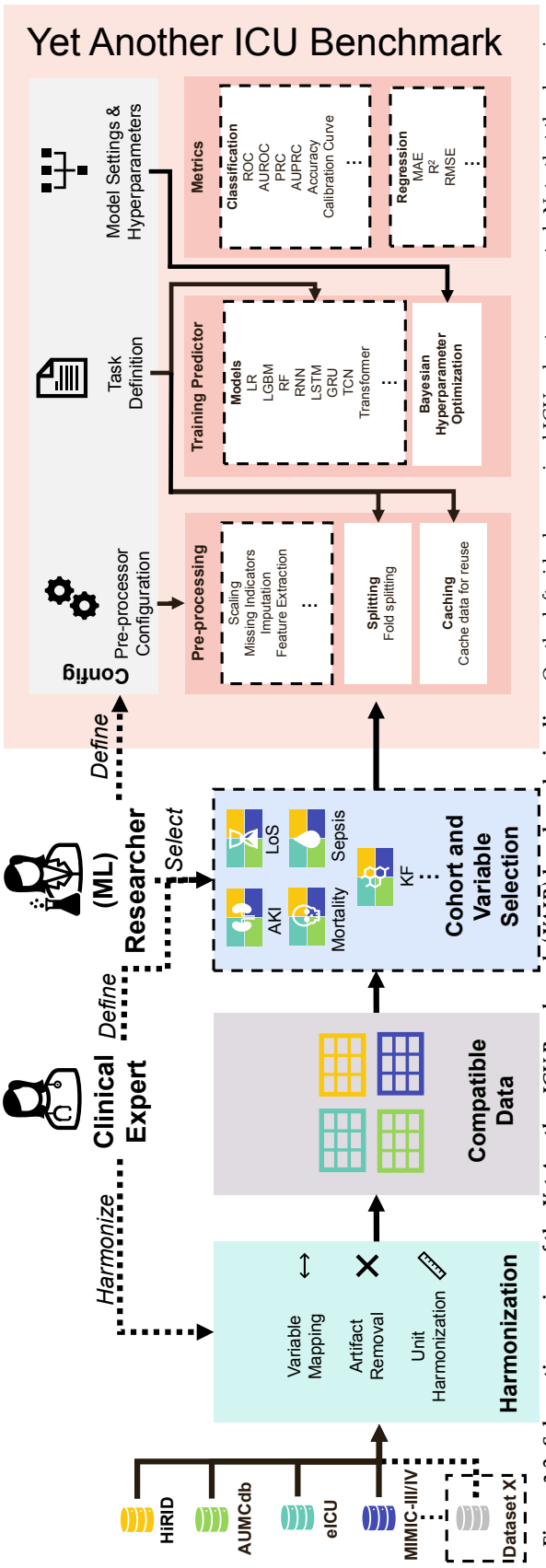


Figure 3.2: Schematic overview of the *Yet Another ICU Benchmark* (*YAIB*) benchmark pipeline. On the left side, harmonized ICU cohorts are created. Note that the domain expertise of clinicians is often necessary for defining clinically useful tasks. The schematic overview of the benchmark stages can be found on the right. Note that the dotted line indicates that this component can be easily extended, as it follows an abstracted interface.

adding functionality to *YAIB*, and users may nevertheless benefit from using parts of our framework. We strongly request users of YAIB to provide their code and a detailed list of the changes they have made to the repository to accurately and transparently provide results for their experiments.

3.3.2 Extensibility and Reproducibility

We designed *YAIB* to be as extensible as possible while retaining full reproducibility. This means easy support of new databases, clinical concepts, tasks, experiment configurations, preprocessing pipelines, imputation methods, models, and evaluation metrics. If changes are necessary, they need to be reproducible and easily shareable across research teams. If the user only requires a few default ICU tasks from a single i.i.d. dataset to test their new method, any existing ICU benchmarks could be sufficient. Users do not need to apply for access to multiple datasets and do not have to deal with the intricacies of the clinical task definition. As long as the integration of a new model is seamless, such simple frameworks are fit-for-purpose and abstract much of the complexity, allowing the user to only worry about one thing: their model. If multiple papers used the exact same benchmark, results are also directly comparable between papers (an "apples-to-apples" comparison).

However, we found that this setup tends to be too restrictive and thus unrealistic. Users often want to highlight a particular aspect of their model, prompting them to adapt to the default task. At other times, they want to show clinical impact and need to adapt the default task to make it more realistic. Given the lack of successful translation of prediction models into clinical practice, reviewers are also increasingly requesting external validation – sometimes with multiple endpoints – which is difficult to shoehorn into most existing solutions. *YAIB* embraces the need to tweak experimental setups. Results will no longer be directly comparable between papers, but we argue that true apples-to-apples comparisons were inherently rare. Instead of forcing users into a rigid framework, it allows for adaptations but requires them to be done in a transparent manner. Absolute performance should be compared only within the same paper or among papers with the same task setup (see our examples in Tables 5 and 6).

To facilitate the transparency of adaptations, we rely on a sophisticated framework to define clinical concepts across multiple datasets (ricu). We have adapted and extended ricu to provide a standard workflow for *YAIB* to integrate new databases and define new clinical concepts. To date, it has been successfully used to bring 4/5 +1 ICU datasets into a common format (including our addition of the Salzburg Intensive Care Database, which is currently in quality control). This approach is flexible enough that we have not yet encountered significant restrictions in mapping admissions, demographics, vital signs, laboratory values, medication (including rates and durations), clinical scores, and outcomes at different time scales across datasets. The main restriction of ricu is that it is currently implemented in the R language only, but we provide guidance on how to access it via rpy2, and we are in the process of porting it to Python; this will make our pipeline even more accessible, especially to clinical researchers. Our cohort definition functionality provides helper functions to apply inclusion/exclusion criteria on top of ricu and report step-by-step attrition numbers. The cohorts can be

used in a modular fashion with custom preprocessing steps, imputations, prediction models, and evaluation metrics, all using the exact same code across multiple datasets.

Even so, there will likely be situations where the user may be better off with a custom solution. We expect this to occur once their use case diverges significantly from standard supervised learning. For example, federated learning or reinforcement learning setups may require significantly different training and evaluation loops. These are not currently supported, but we consider this as future work. In any case, the user can still use our data processing, cohort generation, and possibly other parts of *YAIB* (e.g., by exchanging the default training module with a custom module). Authors using *YAIB* should, therefore, provide their code and a detailed list of the changes they have made to the repository; modern version control allows us to verify this against the original *YAIB* repository easily.

The *YAIB* pipeline has helped us to produce reproducible results quickly and provides the required extensibility for our purposes. We are in touch with some researchers who have used *YAIB* to date and provided feedback, although mainly in an informal way. For concrete examples and guidance for extending *YAIB* , we refer to Appendix D and the wiki documentation.

3.3.3 The Choice of Features

We chose the 52 most common clinical features shared by all datasets. They were readily available in all benchmarked datasets, demonstrating *YAIB* 's adaptability. This is done because our work focuses on the interoperability of datasets and the opportunity for experiments with a.o. transfer learning and domain adaptation. We believe there is the most value in providing a modular setup where the user can add or remove features to suit their needs better and, most importantly, do so reproducibly.

Nevertheless, several medications for eICU and MIMIC-IV are readily available; the ricu package maintains a complete list of the currently available native concepts, which are available[3]. Complex concepts, dependent on several native concepts, such as SOFA scores, are additionally available. Each concept that is available in ricu can be readily used in *YAIB*. Some medications, such as antibiotics and vasopressors, have already been implemented to define the complex Sepsis endpoint. Therefore, we decided to leave those out to have the same features for each task.

We note, additionally, that it is straightforward to implement new concepts in our pipeline; Appendix E.2 describes the addition of Potassium Chloride to the ICU harmonization package ricu. A similar process can be followed for adding new medications, which immediately improves the usability of *YAIB*. Moreover, we are actively working on integrating more features, including comorbidities and medications. We want to note that many features are unavailable across all datasets; this does not mean they can not be valuable in clinical prediction tasks.

Finally, we would like to point out that *YAIB* 's end-to-end pipeline is designed as a solid starting point for 1) clinicians looking for external validation to employ ML in practice, 2) dataset creators looking for a solid platform to facilitate widespread use, and 3) the ML community to contribute novel prediction models. They can use a mature

[3] https://github.com/eth-mds/ricu/blob/main/inst/extdata/config/concept-dict.json

and externally developed framework, which adds to the credibility of any experimental results. Adding new feature concepts for their datasets can also increase their adoption. They are likely domain experts for their respective datasets, meaning fewer errors are made in this process. This process will improve the usability of *YAIB* as an end-to-end benchmarking tool and improve the confidence of health experts in clinical ML.

3.3.4 Using YAIB in Novel Scientific Work

We acknowledge the importance of reproducible ML experiments. This section describes how future work can transparently use *YAIB* as a platform for comparing their contributions. The authors ideally provide one or more open GitHub repositories so it is straightforward to check versioning; this includes:

1. The concept dictionary in JavaScript Object Notation (JSON) format if they add new concepts. The main repository contains the current version of the concept dictionary of the vanilla `ricu`[4].
2. The repository that is used to generate cohorts if they introduce a new task. Ideally, this is forked from the *YAIB*-cohorts repository.
3. The `preprocessing.py` file in case this has been changed.
4. The `model.py` and `dataset.py` file that contains the definition for the model, dataset, and dataloader (if adjusted).
5. `model.gin` file that specifies the used hyperparameters and hyperparameter ranges.
6. `wandb.yml` if Weights and Biases is used for running experiments with this model.
7. Provide versions of `ricu`, `YAIB-cohorts`, and `YAIB` they have used as a base.

If authors cover these aspects when presenting new work; one can easily reproduce their experiments even though they might not have used a "vanilla" implementation of YAIB. An additional benefit of providing these materials is that authors of future work can hereby participate in making YAIB more comprehensive.

3.3.5 Clinical Concepts

We ensured that our benchmark supports existing and future ICU datasets. Working with multiple datasets requires careful data harmonization, as datasets are collected in different locations, with different clinical recording, and may have completely different data structures. We use the `ricu` [Ben+23] R package to bring datasets into a common, semantically interoperable format. This harmonization relies on two things: **1)** a common temporal reference point and **2)** a dataset-independent definition of clinical concepts. `ricu` by default distinguishes measurements recorded for a patient, a hospital admission, or an ICU admission, and supports conversion between these levels of measurement. The definition of reference points facilitates temporal comparability between

[4] `https://github.com/eth-mds/ricu/blob/main/inst/extdata/config/concept-dict.json`

datasets. `ricu` also allows defining clinical concepts such as heart rate or SOFA score independently of any particular dataset, specifying their meaning, plausible min/max ranges, and units of measurement. A concept can be enabled for a dataset by specifying how it should be extracted from the data, for example, by selecting an entire column or subsetting a table based on an item identifier. `ricu` thus acts as an interface to the raw data (stored in a fast, compressed column format), on command returning the data for a concept in a table of ID-time-value pairs. This is still no panacea to make ICU datasets immediately interoperable, but it provides a helpful framework for harmonization. For users unfamiliar with R, we provide an interface to directly access `ricu` concepts from Python. PyICU, a native Python implementation of `ricu`, is in development.

3.3.6 Patient Cohort and Task Definition

Once in a common format, the same task definition can be applied across datasets. This facilitates code reuse and eliminates opportunities for error. Even so, care must be taken to combine clinical concepts, define meaningful prediction targets, and apply appropriate exclusion criteria. We provide default workflows and helper functions to support this process, including a transparent pipeline for applying exclusion criteria and reporting patient attrition. We supplied this functionality in a standalone repositoryto facilitate its use with other modeling frameworks such as Clairvoyance [Jar+21]. The specification of our adaptive and re-definable pipeline is found in section C.3.

3.3.7 Training and Evaluation

A single *YAIB* experiment creates and optimizes a model for a given task and preprocessing pipeline. Experiments are defined using the `gin-config` library [DGS18] in simple Python-like text files. The model configuration defines the architecture and contains information on hyperparameters and optimizers. Every aspect of a model is fully configurable. The task configuration defines the target, the data source, the features, and the preprocessing. Additionally, one can define the cross-validation splits and the number of iterations. By defining the model and task separately, they can be mixed and matched, training the same architecture for multiple tasks or training multiple models for a single task. We provide details for adding new datasets, preprocessing, models, and an example of sepsis prediction in section C.4. Training is supervised by PyTorch Lightning [Ft23], which uses standardized training and logging, GPU parallelism, and advanced debugging. Users can configure hyperparameter ranges and sampling methods for model optimization. A Gaussian Process is fit to the hyperparameters using `scikit-optimize` [Hea+21] as a robust alternative to random search [SLA12]. In later version of the benchmark, we use the Optuna framework [Aki+19] that is in active development, provides more algorithm options, supports checkpointing, and generates visualizations.

3.3.8 Result Tracking

Results are automatically aggregated and written to a JSON file, in addition to optional Tensorboard [Aba+16], PyTorch Lighting [Ft23], and WandB [Bie20] logging for

easy experiment tracking. Performance evaluation records widely-used metrics out of the box (AUROC, AUPRC, calibration curve, accuracy, loss) and supports multiple evaluation libraries: TorchMetrics [Nic+22], Pytorch-Ignite [Fom+20], and Scikit-Learn [Ped+11a] metrics. New metrics, either developed by the user or from existing libraries, can be easily added (see Appendix C.5.6). Results are automatically summarized and accessible as a JSON file.

3.3.9 Reproducibility

For the purposes of reproducibility we would like to refer to the source code of `YAIB`[5] (main benchmark), `YAIB-cohorts`[6] (adaptable cohort extraction) and ReciPies[7] (extensible preprocessing package). Models for each task and architecture are publicly available[8]. In the included source code, a file called `PAPER.md`[9] describes the reproducibility steps of the experiments in this paper. Specifically, one requires the standalone codebase of `YAIB-cohorts` to first create the cohorts from the acquired data, once you have completed the required credentialing (see section C.2 for details). As mentioned, we include demo cohort data for each task (results for these cohorts are shown in section C.1). section C.3 describes the data processing and task creation. The usage of *YAIB* is detailed in section C.4. section C.5 shows how *YAIB* can be extended with new datasets, clinical concepts, tasks, models, and evaluation metrics. Additionally, we refer to the `README.md`[10] and the wiki[11] for the usage of *YAIB*. section C.6 and C.7 detail the experiment design and chosen hyperparameters, respectively. Finally, section C.8 contains the machine learning reproducibility checklist for our work.

3.3.10 New Features

Since publication, we have added several new features that allow more general use of YAIB. Specifically, we use YAIB as our experimental framework of choice in chapter 4.

- To improve performance and usability, the preprocessing backbone has been replaced with the Polars [Vin+24] library. ReciPies now also handles both Pandas and Polars. Preprocessing pipelines are now significantly faster.
- In the process, we have added modality selection capabilities that can, at runtime, change the data modalities that are used to train a model. One can more reliably compare the impact of different modalities on prediction performance.
- Moreover, outcome prediction horizons and tasks are now easily switched out by providing different outcome files for the same dataset and using a flag to select the corresponding file for the current experiment. Static datasets are also supported in the new framework.

[5] https://github.com/rvandewater/YAIB
[6] https://github.com/rvandewater/YAIB-cohorts
[7] https://github.com/rvandewater/ReciPys
[8] https://github.com/rvandewater/YAIB-models
[9] https://github.com/rvandewater/YAIB/blob/master/PAPER.md
[10] https://github.com/rvandewater/YAIB/blob/master/README.md
[11] https://github.com/rvandewater/YAIB/wiki/YAIB-wiki-home

- SHapley Additive exPlanations (SHAP) [LL17a] explainer value aggregation has been added to investigate feature importance across multiple datasets. This allows for a more robust interpretation of the model's predictions and feature importance.
- Also, we have added the option to output the raw prediction values. The output can be used in meta-models, such as alarm systems. It also allows us to, for example, create Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for which we provide utilities.
- More model architectures have been added, including Imbalanced-Learn classifiers [LNA17], CatBoost [Pro+19], and XGBoost [CG16]. These models are now available in the *YAIB* framework; this enables users to experiment with different architectures easily.

3.4 Preprocessing and Feature Engineering with ReciPies

Further preprocessing is often required at runtime, including data normalization, generation of missingness indicators, and imputation. We provide a transparent, flexible way for users to define their preprocessing pipeline. ReciPies [VSR26] is available as a standalone package and has been further developed to be a general-purpose preprocessing library. ReciPies, like *YAIB*, is open source and extensive documentation and the code are available online[12].

3.4.1 Motivation

ML workflows live or die by their data-preprocessing steps, yet in Python, these steps are often scattered across ad-hoc scripts or opaque Scikit-Learn [Ped+11b] (sklearn) snippets that are hard to read, audit, or reuse. ReciPies provides a concise, human-readable, and fully reproducible method to declare, execute, and share preprocessing pipelines, that adheres to Configuration as Code principles. It lets users describe transformations as a recipe made of ordered *steps* (e.g., imputing, encoding, normalizing) applied to variables identified by semantic roles (predictor, outcome, ID, time stamp, etc.). Recipes can be *prepped* once, *baked* many times, and separated between training and new data. ReciPies provides the choice of Pandas and Polars backends and is easily extensible. Data provenance can be tracked and published. Packaging preprocessing as clear, declarative objects, ReciPies lowers the cognitive load of feature engineering, improves reproducibility, and makes methodological choices explicit, benefiting individual researchers, engineering teams, and peer reviewers alike.

Robust machine-learning results in science hinge on transparent, reproducible data-preprocessing—yet in Python, these steps are typically spread across ad-hoc notebooks or are buried inside opaque scripts; that is, if this code is even made available. Additionally, most variable semantics are unclear. These problems confound research results, complicate peer review and hinder reuse. Researchers and engineers working with longitudinal regulated data (e.g., energy production, finance, and environmental

[12] ReciPies website:`https://rvandewater.github.io/ReciPies/`

monitoring) especially need pipelines they are able to audit, serialize, and hand to collaborators without reverse-engineering a tangle of imperative code. The current lack of reproducibility has been documented extensively in the literature [McD+21; JPM17; GK18; Kel+19b; Raf19; Sem+25]; moreover, scientific venues have begun to address this issue [Var24].

ReciPies fills this gap by bringing a tidy, stepwise *recipe* interface to Python. Users declare transformations over variables selected by semantic roles; recipes are *prepped* once on training data and *baked* on new data to eliminate leakage; and every step is inspectable, versionable, and serializable (JSON/YAML). `Recipes` runs on Pandas [McK10] and Polars [Vin+24] for interoperability and performance, and their Object-oriented abstractions enable users to implement custom steps. The framework is declarative and reproducible for data preprocessing, prioritizing human readability and methodological transparency. We demonstrate that there is no need to sacrifice readability for performance or flexibility for simplicity. By reducing the cognitive overhead of feature engineering and making methodological choices explicit, ReciPies enables researchers to focus on their core work. We hope this will broaden the reproducibility discussion in ML from hyperparameters to the entire experiment pipeline. We encourage the development of domain-specific step libraries and integration patterns that can benefit the broader ecosystem.

Our work brings the Recipes [Kuh+24] framework to Python and extends it for the ML community. The design enables straightforward integration as part of a pipeline that includes ML libraries like sklearn [Ped+11b] and PyTorch [Pas+19]. To the best of our knowledge, no other packages comply with the flexibility and reproducibility of ReciPies and its Configuration as Code approach. Sklearn offers composable transformers, but no role-based variable grammar, limited human-readability, and awkward serialization. Feature-engine [Gal21], pyjanitor [J+19], or scikit-lego [War+25] add helpful transformers or cleaning verbs. However, none provide a unified, declarative recipe abstraction with a strict "prep/bake" split and backend flexibility. ReciPies provides a stepwise recipe that is easy to use and read, allowing users to readily preprocess data for a wide range of machine learning pipelines.

3.4.2 Usage

ReciPies is used as the bedrock of reproducible pipelines of Yet Another ICU Benchmark [Van+24a]. The adaptable, configurable code modules that make extensive use of ReciPies can be found at `https://github.com/rvandewater/YAIB/blob/development/icu_benchmarks/data/preprocessor.py`; this demonstrates that ReciPies can be used for arbitrary research domains.

3.4.3 Future Steps

We plan to expand the library of Polars-native steps to fully leverage its columnar execution model, particularly for time-series operations and large-scale aggregations, where Polars shows significant performance advantages. We envision ReciPies recipes as portable preprocessing artifacts that can be versioned, tracked, and deployed across

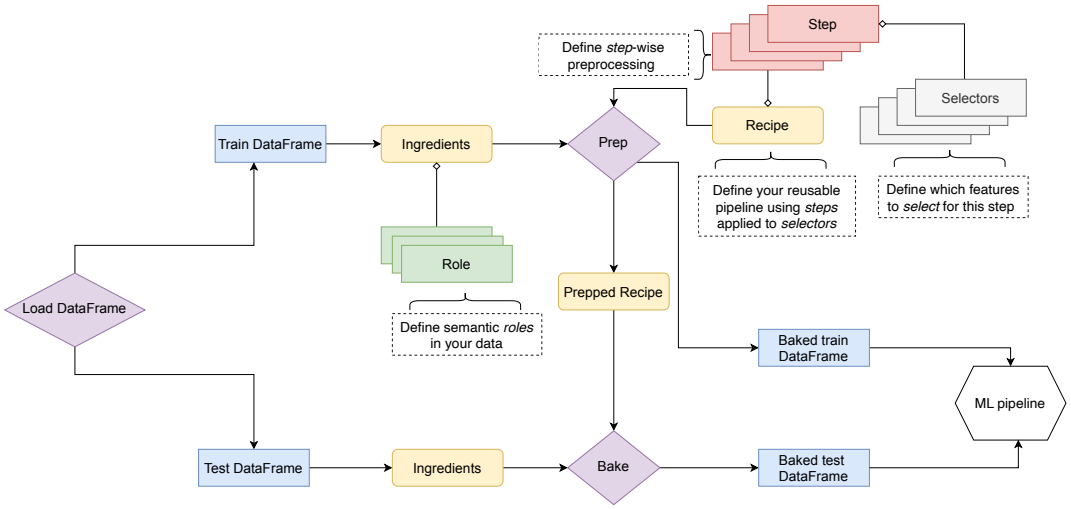


Figure 3.3: A schematic overview of ReciPies' architecture. We 1) load a Pandas or Polars (training) dataframe, then 2) wrap it in an Ingredients object that maintains column role information (i.e., what does this column do in this dataset). Next, we 3) define a Recipe consisting of multiple Steps that operate on selected columns. Finally, we 4) prep the Recipe on the training data and 5) bake it on new data. We can then 6) run our ML pipeline on train and test data.

```
import polars as pl
from sklearn.model_selection import train_test_split
from sklearn.impute import MissingIndicator
from recipies import Ingredients, Recipe
from recipies.selector import has_role
from recipies.step import StepImputeFill, StepScale, StepSklearn

# Load and split Physionet Computing in Cardiology Challenge 2019 dataset
df = pl.read_csv("Physionet_CiCC_2019.csv", sep="|")
df_train, df_test = train_test_split(df, test_size=0.2, random_state=42)

# Define ingredients (training set + roles)
roles = {"outcomes": ["SepsisLabel"], "predictors": ["Age", "HR", "Temp"], "groups": ["PatientID"],
"sequences": ["ICULOS"]}
ing = Ingredients(df_train, roles=roles)

# Define the recipe for processing the ingredients (e.g., scaling and imputation)
rec = Recipe(ing)
rec.add_step(StepScale())
rec.add_step(StepSklearn(MissingIndicator(features="all"), sel=has_role("predictor")))
rec.add_step(StepImputeFill(strategy="forward"), sel=has_role("predictor"))

# Prepare (=fit) recipe on training data and apply the same recipe to the test set
df_train = rec.prep()
df_test = rec.bake(df_test)
```

Figure 3.4: A code snippet demonstrating ReciPies' usage. We 1) load a Polars dataframe, then 2) wrap it in an Ingredients object that maintains column role information (i.e., what does this column do in this dataset). Next, we 3) define a Recipe consisting of multiple Steps that operate on selected columns. Finally, we 4) prep the Recipe on the training data and 5) bake it on new data.

different environments. Tighter integration with experiment tracking and model registries would streamline the transition from research to production, a complex process in many application domains.

Table 3.4: Prediction task overview. Note that the related work is non-exhaustive.

No	Task	Frequency	Type	Related work
1	Mortality	Once per stay*	C	[BXA20; Lu+22; Med+19; Sha+17; Sye+21b]
2	AKI	Hourly	C	[Hua+21; Nik+22; Pan+19; Ran+20; SZC21; Wan+20a; Koy+18b]
3	Sepsis	Hourly	C	[Kok+20; Lau+20; Mer+20a; Fle+20b; Moo+21a; Moo+19; Mur+21; Rey+19; SZC21; Wan+22]
4	KF	Once per stay*	R	[Tom+19; Fut+16; Per+15; Che+18b]
5	LoS	Hourly	R	[Shi+19a; GLC20]

C: Classification, R: Regression, * Using data from 0-24 hours.

3.5 Experiments

We ran experiments for five common prediction tasks: ICU mortality, onset of Acute Kidney Injury (AKI), onset of sepsis, Kidney Function (KF) on day 2, and remaining Length of Stay (LoS) (Table 3.4). Mortality and KF used data from 0-24 hours. All other tasks used all available data until the event or discharge. We ensured adequate data quality by excluding: **1)** patients younger than 18 years; **2)** stays with missing discharge times; **3)** stays with less than six hours in the ICU; **4)** stays with measurements in less than four time bins; and **5)** stays with no measurement for more than 12 consecutive hours in the ICU. We also applied task-specific exclusion criteria. For example, we excluded stays of less than 30 hours for the ICU mortality task, as this could introduce causal leakage from patients already dead or about to die at the time of prediction. For each task, we included 52 features, of which 4 were static and 48 were time series. Various additional features, including prescriptions and diagnoses, can be directly used in *YAIB* by adjusting the cohort generation module (YAIB-cohorts); if features are not available, their implementation is straightforward (section C.5). Information on the datasets, features, and individual cohort definitions can be found in section C.2 and C.3. The code to define these cohorts is publicly available. In addition to the baseline performance for each task, dataset, and model, we used *YAIB* to investigate the effects of small variations in task definitions on predictive performance — a common obstacle to model comparability [Moo+21b; Fle+20b]. Specifically, we **i)** only excluded stays of less than 24 hours to assess the effects of causal leakage by aligning our mortality task with [Yèc+22], **ii)** omitted static and dynamic historical features (i.e., `min`, `max`, `count`, `mean`) to simulate access to fewer input data, and **iii)** compared alternative definitions for sepsis . We, additionally, evaluated transfer learning with the harmonized datasets (**iv**).

3.5.1 Preprocessing

1. Scaling: The data was scaled to zero mean and unit variance.

2. Imputation: After adding missing indicators, we forward-filled all columns for the dynamic data, replacing missing values with the last known values of the same stay. Missing values without a prior measurement were filled with the sample mean. To

prevent data leakage, we used the mean of the train split as the sample mean for all splits.

3. Feature generation: We generated the `min`, `max`, `mean`, and `count` of measurements for each feature in the dynamic data. We only applied this step for the conventional ML models, e.g., Light Gradient Boosting Machine (LGBM), as they cannot capture sequential information natively.

3.5.2 Models and Experimental Setup

We considered a range of algorithms used in previous benchmarks (see Table 3.1) and applied work [Hyl20; Pir+15; Sil+12; Sye+21b], including regularized logistic regression (LR) and elastic net (EN) (used for classification and regression, respectively [Ped+11a]), LGBM [Ke+17], and four variations of neural networks: Gated Recurrent Unit (GRU) [Cho+14b], Long Short-Term Memory (LSTM) [HS97], Temporal Convolutional Network (TCN) [BKK18] and Transformer (TF) [Vas+17]. LR, EN, and LGBM were used with the feature generation described above, as they are unable to utilize time series. The implementation of neural networks was adapted from [Yèc+22].

For our experiments, unless stated otherwise, we used 5 iterations of 5-fold cross-validation. Hyperparameters were tuned on the training set using 30/50 (DL/ML, respectively) iterations of Bayesian hyperparameter optimization [SLA12]. For computational reasons, hyperparameter tuning used only the first 2/3 folds, respectively (see section C.7 for a definition of all searched and selected hyperparameters). The final validation of the best hyperparameters used all 5 folds. Each model was optimized for a maximum of 1000 epochs. Training was stopped early if performance on the validation set did not improve for 10 epochs. The epoch with the best performance on the validation set was retained and evaluated on the test set. This process was repeated for 5 iterations, after which the results were averaged, and the standard deviation was calculated.

3.5.3 Benchmarking Baseline Models on Major ICU Datasets

Baseline results for all tasks can be found in Table 3.5 and 3.6. Note that we have also benchmarked our tasks for two openly available demo datasets from MIMIC-III and eICU; these can be directly accessed without completing a credentialing procedure (see Table C.4 and C.5).

ICU Mortality The performance of traditional ML and DL models was highly comparable among each other and across datasets when predicting mortality based on data from the first 24 hours. Notably, AUPRC was higher in AUMCdb due to a higher outcome prevalence (Table 3.2). Surprisingly, Transformers performed poorly on MIMIC-IV for this task, which could be due to poorly chosen hyperparameter ranges.

Acute Kidney Injury (AKI) Maximum achievable performance was also similar across datasets when predicting the hourly onset of AKI, with the notable exception of HiRID,

Table 3.5: Baseline performance on the classification tasks. We **embolden** the best mean AUROC × 100 (↑, i.e., higher is better) and AUPRC × 100 (↑) per dataset and those within a standard deviation (±).

	AUMCdb		HiRID		eICU		MIMIC-IV	
Algorithm	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Mortality								
LR	83.7±0.6	52.9±1.2	84.0±0.3	36.9±1.1	84.8±0.2	33.0±0.7	86.1±0.1	39.7±0.6
LGBM	**84.5±0.5**	**53.7±1.2**	84.4±0.3	**40.6±0.8**	**85.7±0.2**	**36.0±0.6**	**87.7±0.2**	**44.2±0.7**
GRU	83.9±0.3	**53.8±0.7**	**84.8±0.2**	39.4±0.4	**86.0±0.1**	**35.6±0.1**	**87.6±0.1**	42.8±0.3
LSTM	83.7±0.7	**53.6±1.4**	84.0±0.7	37.8±1.0	85.5±0.2	**35.7±0.8**	86.7±0.4	41.0±0.7
TCN	**84.0±0.6**	**54.2±1.4**	**84.6±0.7**	39.2±1.3	85.4±0.2	34.3±0.6	87.1±0.3	41.4±0.8
TF	84.1±0.2	**54.4±1.1**	**84.9±0.7**	39.3±1.5	**85.9±0.2**	34.7±0.8	86.9±0.3	42.2±0.3
AKI								
LR	85.5±0.3	45.1±0.4	79.6±0.1	31.8±0.8	72.8±0.1	32.2±0.2	77.1±0.2	37.7±0.3
LGBM	85.8±0.3	48.4±0.6	80.2±0.2	32.8±0.4	84.6±0.1	50.8±0.2	83.8±0.1	53.3±0.2
GRU	**90.6±0.3**	**52.8±0.7**	**82.2±0.2**	**33.9±0.4**	**90.9±0.0**	**72.2±0.1**	**90.7±0.1**	**69.6±0.2**
LSTM	86.5±0.4	40.6±0.6	81.0±0.4	31.8±0.4	90.2±0.1	69.9±0.2	89.7±0.1	66.5±0.2
TCN	89.6±0.2	50.0±0.9	81.2±0.2	32.3±0.4	90.4±0.0	70.4±0.2	89.8±0.1	66.8±0.2
TF	88.2±0.2	48.2±0.7	81.5±0.2	33.4±0.5	89.9±0.1	68.0±0.3	89.6±0.1	65.6±0.2
Sepsis								
LR	74.7±1.0	4.0±0.4	76.5±0.6	8.4±0.3	71.8±0.3	2.9±0.1	77.1±0.4	4.6±0.1
LGBM	74.0±0.8	5.2±0.7	76.1±0.4	10.4±0.5	69.1±0.3	3.3±0.1	77.5±0.3	5.9±0.2
GRU	79.7±0.9	7.7±0.7	**80.6±0.5**	**12.6±0.5**	**77.4±0.2**	**5.1±0.1**	**83.6±0.3**	**9.1±0.3**
LSTM	77.1±0.8	6.4±0.5	78.8±0.4	11.1±0.5	74.0±0.2	4.0±0.1	82.0±0.3	8.0±0.2
TCN	78.7±0.7	7.1±0.6	**80.8±0.5**	**13.0±0.4**	76.7±0.1	4.9±0.1	82.7±0.3	**8.8±0.2**
TF	**80.7±0.9**	**8.6±0.8**	**80.8±0.3**	**12.6±0.6**	76.2±0.1	4.6±0.1	80.0±0.8	6.6±0.2

Table 3.6: Baseline performance on the regression tasks. Results are reported in Mean Absolute Error (↓)

	Kidney Function *in mg/dL*				**Length of Stay** *in hours*			
Algo.	**AUMCdb**	**HiRID**	**eICU**	**MIMIC-IV**	**AUMCdb**	**HiRID**	**eICU**	**MIMIC-IV**
EN	**0.24±0.00**	0.28±0.00	0.31±0.00	0.25±0.00	54.9±0.0	47.2±0.1	43.6±0.0	46.5±0.0
LGBM	0.32±0.00	0.34±0.00	**0.29±0.00**	**0.24±0.00**	44.7±0.0	**39.2±0.1**	39.3±0.0	40.1±0.0
GRU	0.29±0.00	0.32±0.01	0.34±0.01	0.30±0.01	42.9±0.1	39.6±0.1	38.9±0.1	39.9±0.1
LSTM	0.29±0.00	0.33±0.00	**0.28±0.01**	0.28±0.01	44.8±0.1	39.8±0.1	39.2±0.1	40.6±0.1
TCN	0.28±0.01	**0.23±0.01**	0.31±0.00	0.28±0.01	43.7±0.1	39.9±0.1	38.9±0.0	40.4±0.1
TF	0.26±0.00	0.31±0.01	0.33±0.01	0.32±0.01	**41.8±0.1**	**39.1±0.1**	**38.2±0.1**	**39.0±0.1**

We provide the average and [Interquartile Range (IQR)]. For **KF**: AUMCdb: 0.97 [0.70, 1.61], HiRID: 0.92 [0.67, 1.50], eICU: 1.00 [0.71, 1.68], MIMIC-IV: 1.00 [0.70, 1.60]. For **LoS**: AUMCdb: 24 [19, 77], HiRID: 24 [19, 50], eICU: 42 [23, 76], MIMIC-IV: 48 [26, 89]. See Table 3.3 for details.

which had both lower AUROC and AUPRC for all models. GRU models consistently achieved the best performance.

Sepsis The performance of baseline models was worse for the hourly onset of sepsis, both for AUROC and especially AUPRC. This may be explained by the particularly low prevalence of ∼ 1% hourly bins classified as septic and the relative difficulty of predicting sepsis in general [Moo+21b].

Kidney Function (KF) Classical ML models achieved relatively good performance for this task, which may reflect the dependence of KF on a limited number of features [GOV22b].

Length of Stay (LoS) The performance of ML and DL models was also comparable across datasets. Nevertheless, predicting the length of stay seems difficult, given that the average MAE is almost two days. Transformers consistently outperformed most other model types.

Table 3.7: ICU mortality prediction on HiRID with ($> 24h$) and without ($> 30h$) possibility of causal leakage. We compare the standard definition, used for our baseline results, with the extended version (which includes patients that have spent less than 30 hours in the ICU). For comparison, we also included the similarly defined HiRID-Benchmark [Yèc+22] cohort. Note that this work reports higher performance on the same task; they made use of a more inclusive cohort, a higher sampling resolution, and included administered drugs as features. This could, besides including more patients in the cohort, positively impact the performance.

	Cohort definition					
	w/o leakage		**w/ leakage**		**[Yèc+22]**	
Algorithm	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
LR	84.0±0.3	36.9±1.1	87.2±0.4	43.1±1.3	89.0±0.0	58.1±0.0
LGBM	**84.5±0.3**	40.6±0.9	**87.9±0.5**	**47.7±1.2**	88.8±0.2	54.6±0.8
GRU	**84.8±0.2**	39.4±0.4	**88.2±0.3**	46.1±1.2	90.0±0.4	**60.3±1.6**
TCN	**84.6±0.7**	39.2±1.3	87.8±0.2	45.2±1.0	89.7±0.4	**60.2±1.1**
TF	**84.9±0.7**	39.4±1.5	**88.2±0.3**	**47.1±1.2**	**90.8±0.2**	**61.0±0.8**

3.6 Using YAIB as an Experimental ML Framework

Now, we demonstrate how YAIB can be used to validate hypotheses quickly in EHR prediction task modelling.

3.6.1 Changing Exclusion Criteria for Mortality Cohorts

As hypothesized, the choice of exclusion criteria could majorly impact achievable prediction performance (Table 3.7). Compared to the peak performance achieved with the HiRID-benchmark [Yèc+22], our baseline performance for the mortality task was noticeably lower. Further investigation revealed that while the outcome definition is identical, there were two important differences: Yèche et al. [Yèc+22] 1) did not exclude patients that spent less than 30 hours in the ICU – we, instead, decided to exclude these patients as this prediction is usually not clinically insightful – and 2) used additional predictors, including medication information. Aligning the exclusion criteria accounted for half of the performance difference. The remaining difference was likely due to the inclusion of additional predictors — most notably drug usage — in the HiRID-benchmark. This highlights the difficulties of comparing works that ostensibly address the same task, even using the same dataset and model implementation.

Table 3.8: AUROC performance comparison of including static data for ICU mortality prediction.

	Dataset							
	AUMCdb		HiRID		eICU		MIMIC-IV	
Inclusion	w/ static	w/o static	w/ static	w/o static	w/ static	w/o static	w/ static	w/o static
LR	83.7±0.6	82.9±0.6	84.0±0.3	82.8±0.7	84.8±0.2	84.3±0.2	86.1±0.1	84.3±0.3
LGBM	**84.5±0.6**	83.4±0.4	**84.4±0.3**	83.9±0.7	**85.7±0.2**	84.7±0.2	84.7±0.2	86.2±0.2
GRU	83.7±0.7	83.4±0.6	**84.3±0.7**	84.0±0.8	**85.9±0.2**	85.7±0.2	**87.4±0.2**	**86.9±0.3**
LSTM	83.7±0.7	82.9±0.6	84.0±0.7	83.4±0.7	85.5±0.2	85.1±0.2	86.7±0.4	86.1±0.3
TCN	**84.0±0.6**	83.5±0.6	**84.6±0.7**	83.9±0.8	85.4±0.2	**86.4±0.3**	87.1±0.3	86.4±0.3
Transformer	**84.1±0.2**	83.7±0.4	**84.9±0.7**	84.4±0.7	**85.9±0.2**	85.7±0.2	86.9±0.3	86.5±0.3

3.6.2 The Impact of Data Features

The RECIPIES package provides the functionality of assembling different preprocessing steps to be supplied by the user (i.e., a recipe). We leveraged this customizable preprocessing to perform training prediction models without static features (i.e., age, sex, height, and weight). The AUROC results can be found in Table 3.8 (the AUPRC is recorded in Table C.3). From these results, we can see that including the static data seems to result in better performance across all models and datasets. A notable exception is the transformer model on MIMIC-IV: we notice that training without the static data performs significantly better; we, again, assume the choice of hyperparameters might have limited the performance of the original cohort.

Restricting Input Features Additionally, we reran the experiment without generating dynamic historical features (i.e., `min`, `max`, `count`, `mean`) on LGBM and logistic regression. We observed that dynamic feature generation consistently outperformed task definitions that did not include them (Table C.1 and C.2). LR on MIMIC-IV showed a considerable performance gap, whereas AUMCdb remained stable. We noted a performance decrease that ranges between 4.0% and 19.1% for LR and between 5.2% and 13.1% for LGBM. Omitting static features led to minor drops in performance (Table 3.8 and C.3); averaged across datasets, we observe performance differences ranging between 0.5% and 0.2% for the transformer model. A similar experiment for using static data can be found in Table 3.8 and C.3.

Comparing Sepsis Definitions Sepsis has been defined in several ways [Fle+20b], mainly because a clinical gold standard that can be transferred to ML models is currently lacking. Our sepsis definition (adapted from [Sey+16], see section C.3) can be considered closely related to that used by [Moo+21a], who implement a variant of Sepsis-3 [Sin+16]. However, we required that antibiotics were administered continuously for ≥ 3 days [Rey+19]. We judged that this would increase the clinical usability of the task. However, we found that it also severely reduced the achievable AUPRC — likely due to a much lower prevalence (Table 3.9). The definition used by Calvert et al. [Cal+16], on the other hand, adapted Sepsis-2 [Lev+03], which differs fundamentally from Sepsis-3 and resulted in a notably higher AUROC [Eng+20]. This highlights the

importance of precise cohort definitions, as some definitions may, by design, be more difficult to predict.

Table 3.9 shows the number of stays and the prevalence of sepsis per stay for all of the above cohort definitions; the incidence of sepsis ranges from 3.7 to 16.7 percent, which can be explained by both differences in patient characteristics as well as antibiotic prescribing behavior. In Table 3.10, we can see significant differences in AUPRC

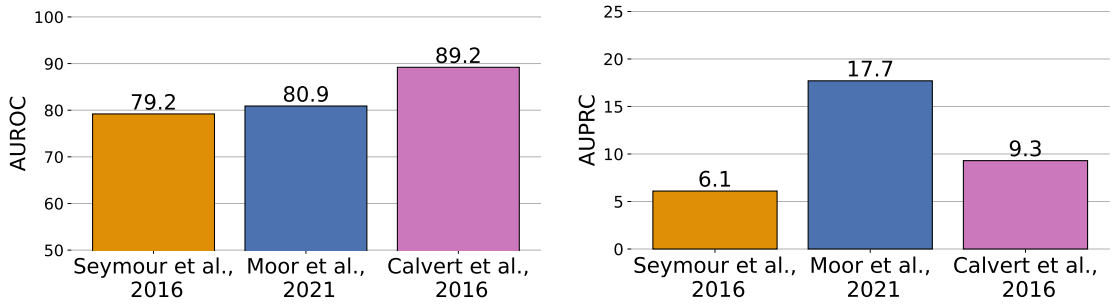


Figure 3.5: Performance comparison of GRU models trained with cohorts based on different sepsis definitions on MIMIC-IV. Left: AUROC performance for different sepsis definitions. **Right**: AUPRC performance for different sepsis definitions. We can see that the choice of sepsis definition has a considerable impact on achievable performance.

outcome, depending on the cohort definition, though the GRU consistently outperforms the LGBM. The results show that it is important to agree upon precise cohort definitions for comparability and, by extension, the progression of developing clinically applicable models. Figure 3.5 shows the results on the GRU model.

Table 3.9: Comparing MIMIC-IV sepsis cohorts according to three different definitions.

Cohort	Seymour et al. [Sey+16]*	Moor et al. [Moo+21a]	Calvert et al. [Cal+16]
Stays	67,056	53,642	65,901
Prevalence	3,730 (5.6%)	8,919 (16.7%)	2,406 (3.7%)

*Our default definition.

3.6.3 Transfer Learning

The *YAIB* framework supports transfer learning by allowing models trained on one dataset to be evaluated and fine-tuned on another. This is particularly useful in settings where data might be scarce.

Table 3.10: Sepsis prediction on MIMIC-IV for different definitions of sepsis.

	Sepsis definition					
	[Sey+16]*		[Moo+21a]		[Cal+16]	
Algorithm	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
LGBM	75.9±0.2	4.3±0.0	72.4±0.0	10.5±0.0	62.2±0.2	1.8±0.0
GRU	**79.2±0.1**	**6.1±0.0**	**80.9±0.0**	**17.7±0.0**	**89.2±0.0**	**9.3±0.2**

* Our definition; adapted to be more clinically actionable, see section C.3.

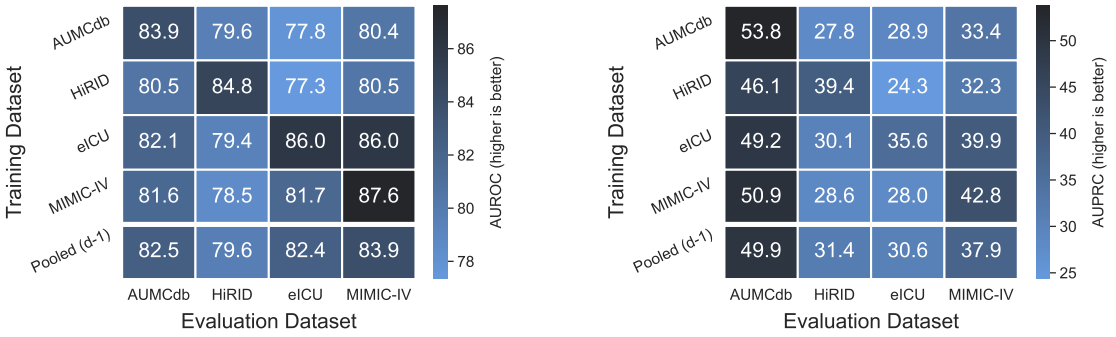


Figure 3.6: Performance of prediction models when trained on one dataset (row) and evaluated on all others (columns). Left: Performance in AUROC of the GRU model on *ICU mortality*. **Right**: Performance in AUPRC for the same models. Pooled (d-1) refers to training a model on every dataset except the evaluation dataset.

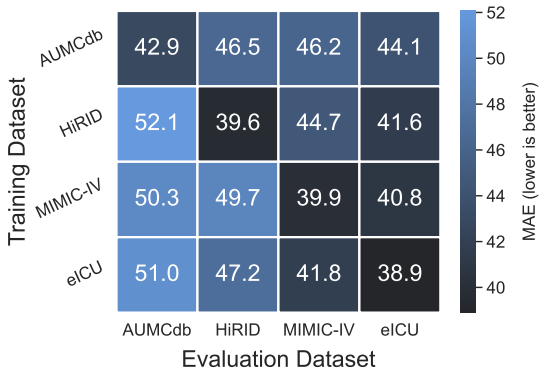


Figure 3.7: Performance of prediction models when trained on one dataset (row) and evaluated on all others (columns). Mean Absolute Error (MAE) for the Transformer model on Length of Stay prediction.

External Validation *YAIB*'s common dataset format allowed us to evaluate a model trained on an equal sample of one dataset on data from all other datasets. We additionally trained a model on pooled (d-1) data from three datasets and evaluated on the fourth, held-out dataset. For the ICU mortality task (Figure 3.6), models, as expected, performed best on independent test data from their training dataset (diagonal). Performance could drop considerably when models were evaluated in another database (off-diagonal). Notably, AUPRC performance could increase in the evaluation dataset (rows) but always remained lower than the highest achievable performance for that dataset (columns). We found that MIMIC-IV and eICU transferred well among each other. The pooled model usually performed as well as the best single-dataset model. Notably, AUMCdb AUPRC results demonstrate decidedly higher performance than evaluation on other datasets, which could be the result of a patient case mix and outcome prevalence (see Table 3.3).

Fine-tuning In Figure 3.6, we saw that eICU resulted in the most generalizable model for ICU mortality, which may serve as a strong pretraining for transfer learning. Since it worked worst for HiRID, we further fine-tuned the eICU GRU model (source) for HiRID (target) by retraining it using an increasing number of samples from the HiRID dataset. We compared the results to a model trained from scratch on the same amount of HiRID samples (Figure 3.8). Fine-tuning was profitable for any number of additional samples and especially for $< 4,000$ samples.

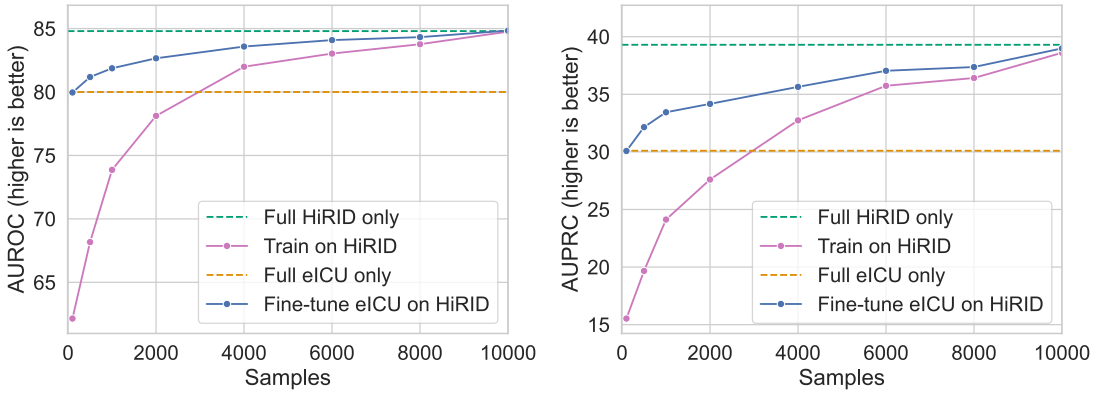


Figure 3.8: Fine-tuning an eICU model for ICU mortality prediction on HiRID. Left: AUROC and **Right**: AUPRC performance of the GRU model on HiRID when fine-tuned with an increasing number of samples from the HiRID dataset. The dashed line shows the performance of a model trained from scratch on the same amount of HiRID samples.

3.7 Discussion

We provide extensive ML and DL baselines for five clinical prediction tasks trained across four major open-source ICU datasets. While we frequently obtained comparable results across model architectures, seemingly small differences in cohort definition could substantially impact the achieved accuracy. Our findings highlight not only the need for standardized training pipelines but also for harmonized cohort definitions to allow for a meaningful comparison of clinical prediction models. Our work provides the first international, multi-center ICU benchmark, including the first-ever benchmark for the Amsterdam University Medical Center Database dataset. It naturally facilitates sorely needed external validation of model performances and allows fine-tuning of pre-trained models for new datasets. This makes *YAIB* relevant to a wide range of research areas beyond classical supervised learning, including domain adaption and generalization. We hope this broad reach encourages ICU data providers to ensure compatibility with *YAIB*, as they can expect a larger overall research impact. This simplifies the use of novel datasets by the clinical and ML community.

YAIB aids researchers in training baseline models by providing them with ready-to-use implementations of state-of-the-art model architectures; new model implementations can therefore be easily compared. While most existing benchmarking studies are hard-coded, we utilize flexible, *dataset-independent* cohort definitions and configurable preprocessing facilities linked via a common, shareable syntax. This setup acknowledges that task definitions inevitably involve arbitrary decisions, without one "size" that fits all. In our work, we embrace this idea and aim to equip researchers — both applied and theoretical — with the tools to quickly adapt a task to their individual needs (including the use of custom proprietary data) while maintaining reproducibility and reusability across studies. Models can thus be compared across multiple, slightly different task definitions and datasets, still ensuring an apples-to-apples comparison. We hope this lowers the bar for researchers to test their approaches across a range of configurations and datasets.

3.7.1 Limitations

YAIB's open-access benchmarking is currently limited to ICU settings, where several datasets are publicly available. A similar setup could be beneficial for data from other medical settings, such as inpatient wards. Although created for critical care, *YAIB* is not specific to the ICU and can be readily extended to other settings, provided a suitable configuration is defined (see chapter 4). Features included in *YAIB*, at the time of writing, mainly relate to vital signs, lab tests, and data relevant to outcome definitions. Further clinician-assisted harmonization efforts will be necessary to increase the breadth of features, most notably medications and comorbidities. If *YAIB* is adapted to general EHR, including medical imaging is a logical next step (it is adapted to handle wearable vital signs and text in chapter 4). We also note that we compared these models on the basis of commonly used ML metrics; we leave the comparison with respect to clinical fairness and bias as an easy future extension to our framework (see section C.5). Finally, we advise users of our benchmark to carefully consider the compromises made to allow for cohort harmonization; we strongly recommend clinical validation before making practical decisions based on the developed models.

3.8 Conclusion

Routine medical data is highly complex. Without clear ground truth, researchers are inevitably forced to make arbitrary design choices when defining outcomes and populations of interest. To promote comparable and reproducible models in this setting, we believe that further tools are needed that allow researchers to define clinical prediction tasks transparently, share experimental setups easily, and validate results against various data sources. As a flexible and extensible framework for clinical modeling on ICU data, *YAIB* is meant to be a step towards that goal. Since publishing this work as preprint, we have already seen community driven use and extensions of *YAIB* for research purposes [SRC24; San+25a; Roc+24; FXC25].

4

Creating a Continuous Multimodal AI with Wearable Vital Signs to Predict Postoperative Complications in the General Ward

This chapter largely mirrors the work Van de Water et al. [Van+25c]. In the prospective study conducted at the Charité Virchow Clinic Visceral Surgery department, which collected the data for this work, I served as the lead Machine Learning scientist, performed extensive data transformation and analysis, and was the primary ML author. It builds upon earlier work I authored Van de Water et al. [Van+24c] and Van de Water et al. [Van+24b]. Moreover, the alarming approach has been conceptualized by Daniela Zuluaga Lotero in their Master thesis that I supervised. A retrospective analysis where I conceptualized, developed and analyzed an ML system [Rie+25b] forms one section of this chapter; this was a retrospective work that was prepared while the prospective study was ongoing. Furthermore, more retrospective work is in preparation [Van+25b]. Lastly, I contributed to various ML-related work in visceral surgery as part of my affiliation of guest scientist and collaboration with the Charité that is not covered in this chapter [Win+24; Pfi+24; Mau+25; Win+25; Pfi+25]. This chapter has been extended with related work, clinical context, and results to provide a comprehensive overview.

4.1 Motivation

More than 300 million surgical procedures are performed worldwide each year [Dob20], linking one-third of all hospital expenditures to surgical interventions [Shr+15]. Moreover, about half of people undergo abdominal surgery over their lifetime [Nun+09]. Despite considerable advances in surgical techniques, more than 25% of patients having major abdominal surgery experience a life-threatening event during hospitalization [Bau+19]. Postoperative complications are a major determinant for surgical outcomes, including higher mortality, reduced disease-free survival, and impaired quality of life

for patients and increased costs for society [Bro+14b; Dow+23; Kon+21a; Mua+21; Sar+14; Von+11; Wan+19].

Early detection of postoperative complications is a cornerstone of failure-to-rescue strategies because it enables immediate therapeutic intervention. Nevertheless, timely recognition of complications remains challenging. Physiological patterns are often subtle, nonspecific, and time-varying, with substantial heterogeneity between patients [RKK24]. Deterioration is often detected late, after a self-reinforcing cascade of aggravation has already begun.

AI and, specifically, ML can predict complex adverse outcomes, such as cardiac arrest or sepsis, in the intensive care unit (ICU) [Hyl20; Nem+18]. However, in abdominal surgery, most complications emerge only after the patient has been transferred to the general ward (GW). Compared to the ICU, GW care is constrained by lower medical staffing and less monitoring equipment. Vital-sign monitoring is often confined to intermittent spot-checks, resulting in long intervals of insufficient observation that hinder early detection [Hym+07].

We hypothesized that incorporating telemetric monitoring into digitalized clinical data would yield accurate AI-driven digital patient representations, enabling the earlier identification of clinically relevant complications after major gastrointestinal, hepatobiliary, and pancreatic surgery compared with standard monitoring. We prospectively developed a multimodal artificial intelligence (AI) prediction model for surgical ward patients, serving as a scalable foundation for future semi-autonomous care frameworks. The system integrates routinely generated data — including patient characteristics, intraoperative parameters, ICU data, laboratory values, semantic clinical notes, and medications — with high-resolution telemetric vital signs to create a digital representation of the patient's physiological state. Aiming to detect clinically relevant yet highly granular complications, these include deep organ space surgical site infections (SSI-III) [McL+23], bile leakages [Koc+11], and pancreatic fistulas [Bas+17; Nah+18]; to simplify prediction, we categorize all three endpoints as SSI-III. We translated our model into an actionable alarm framework for medical staff to improve failure-to-rescue management on GWs.

While the use of AI and ML is increasingly researched and employed in many areas of healthcare [Lam+23; Var+24; RKK24], many — often life-threatening diagnoses — remain undetected. Development and deployment of the Early Warning System (EWS) in contexts where data is scarce or of insufficient quality, such as post-surgery nursing wards, is particularly challenging.

4.2 Related Work

In this section, we put our work into context by discussing prospective clinical studies with machine learning and AI applications in mind.

4.2.1 Conducting a Prospective Clinical Study

Publicly available datasets, primarily containing ICU data, have proven valuable for the development and validation of proof-of-principle models. However, due to the high

costs, treatment in the intensive care unit (ICU) remains a highly limited resource in clinical practice, reserved for only a few critically ill patients. Additionally, the predominantly invasive monitoring of vital signs in the ICU is associated with limitations on mobility, which can adversely affect recovery. Consequently, most patients spend their hospital stays on general wards with significantly reduced monitoring. This poses a particular risk for patients undergoing major surgeries, as they are susceptible to rapidly life-threatening complications in the postoperative period, and early detection with subsequent initiation of countermeasures is crucial for morbidity and mortality. One potential solution could be a hybrid ward, which utilizes non-invasive telemonitoring combined with continuous ML-based risk stratification to provide an optimized and resource-efficient monitoring approach.

4.2.2 An Extensive Hybrid Dataset Collected Over the Entire Stay

The dataset used in this study is unique in its comprehensiveness, as it integrates a wide range of data types collected throughout the patient's entire hospital stay. This includes not only clinical data from electronic health records but also real-time monitoring data from wearable devices and patient-reported outcomes. Such a multimodal approach allows for a more holistic understanding of patient health and the factors contributing to postoperative complications.

The Electronic Health Record (EHR) provides considerable administrative benefits. In addition to efficient administration, the data can be repurposed to improve patient treatment and predict clinical outcomes using ML. Using this retrospective data, several datasets have been developed (e.g., Johnson et al. [Joh+22], Thoral et al. [Tho+21], Hyland [Hyl20], and Rodemund et al. [Rod+23a]). Often, these datasets primarily contain heavily monitored Intensive Care Unit (ICU) data due to its abundance and ease of extraction [Shi+19a]. Despite the increased popularity and usage of EHR for ML [Sta+22; Mur+21; Sou+20], there are major limits to retrospective data collection. Due to bias towards ICU data, the monitoring timespan is severely limited, and cohorts often include patients with considerably different treatment histories, which can influence model usability. Additionally, many clinically relevant endpoints are often not available. This means machine learning researchers are restricted to predicting retrospectively defined endpoints [Van+24a] that are not always clinically relevant and, therefore, have no impact on improving treatment outcomes.

After a surgical procedure, patients are stabilized in the ICU for about 24 hours and transferred to the nursing ward afterward. However, complications caused by the surgery often do not occur immediately, but days or even weeks later [Fel+16; Chi+22]. At this stage, the patient is in the nursing ward where, unlike the ICU, continuous monitoring of vital signs is not available. There are several reasons for this seemingly early discharge from the ICU: patient discomfort, such as high noise levels, invasive monitoring, and poorer mobilization — which impedes recovery [JJL18; Mor21]. In addition, ICU resources are costly, limited, and reserved for critically ill patients [Bag+20].

To our knowledge, it is the largest dataset that combines wearable and clinical modalities [Cli+14; Lee+24]. We note that our dataset contains considerable amounts of data points per subject. The dataset is, however, sparse over time as we collect different types of data in each stage of the patient's stay. It contains a mix of high-frequency

time series, such as ICU and wearable data, as well as events that encapsulate a large amount of information, such as lab results.

4.2.3 Integrating Wearables into Clinical Practice

We recognize several earlier works that have worked to integrate wearable data into clinical practice [Bro+14a; Cli+14; Dow+19; Car+19; Edd+22; vRos+23; Pos+24]. These works mainly employed wearable devices as standard monitoring equipment which limits their usability due to low overall data quality. For example, Zahradka et al. [Zah+22] notes that information overload is likely to occur for sensitive, non-ML-powered alarms. Moreover, Posthuma et al. [Pos+24] noted no significant difference between a cohort with wearables and without in normal general wards. van Rossum et al. [vRos+23] describes the development of a sensor-based EWS but, like most studies, does not employ an ML framework that can scale to multiple modalities.

4.2.4 Predicting Surgical Complications

There have been earlier attempts at predicting surgical complications [vBoe+24; Shi+23; Smi+22] of which most use retrospective data. Moreover, our earlier work has focused on performing preoperative risk [Win+24] and 90-day mortality predictions [Rie+25b; Van+25b] using retrospective data. Another avenue, to which we have contributed, is the use of federated learning to build models across multiple hospitals while preserving patient privacy [Pfi+24; Mau+25; Win+25; Pfi+25].

However, most earlier works have focused on predicting complications using only static clinical data or limited vital sign data from the ICU. In contrast, the work in this chapter leverages a wide variety of modalities recorded during the entire patient stay and provides continuous predictions designed to provide an Early Warning System.

4.2.5 Alarm Systems

Clinical alarm systems are designed to alert healthcare providers to significant changes in a patient's condition that may require immediate attention. Traditional alarm systems often rely on fixed thresholds for vital signs, which can lead to high false alarm rates and contribute to alarm fatigue among healthcare staff. Early Warning System have been developed to improve the detection of patient deterioration by combining multiple physiological parameters into a single score [Ede+24]. These scores, such as the National Early Warning Score (NEWS) [Smi+13], have been shown effective in predicting adverse outcomes in hospitalized patients. However, they are typically based on intermittent measurements and may not fully capture the dynamic nature of patient physiology.

Various non-ML alarm systems have been proposed for monitoring patients with wearable devices [Kni+21; McL+21; Zah+22; van+23; vRos+23; Pos+24].

Converting predictive algorithms into practical alarm frameworks requires striking a balance between clinical effectiveness and operational feasibility. Previous research in hospital ward-based patient deterioration monitoring has largely depended on wearable monitoring technologies as independent surveillance tools or on computed EWS

for indicating the physiological state. Nevertheless, these methodologies face limitations due to poor signal quality, as they mostly rely on this single source; this compromises clinical implementation [Dun+20]. While continuous vital sign monitoring (CVSM) may detect physiological modifications before clinical decline, not every patient demonstrates substantial changes, emphasizing the tension between sensitivity and specificity in alarm architecture [Jen+25].

Despite sensor-driven EWS platforms [vRos+23; van+23; Pos+24; Yad+24] or AI-augmented EWS architectures[Ede+24] demonstrating enhanced performance relative to conventional scoring systems, they continue to be restricted by their dependence on a limited spectrum of physiological parameters. Through their absence of multimodal versatility and situational responsiveness, these frameworks risk missing the intricate relationships among demographic, procedural, and clinical variables that influence individual postoperative courses.

4.3 Retrospective Multicenter Prediction of Postoperative Complications

4.3.1 Introduction

Riepe et al. [Rie+25b] and the work in progress Van de Water et al. [Van+25b], both analyze retrospective data collected in a similar clinical setting of visceral surgery as the prospective Clinical ASSist AND aleRt Algorithms (CASSANDRA) study [Van+25c], which is described from the next section onwards. The retrospective work was performed largely during the data collection and infrastructure-building phase of the prospective study. Here, we adapt from the findings of Riepe et al. [Rie+25b], where I performed the ML development and analysis, as this work has already been published. Emergency general surgery cases are covered Van de Water et al. [Van+25b], whereas Riepe et al. [Rie+25b] and this section cover subjects from elective visceral surgery. Although we are severely limited in the number of parameters and the frequency of the collected data compared to the prospective dynamic data later in this chapter, we have a large cohort of 7,711 patients who underwent major elective visceral surgery at two tertiary care centers.

4.3.2 Methods

Ethics approval This study was approved by the Institutional Ethics Committee on July 26, 2023, and conducted in accordance with the Declaration of Helsinki. An additional data protection consultation was performed by the supervising institution (DSO_888), and the study adheres to the TRIPOD+AI statement [Col+24]: updated guidance for reporting clinical prediction models that use regression or machine learning methods [Col+15].

Setting This retrospective cohort study enrolled patients of legal age undergoing major elective visceral surgery at two independent, high-volume, tertiary, university-affiliated surgical centers between January 2014 and December 2022. Center 1 (C_1) served as

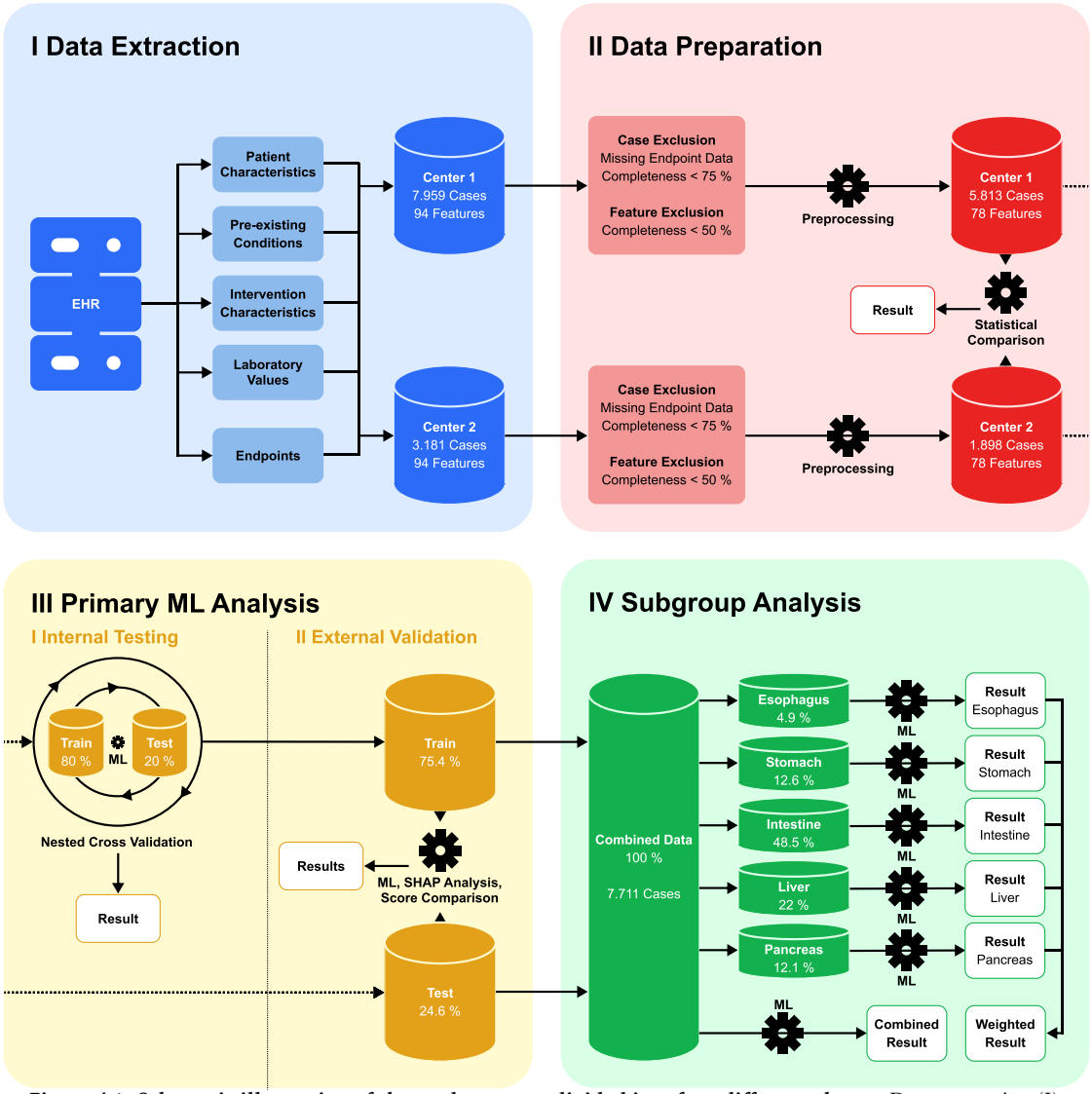


Figure 4.1: Schematic illustration of the study concept divided into four different phases. Data extraction (I), data preparation (II), primary ML analysis (III), and subgroup analysis (IV). EHR = Electronic Health Record, ML = Machine Learning, SHAP = SHapley Additive exPlanations. Used under CC BY-NC 4.0 license from Riepe et al. [Rie+25b].

the internal training cohort, while Center 2 (C_2) was used for independent validation. Major visceral surgery encompassed procedures of the esophagus, stomach, (large and small) intestine, liver, and pancreas, defined based on the German adaptation of the International Classification of Procedures in Medicine [Wor23] (OPS [Bun25], Appendix D.1.1). Surgery involving the esophagus, stomach, liver and pancreas was performed by surgeons who subspecialized in these respective organ systems, whereas general surgeons handled intestinal surgery. 90-Day mortality (90DM) was selected as the primary endpoint, and 30-Day Mortality (30DM) as the secondary endpoint. Exclusion criteria were specified as missing information regarding an endpoint or data completeness below 75% across all features (Figure D.1).

Table 4.1: Statistical analysis of selected features and statistical comparison between the two included centers. Numerical features are given as an average with a standard deviation. Used under CC BY-NC 4.0 license from Riepe et al. [Rie+25b].

#	Feature	Complete (%)	Combined (C_1 + C_2)	Internal (C_1)	External (C_2)	P
1	Gender	94.71				0.27
	Male		54.88% (n = 4232)	54.24% (n = 3153)	56.85% (n = 1079)	
	Female		39.83% (n = 3071)	38.74% (n = 2252)	43.15% (n = 819)	
2	Age (y)	100	57.93 (SD = 15.69)	58.82 (SD = 14.83)	55.21 (SD = 17.79)	<0.01
3	Height (m)	85.25	172.91 (SD = 10.09)	172.91 (SD = 10.17)	172.89 (SD = 9.86)	0.39
4	Weight (kg)	85.44	79.85 (SD = 25.09)	81.97 (SD = 26.9)	73.52 (SD = 17.2)	<0.01
5	BMI (kg / m 2)	85.25	26.62 (SD = 7.71)	27.32 (SD = 8.3)	24.52 (SD = 5.04)	<0.01
6	ASA score	88.33				<0.01
	1		4.92% (n = 379)	4.27% (n = 248)	6.90% (n = 131)	
	2		45.18% (n = 3484)	41.96% (n = 2439)	55.06% (n = 1045)	
	3		36.77% (n = 2835)	40.63% (n = 2362)	24.92% (n = 473)	
	4		1.45% (n = 112)	1.60% (n = 93)	1.00% (n = 19)	
	5		0.01% (n = 1)	0.02% (n = 1)	0.00% (n = 0)	
23	Malignancy	100	65.02% (n = 5014)	71.01% (n = 4128)	46.68% (n = 886)	<0.01
46	Primary system	100				<0.01
	Esophagus		4.94% (n = 381)	4.94% (n = 287)	4.95% (n = 94)	
	Stomach		12.55% (n = 968)	14.83% (n = 862)	5.58% (n = 106)	
	Intestine		48.49% (n = 3739)	39.26% (n = 2282)	76.77% (n = 1457)	
	Liver		21.96% (n = 1693)	26.84% (n = 1560)	7.01% (n = 133)	
	Pancreas		12.06% (n = 930)	14.14% (n = 822)	5.69% (n = 108)	
69	CRP (mg / L)	63.05	19.19 (SD = 40.67)	17.15 (SD = 38.3)	27.73 (SD = 48.46)	<0.01
70	Hemoglobin (g / dL)	99.47	12.53 (SD = 2.09)	12.60 (SD = 2.05)	12.29 (SD = 2.2)	<0.01
76	Erythrocytes	99.31	4.39 (SD = 0.66)	4.40 (SD = 0.65)	4.35 (SD = 0.69)	0.03
79	90-day mortality	100	4.01% (n = 309)	4.04% (n = 235)	3.90% (n = 74)	0.83
80	30-day mortality	100	2.02% (n = 156)	1.96% (n = 114)	2.21% (n = 42)	0.56

Data A total of 94 preoperatively available features were selected for analysis, encompassing binary, numerical, and categorical data referring to patient characteristics, pre-existing conditions, intervention characteristics, and laboratory values. Pre-existing conditions and the revised Charlson Comorbidity Index (rCCI) [Cha+87; Qua+11] were obtained retrospectively as outlined by Quan et al. [Qua+05]. Data preprocessing included the elimination of features with completeness below 50% (Table D.1) and the definition of upper and lower boundaries for numerical features. Categorical parameters were ordinally or one-hot encoded, based on the presence of an intrinsic order. Prior to ML analysis, missing values were imputed using the k-Nearest Neighbors (k-NN) imputer, and numerical parameters were normalized. Finally, binary features were added to indicate missing values. Data processing and analysis were performed using Python (version 3.10). The libraries used are listed in Table D.7.

Statistical methods We compared Patient data between the two included centers. Continuous variables are presented as mean values with standard deviation (SD) and were analyzed using Student's t-test [Stu08] or Mann-Whitney U-test [MW47], depending on the distribution determined using the Anderson-Darling test [AD52]. Binary and categorical features are shown as class frequencies and were compared using the Chi-Square test [Pea00]. Statistical significance was defined at $p < 0.05$. The discriminatory performance of all models was assessed using the Area Under the Receiver Operating Characteristic (AUROC), the Area Under the Precision Recall Curve (AUPRC), the Matthews correlation coefficient (MCC), and the F1-Score. Where different folds or seeds yielded multiple results, these are presented as an average with standard deviation. A schematic illustration of the methodology is presented in Figure 4.1.

Conventional scoring systems The ASA score and rCCI were used as conventional indicators of perioperative risk. To calculate the prediction performance, all score samples of C_1, including imputed values, were used to determine their respective mortality rates by class. These were subsequently applied to the outcome data of C_2 as an external validation to calculate the described evaluation metrics and assess the predictive performance in an independent cohort.

Primary machine learning analysis Internal testing utilized a repeated nested cross-validation setup to split the data of C_1 into a training and test subset at a ratio of 80:20 [CT10]. This approach nests the hyperparameter optimization cross-validation (k=5) inside a second cross-validation (k=5) loop, thereby reducing the risk of overfitting and providing a more reliable estimate of the performance and variability [CT10]. The optimal hyperparameters — including those related to cost-sensitive learning for handling class imbalance — were subsequently selected from a predefined range based on the average performance across 30 iterations. The best-performing model on this inner training data was then evaluated on the left-out test fold. The process was repeated 5 times for each fold, and the entire experiment was repeated for 5 random seeds to prevent bias from initial fold splitting. We trained the following classifiers: eXtreme Gradient Boosting (XGBoost) [CG16], logistic regression (LR), and Balanced Random Forest (BRF) [LNA17]. Finally, it was tested whether the Synthetic Minority Over-sampling Technique (SMOTE) [Cha+02] could further increase performance.

External validation employed the complete C_1 cohort for model training across 5 random seeds to ensure the maximum amount of available training data was utilized. Subsequently, the model was assessed across all C_2 samples. The ML setup and analysis code is publicly available[1].

Additionally, feature importance analysis was performed during external validation using SHapley Additive exPlanations (SHAP) values [LL17b]. At its core, this approach investigates the alteration in model performance by deleting each feature individually, thus allowing the calculation of the outcome contribution of features separately within the final model [LL17b; Sha53]. Feature weights were averaged across 10 random seeds and normalized for final reporting.

Subgroup analysis To evaluate the potential of a combined large-scale model against individual organ-specific approaches, the data from both centers was merged. Using the XGBoost classifier, separate models for esophageal, gastric, intestinal, liver, and pancreatic surgery, as well as a joint multi-organ model, were trained and validated in a manner analogous to the preceding internal testing. The organ-specific performance results were then weighted according to the proportion of their respective organ system to obtain a separately trained weighted average across all individual models. This ensures the same organ distribution as in the combined multi-organ approach, against which the weighted result was then compared. The weighted average of the standard deviation was derived mathematically from the calculated weighted average of the variance.

[1] https://github.com/rvandewater/CASS-PROPEL

4.3.3 Results

A total of 11,140 cases met the inclusion criteria. Following the application of the exclusion criteria, 7,711 (nC_1 = 5,813; nC_2 = 1,898) cases remained eligible and were enrolled in the analysis. The overall completeness across the 78 included features was 93.2% (C_1: 93.3%; C_2: 93%).

Baseline characteristics Statistical key figures between the two centers are given in Table 4.1; additional statistical feature information are found in Table D.2 (patient characteristics), D.3 (comorbidities), D.4 (surgical intervention characteristics), and D.5 (lab features). Most patients in both cohorts were male (C_1: 54.2% [n = 3,153]; C_2: 56.9% [n = 1,079]) without significant distribution differences (p = 0.27). Patients in the internal training cohort were significantly older (C_1: 58.8 [SD = 14.8]; C_2: 55.2 [SD = 17.8]; p < 0.01). Consistently in all organ systems, most cases were oncologic patients (Esophagus: 96.1% [n = 381]; Stomach: 49% [n = 968]; Intestine: 54.6% [n = 3,739]; Liver: 79% [n = 1,693]; Pancreas: 85.5% [n = 930]), yet the overall proportion was significantly higher in the internal (71% [n = 4,128]) than the external cohort (46.7% [n = 886]; p < 0.01). The predominant comorbidities in the internal training cohort were arterial hypertension (45.6% [n = 2,651]), followed by metastasis of a solid tumor (36.7% [n = 2,135]) and diabetes without chronic complications (21.4% [n = 1,241]). The rCCI differed significantly with an average of 5.8 (SD = 3.9) in the internal training and 4 (SD = 3.8) in the external validation cohort (p < 0.01). ASA 2 was the predominant category in both cohorts (C_1: 42%; C_2: 55.1%).

The distribution of organ systems diverged significantly between the two cohorts (p < 0.01). The internal training cohort showcased higher proportions for gastric (C_1: 14.8% [n = 862]; C_2: 5.6% [n = 106]), pancreatic (C_1: 14.1% [n = 822]; C_2: 5.7% [n = 108]) and particularly liver surgery (C_1: 26.8% [n = 1,560]; C_2: 7.0% [n = 133]). Surgery on the intestine, including colorectal surgery, made up the majority of operations in both cohorts. However, they were particularly prevalent in the external validation cohort (C_1: 39.3% [n = 2,282]; C_2: 76.8% [n = 1,457]).

Overall, 30DM reached 2% (n = 156; C_1: 2%; C_2: 2.2%) and almost doubled to 4% regarding 90DM (n = 309; C_1: 4.0%; C_2: 3.9%). No significant differences were found between the two cohorts (p90DM = 0.83; p30DM = 0.56).

Conventional risk stratification The ASA score achieved an AUROC of 0.72 and AUPRC of 0.08 for 90DM as well as 0.75 and 0.05 for 30DM, respectively. In comparison, the rCCI demonstrated higher discriminatory performance in terms of both AUROC (90DM: 0.81; 30DM: 0.78) and AUPRC (90DM: 0.11; 30DM: 0.05). Due to the low mortality prevalence of less than 50% across all categories, the default MCC and F1-Score defined at a decision boundary of 0.5 equaled 0 in all cases. To bypass this, the maximum F1-Score was employed, resulting in 0.15 for the ASA score and 0.2 for the rCCI, both with a threshold of 0.06 for 90DM. Regarding 30DM, a 0.1 F1-score with a threshold of 0.03 was achieved for the ASA score, while the rCCI reached 0.11 for a 0.02 threshold.

Table 4.2: Model performance for the multicenter retrospective analysis for both endpoints following external validation. Results are shown as mean and standard deviation across all seeds in brackets. We Compared with the conventional scoring systems ASA and rCCI as well as chance level. The best performance is highlighted in bold. Used under CC BY-NC 4.0 license from Riepe et al. [Rie+25b].

Endpoint	Classifier/score	AUROC	AUPRC	MCC	F1-score
90DM	XGBoost	**0.86 (0.01)**	**0.20 (0.04)**	**0.28 (0.01)**	**0.30 (0.01)**
	Logistic regression	0.84 (<0.01)	0.19 (<0.01)	0.27 (<0.01)	0.29 (<0.01)
	Balanced random forest	0.84 (0.01)	0.18 (0.01)	0.26 (0.01)	0.28 (0.01)
	ASA score	0.72	0.08	0.00	0.00
	rCCI	0.81	0.11	0.00	0.00
	Chance level	0.50	0.04	0.00	0.04
30DM	XGBoost	0.81 (0.01)	**0.11 (0.01)**	**0.18 (0.02)**	**0.18 (0.02)**
	Logistic regression	0.78 (0.02)	0.10 (<0.01)	0.16 (0.01)	0.17 (<0.01)
	Balanced random forest	**0.83 (0.01)**	0.09 (0.01)	0.17 (0.01)	0.17 (0.01)
	ASA score	0.75	0.05	0.00	0.00
	rCCI	0.78	0.05	0.00	0.00
	Chance level	0.50	0.02	0.00	0.02

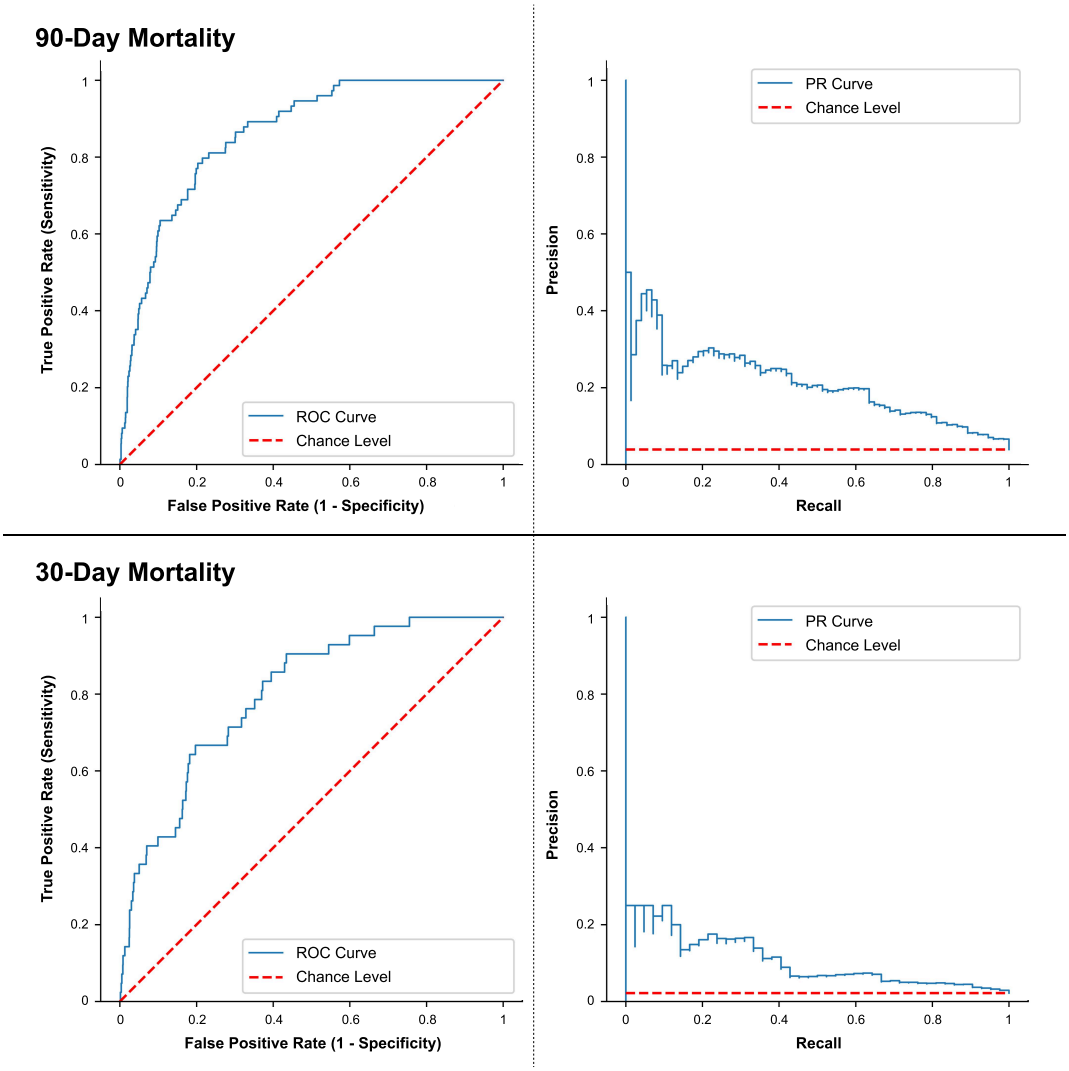


Figure 4.2: Performance of the XGBoost model for 90-day mortality prediction. Left: Receiver Operating Characteristic (ROC) curve. **Right**: Precision-Recall (PR) curve. Both are the result of one fold and show the chance level predictor in this case. Used under CC BY-NC 4.0 license from Riepe et al. [Rie+25b].

Table 4.3: Organ-specific subgroup analysis for 90-day and 30-day mortality prediction. Results of the subgroup analysis for both endpoints are analogous to internal testing using the XGBoost classifier. Values, here, are presented as mean and standard deviation across all folds and seeds. Used under CC BY-NC 4.0 license from Riepe et al. [Rie+25b].

Endpoint	Organ system	Fraction	Prevalence	AUROC	AUPRC	MCC	F1-score
90DM	Esophagus	0.05 (n = 381)	0.05	0.83 (0.04)	0.20 (0.05)	0.24 (0.07)	0.27 (0.06)
	Stomach	0.13 (n = 968)	0.02	0.88 (0.13)	0.24 (0.26)	0.29 (0.30)	0.30 (0.28)
	Intestine	0.48 (n = 3739)	0.03	0.87 (0.05)	0.17 (0.09)	0.24 (0.14)	0.25 (0.12)
	Liver	0.22 (n = 1693)	0.05	0.78 (0.07)	0.24 (0.10)	0.31 (0.10)	0.30 (0.10)
	Pancreas	0.12 (n = 930)	0.09	0.66 (0.08)	0.22 (0.12)	0.25 (0.14)	0.32 (0.11)
	Weighted result	1 (n = 7711)	0.04	0.82 (0.07)	0.20 (0.13)	0.26 (0.16)	0.28 (0.14)
	Combined result	1 (n = 7711)	0.04	0.84 (0.04)	0.21 (0.06)	0.28 (0.06)	0.29 (0.05)
30DM	Esophagus	0.05 (n = 381)	0.02	0.84 (0.04)	0.16 (0.1)	0.24 (0.14)	0.22 (0.13)
	Stomach	0.13 (n = 968)	0.01	0.84 (0.21)	0.30 (0.38)	0.18 (0.33)	0.19 (0.32)
	Intestine	0.48 (n = 3739)	0.01	0.85 (0.07)	0.18 (0.11)	0.26 (0.17)	0.25 (0.14)
	Liver	0.22 (n = 1693)	0.02	0.72 (0.13)	0.09 (0.11)	0.09 (0.13)	0.10 (0.10)
	Pancreas	0.12 (n = 930)	0.05	0.62 (0.13)	0.17 (0.12)	0.23 (0.22)	0.26 (0.17)
	Weighted result	1 (n = 7711)	0.02	0.79 (0.12)	0.17 (0.17)	0.21 (0.20)	0.21 (0.17)
	Combined result	1 (n = 7711)	0.02	0.83 (0.04)	0.14 (0.08)	0.20 (0.11)	0.19 (0.09)

Primary ML analysis During internal testing, all classifiers yielded comparable results for 90DM with an AUROC of 0.85 across all models and AUPRC results ranging from 0.23 (0.04; XGBoost, BRF) to 0.24 (0.05; LR). The MCC varied between 0.28 (0.03; BRF) and 0.31 (0.05; LR), whereas the highest F1-Score was 0.34 (0.05, LR). Prediction models for 30DM showed higher AUROC values (0.87 [0.03]; LR), though lower AUPRC results (0.17 [0.06]; LR). The SMOTE technique did not further increase performance. Overall, LR demonstrated the best model performance during internal testing. Extended results can be found in Table D.6.

The results of external validation demonstrated high reliability and robustness, with only minor performance decreases compared to the preceding internal testing. For 90DM, the XGBoost classifier performed best across all metrics with an AUROC of 0.86 (0.01), AUPRC of 0.2 (0.04), MCC of 0.28 (0.01), and F1-Score of 0.3 (0.01). For 30DM, the discriminatory power was lower, reaching a maximum AUROC of 0.83 (0.01; BRF) and AUPRC of 0.11 (0.01; XGBoost). Detailed results are presented in Table 4.2, and corresponding ROC and PR curves are shown in Figure 4.2. Feature importance analysis identified the rCCI as the highest impact parameter for 90DM prediction (SHAP = 0.17). Thereafter, patient age (SHAP = 0.14), C-reactive protein (CRP; SHAP = 0.14), hemoglobin (SHAP = 0.11) and erythrocytes (SHAP = 0.09) showed high feature weights. Pancreatic surgery was identified as the most relevant organ-related feature (SHAP = 0.06). A complete overview of the 20 most important features according to SHAP values is presented in Figure 4.3.

Subgroup analysis The performance of ML models in organ-specific subdomains varied substantially across the investigated domains. Gastric surgery achieved the best result for 90DM (AUROC: 0.88 [0.13]; AUPRC: 0.24 [0.26]), although the particularly high SD must be considered. Contrary to the 90DM prediction for pancreatic surgery, which yielded the lowest model performance (AUROC: 0.66 [0.08]; AUPRC: 0.22 [0.12]). Considering 30DM, the organ-specific model for surgery of the intestine performed best, as indicated by the highest AUROC of 0.85 (0.07) and good AUPRC of 0.18 (0.11), especially considering the low mortality rate. Conversely, surgery on the liver (AUROC: 0.72 [0.13]; AUPRC: 0.09 [0.11]) and pancreas (AUROC: 0.62 [0.13]; AUPRC: 0.17 [0.12])

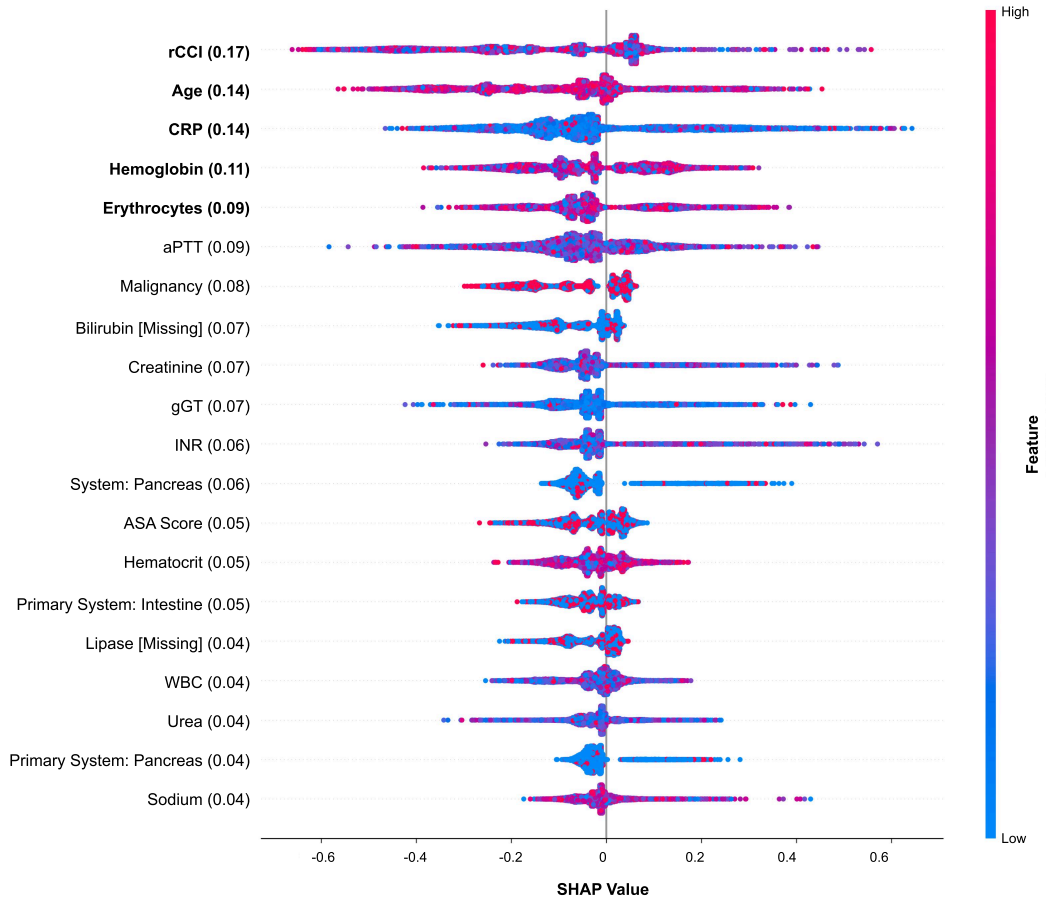


Figure 4.3: Results of SHapley Additive exPlanations (SHAP) analysis with the 20 most important features for 90-day mortality prediction during external validation. The results are shown as a Beeswarm plot with corresponding mean SHAP values. rCCI = revised Charlson Comorbidity Index, CRP = C-Reactive Protein, aPTT = activated Partial Thromboplastin Time, gGT = gamma-Glutamyl Transpeptidase, INR = International Normalized Ratio, ASA = American Society of Anesthesiologists, WBC = White Blood Count. Used under CC BY-NC 4.0 license from Riepe et al. [Rie+25b].

demonstrated a relatively weak prediction performance. Finally, the weighted average across the five organ-specific models yielded an AUROC of 0.82 (0.07) and AUPRC of 0.2 (0.13) for 90DM as well as 0.79 (0.12) and 0.17 (0.17) for 30DM, respectively.

In contrast to organ-specific training, an aggregated multi-organ approach using all samples achieved an AUROC of 0.84 (0.04) and an AUPRC of 0.21 (0.06) for 90DM as well as 0.83 (0.04) and 0.14 (0.08) for 30DM, respectively. Hence, all performance metrics for 90DM and the AUROC for 30DM of the aggregated approach are superior compared to the weighted organ-specific average. Moreover, the combined approach demonstrated a considerably lower SD compared to organ-specific training across all metrics. Detailed results of the subgroup analysis are presented in Table 4.3.

4.3.4 Discussion

This study aimed to evaluate, validate, and compare the potential of ML in preoperative risk stratification for major elective visceral surgery across all as well as for

five individual organ systems. To date, 90DM has been rarely considered for prediction tasks [Ber+18; Hil+19; Chi+20; van+21; GM21; Gra+23; Lee+22b]. However, it is increasingly recognized as a more objective benchmark for perioperative risk assessment compared to in-hospital mortality or 30DM [Res+21; JM21]. Moreover, the lack of external validation still presents a major shortcoming of many surgical ML studies [Ber+18; Ehl+17; Che+18a; Hil+19; Mer+20b; Chi+20; van+21; GM21; Che+22; Che+23; Gra+23; Jun+23].

ML analysis for 90DM achieved AUROC results of 0.85 for internal testing and 0.86 during external validation, indicating high robustness and generalizability of the trained models. Although it remains the most frequently reported metric for ML model performance, solely using AUROC can be limited in discrimination tasks involving highly imbalanced data [SR15]. Consequently, the AUPRC provides a more elaborate evaluation since it is less affected by the large proportion of *true negative* predictions [SR15]. Notably, the AUPRC reached 0.2 during external validation, demonstrating a discriminatory performance five times higher than the baseline chance level of 0.04 (mortality rate). The XGBoost classifier achieved the best results, supporting recent findings that gradient-boosted decision trees are particularly suitable for this use case [McE+23]. Model performance for 30DM was considerably lower across all metrics, despite its shorter prediction time frame. This finding may be attributed to the naturally lower proportion of positive samples compared to 90DM, which has adverse effects on model training and may thus outweigh the shorter prediction horizon. All ML models surpassed the ASA score and rCCI as conventional risk stratification methods for both endpoints. This observation is in accordance with previous studies suggesting that ML can indeed outperform conventional risk scores such as the ASA score [Ber+18; Hil+19; Mer+20b; Chi+20; van+21; GM21; Lee+22b], Preoperative Score to Predict Postoperative Mortality (POSPOM) [Hil+19; van+21], American College of Surgeons (ACS) risk calculator [Ber+18; Mer+20b; GM21] and rCCI [Ehl+17; Hil+19; van+21] for mortality prediction.

Large-scale ML models for outcome prediction in the domain of surgery are still scarce. An analysis involving nearly 500,000 patients after non-cardiac surgery yielded an AUROC of 0.94 for 30DM [Lee+22b]. However, this study also included minor surgeries, resulting in a very low mortality rate of less than one percent [Lee+22b], which is unfavorable in terms of AUROC assessment. Contrary to the reported AUPRC of 0.16, which was lower than in this study's model [Lee+22b], though the lower mortality rate must be considered for comparison. Furthermore, two studies investigating IHM with cohorts exceeding 50,000 cases also achieved AUROC results over 0.9 [Hil+19; Gra+23]. However, they did not differentiate between major and minor procedures, thus yielding a considerable class imbalance. While Graeßner et al. [Gra+23] reported a model AUPRC of 0.11, this metric is not given by Hill et al. [Hil+19], making the evaluation of results based on this unbalanced data challenging. Moreover, neither study considered external validation.

Interestingly, organ-specific training demonstrated considerable performance differences across individual organ systems; however, the different cohort sizes and mortality rates should be noted. The best model was achieved for gastric surgery, outperforming the results of two large coherent multicenter ML studies for 90DM after oncologic gastrectomy in terms of AUROC [Per+22; Cer+24] and AUPRC [Cer+24].

Likewise, the individual model for 30DM prediction in patients undergoing surgery of the intestine demonstrated a high AUROC of 0.85, surpassing the performance of an extensive nationwide ML study in colorectal cancer surgery [van+21]. Yet, a more elaborate comparison is again challenging as no AUPRC is given. In contrast, organ-specific models for patients undergoing liver and pancreatic surgery showed relatively weak discriminatory power for both endpoints. A study applying a multilayer perceptron (MLP) for surgical outcome prediction after hepatocellular carcinoma resections reported an AUROC of 0.84; however, comparison is difficult, as the study is limited to in-hospital mortality [Shi+12].

Notably, the combined multi-organ training approach conducted in this study surpassed the weighted average of organ-specific training across all metrics for 90DM and the AUROC for 30DM. Although the benefit is minor, it suggests that aggregating several heterogeneous groups into one large cohort may be advantageous for the sake of increasing the training sample size. More importantly, the combined approach also showed a substantially lower SD compared to the weighted average of organ-specific training. This demonstrates the increased reliability and robustness following this strategy. To optimize combined approaches more effectively towards individual organ systems in the future, transfer learning might be a promising approach, as discussed in our other work Winter et al. [Win+25]. Here, a source model can be derived from training based on a large, though less specific, cohort, with subsequent organ-tailored fine-tuning for each subdomain. Additionally, collaborative techniques such as federated learning could be employed to establish even larger cohorts in a privacy-sensitive environment in the future; this is discussed in our other work Pfitzner et al. [Pfi+25].

To date, decision support tools are considered by less than a quarter of surgeons to aid their decision-making process [Pra+22; Blo+20]. Low confidence in accuracy as well as a lack of transparency regarding internal result calculation have been identified as major barriers to a broader acceptance and clinical implementation [Pra+22]. Therefore, it is crucial to provide medical professionals with more detailed model insights to increase confidence.

In this section, the rCCI was identified as the most decisive model factor, which aligns with previous clinical studies using conventional statistics [Bha+23; Cha+16]. Additionally, advanced age, being the second most important feature, is a well-recognized risk factor for postoperative mortality [Pea+06; Jha+08]. Notably, hemoglobin and erythrocytes both ranked among the top five model weights. This is particularly interesting, as preoperative anemia has been associated with increased perioperative morbidity. Implementing optimized perioperative blood management has been shown to improve patient outcomes [Luo+20]. Finally, pancreatic surgery was identified as a feature of major importance, which is consistent with the well-documented high morbidity and mortality rates associated with these procedures both in this study and previous research [Nun+09]. Comparing the identified predictors with established clinical evidence reveals a high level of concordance between the respective risk factors. This may increase confidence in these models and promote integration into clinical practice.

The study in this section encountered several limitations. The retrospective study design may have had quantitative and qualitative effects on the data. However, automated data extraction methods were employed to minimize human error and ensure a high-quality, complete dataset. Additionally, only patients providing follow-up data for

at least 90 days were considered for analysis. As patients with uncomplicated courses are less likely to attend follow-up appointments after discharge, the actual 90-day mortality rate may be lower than it appears in our cohorts. Moreover, corresponding ASA values were not available for all cases, resulting in missing values that needed to be imputed; therefore, these values no longer accurately reflected the actual medical assessments performed by a physician. Finally, this study does not differentiate between different socioeconomic subgroups, as advocated by the TRIPOD+AI guideline [Col+24]. However, given the low-threshold access to healthcare in Germany due to generalized public health insurance, we assume that the local population structure is well represented in the data.

Conclusion ML presents a promising preoperative risk stratification approach for the prediction of 90DM in patients undergoing major elective visceral surgery, surpassing conventional risk stratification approaches. Leveraging training across multiple organ cohorts may improve ML performance, particularly model robustness, and hold potential for advancing both further research and clinical applications. However, future research, particularly prospective clinical studies, using a wide range of modalities and clinically actionable endpoints, is needed to assess the potential of such models fully.

4.4 Prospective Clinical Study Design and Population

As noted by the previous section which described a retrospective study in the visceral surgery domain, there is a lack of prospective clinical studies investigating the potential of machine learning (ML) in a complex multimodal setting. Moreover, the previous section and related work has mostly addressed severe endpoints that, in many cases, are not clinically actionable. To address these gaps, we designed and conducted a prospective clinical study, the Clinical ASSist AND aleRt Algorithms (CASSANDRA), to evaluate the feasibility of continuous multimodal AI supported by wearables for predicting clinically relevant postoperative complications during the general ward (GW) stay in patients undergoing major abdominal surgery. The ethics committee of Charité University Hospital provided formal approval (EA4/198/21) on 20 October 2021 for the CASSANDRA study. The study was registered in the German Clinical Trial Register (DRKS - Deutsches Register Klinischer Studien) under the study ID DRKS00034973[2]. This study was funded by the "Gemeinsamer Bundesausschuss (G-BA) Innovationsausschuss" in the framework of "CASSANDRA – Clinical ASSist AND aleRt Algorithms" (project number 01VSF20015). Participants provided written informed consent for the use of their data in this study.

A total of 1,285 patients were prospectively enrolled in the study after providing written informed consent before the surgery, either in the outpatient clinic or on the surgical ward. These patients were consecutively identified upon admission for a planned major surgery, which fell within the predefined list of eligible surgical procedures.

Exclusions were applied as follows(see Figure 4.4):

[2] https://drks.de/search/de/trial/DRKS00034973

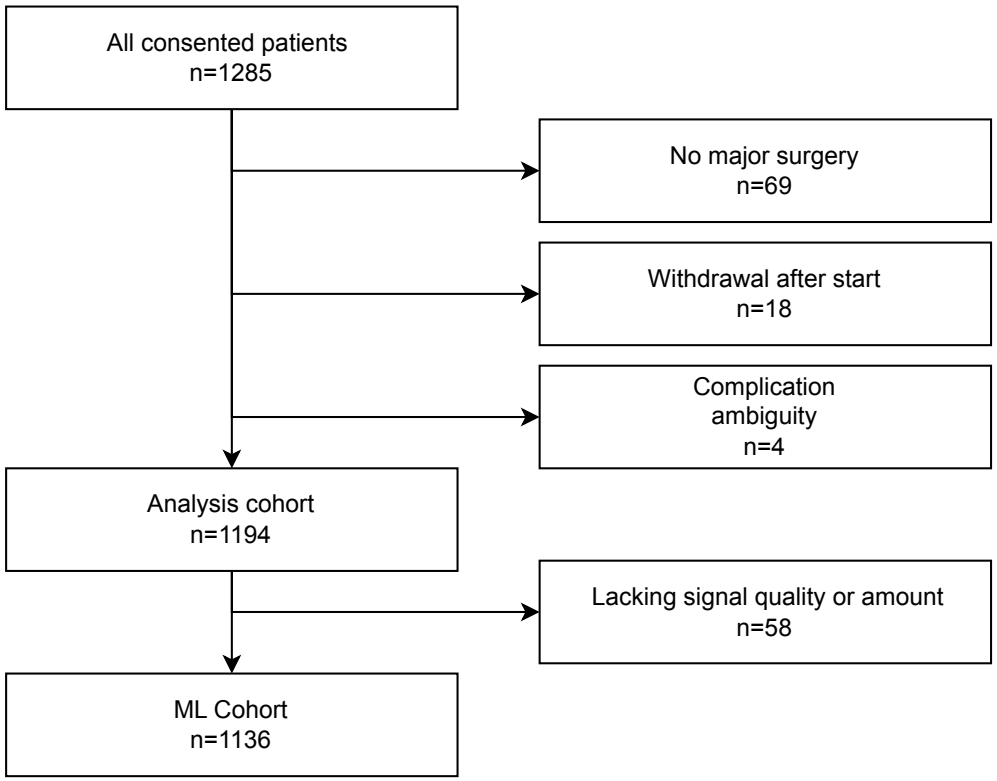


Figure 4.4: Schematic exclusion criteria. We use the analysis cohort for data exploration and analysis purposes, and created a final ML Cohort used to train all models.

- Major surgery not performed (n=69): Patients who were scheduled for a major operation, but ultimately did not undergo the procedure.
- Withdrawal from study (n=18): Patients who withdrew from the study after enrollment, either declining to continue participation or refusing to wear the required monitoring devices.
- Outcome adjudication (n=4): Patients who were excluded during the outcome adjudication process due to ambiguity in defining the presence or timing of postoperative complications. Their clinical trajectory was marked by different complications, and the censoring time point was not precisely defined.
- Lacking signal quality or amount (n=58): after retrieving, cleaning and analyzing the signal quality, if there was no PPG data at all, patients were excluded. For example, patients did not wear the device properly, which prevented it from getting a reading.

After exclusions, the final cohort included 1,136 patients for model development and evaluation. Patients were classified into four mutually exclusive groups based on the presence, severity, and timing of postoperative complications identified during their hospital stay. Patients are censored after reaching their first major endpoint and after a

clinical prediction timeline of 20 days. Additionally, patients are censored after experiencing a major complication different from our endpoint.

The *target complications* are the most common highly granular complications after major abdominal surgery of the upper and lower gastrointestinal tract, major hepatobiliary and pancreatic surgery, are deep organ space surgical site infections (SSI-III), bile leakages [Koc+11], and pancreatic fistulas [Bas+17; Nah+18]. These endpoints were combined because they all represent intra-abdominal pathologies associated with fluid collections that required intervention. These endpoints were subsumed under SSI-III for clinical reasons, reflecting everyday clinical practice.

Patients were classified into four mutually exclusive groups based on the presence, severity, and timing of postoperative complications identified during their hospital stay. Complications were first identified prospectively and later confirmed through a structured chart review using strict clinical definitions based on the Clavien-Dindo classification (CD)[Kat+16] classification:

- Group 1 – *No Complications:* Patients with no documented postoperative complications of any kind throughout their hospital stay.
- Group 2 – *Only Minor Complications (CD I&II):* Patients who had only minor complications during their stay—those not requiring significant interventions (e.g., conservative treatment only)—and no major complications at any time.
- Group 3 – *Target Major Complication First (CD III & IV)*: Patients who experienced at least one major complication, with the first major event matching the predefined target complication (e.g., intra-abdominal infection, anastomotic leak). This group reflects early clinical deterioration directly related to the study outcome of interest.
- Group 4 – *Other Major Complication First (CD III & IV):* Patients who also experienced a major complication, but whose first major event was a different complication type than the predefined target. This group helps identify alternate deterioration patterns and competing risk events.

In cases where patients experienced both minor and major complications, the classification was based on the first major complication, as it marks the most clinically relevant point of deterioration. This logic ensures stratification by both severity and temporal sequence, aligning with the intended evaluation of early warning performance.

4.4.1 Outcome Annotation

The primary endpoint of this study (target major complication) was the occurrence of clinically relevant intra-abdominal complications. In this context, *clinically relevant* referred to complications that required an intervention—surgical, radiologic, or endoscopic. This definition corresponds to complications graded higher than CD II, the most widely used classification system for postoperative complications among clinicians. Other major complications were defined as complications graded higher than

CD II, excluding surgical site infections grade III (SSI-III), bile leakage, and postoperative pancreatic fistula (POPF). A typical example would be postoperative respiratory failure requiring readmission to the intensive care unit and noninvasive ventilation (CD IVa). Minor complications were defined as all complications with a CD Grade less than III (i.e., pneumonia requiring antibiotic therapy). The precise definitions of the target complications are below.

SSI-3 : Surgical Site Infection grade III (SSI-III) is defined by the presence of at least one of the following criteria:

1. Purulent drainage from an organ/space drain (e.g., closed suction drain, open drain, T-tube, or CT-guided drain).
2. Isolation of one or more organisms from fluid or tissue within the organ/space by culture or non-culture–based microbiological testing performed for clinical diagnostic or therapeutic purposes (excluding active surveillance testing).
3. Presence of an abscess or other evidence of organ/space infection detected by
    - gross anatomic or clinical examination,
    - histopathological assessment, or
    - imaging studies demonstrating definite or possible infection.

Bile leakage: In this study, Bile Leakage is defined as bilirubin concentration in the drain fluid at least 3 times the serum bilirubin concentration on or after postoperative day 3 or as the need for radiologic or operative intervention resulting from biliary collections or bile peritonitis.

Postoperative Pancreatic Fistula (POPF): The patient is clinically unstable, often requiring management in the intensive care unit. The patient remains *nil per os* (no food or drink) with complete parenteral nutrition. Computed Tomography (CT) imaging typically shows worrisome pancreatic fluid collections necessitating placement of a new drain. Antibiotic therapy and somatostatin administration are often required. Since the primary endpoint required an intervention, only type C pancreatic fistulas were included in the primary outcome analysis. Postoperative complications were identified prospectively by a visceral surgeon and a general physician through daily review of clinical data. To ensure accuracy, a second visceral surgeon independently performed a manual chart review after the availability of final laboratory and microbiological results following the interventions (typically within 2–5 working days). This multi-step process was implemented to guarantee rigorous and unbiased outcome ascertainment. These outcome assessments were conducted by physicians not involved in the treatment of study patients, ensuring objectivity. Wearable devices used to monitor vital signs had no visible display, making the data inaccessible to both patients and treating physicians. This ensured that clinical decisions, such as ordering Computed Tomography (CT) scans, were based solely on the treating physicians' clinical judgement, entirely independent of the study's monitoring data, preserving unbiased outcome identification. In a subset of cases, a Computed Tomography (CT) or Magnetic Resonance

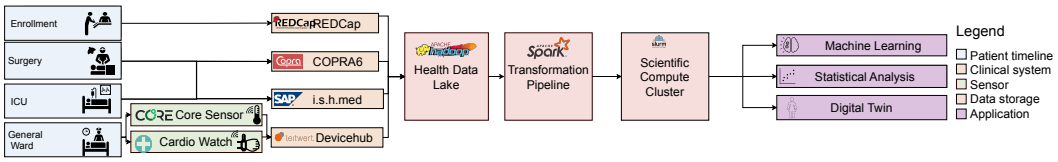


Figure 4.5: Data collection and processing pipeline. From left to right: **1)** simplified time steps of patient treatment process **2)** patient monitors worn on nursing ward, **3)** the respective Patient Data Management System (PDMS) involved with data collection and storage, **4)** the extraction pipeline that ingests data from the PDMS using proprietary APIs, **5)** on-premises Hadoop-based data lake **6)** Spark-based transformation pipeline and further preprocessing on compute cluster, **6)** High-Performance Computing infrastructure used to construct ML pipelines, perform patient analytics, and produce visualizations.

Table 4.4: Patient journey described in terms of the interval spent in each part. Values shown are the median and [25/75 percentile]. D: days, H: hours. Adapted and modified from Daniela Zuluaga Lotero's Master thesis.

	All Patients (n = 1194)	**No Complication** (n = 813)	**Major Complication** (n = 381)
Preoperative period (D)	1.00 [0.92–1.12]	0.99 [0.89–1.11]	1.02 [0.95–1.15]
Surgery duration (H)	4.46 [2.92–5.92]	4.05 [2.67–5.55]	5.28 [3.90–6.78]
Postoperative period (D)	7.82 [4.97–13.91]	5.88 [4.22–8.28]	17.95 [12.22–27.87]
ICU stay (D)	1.72 [0.88–2.91]	1.00 [0.87–2.13]	2.03 [0.91–3.91]
Ward stay (D)	6.04 [4.08–11.00]	4.99 [3.96–6.94]	14.98 [9.97–24.04]
Length of stay (D)	9.18 [6.17–16.26]	7.17 [5.34–10.16]	20.26 [14.15–30.21]

Imaging (MRI) scan was performed before any intervention was indicated. The most common triggers for imaging, in descending order, were: highly abnormal laboratory values, Suspicious Fluid in the drainage, physical condition of the patient, and incidental findings on imaging performed for other indications.

4.4.2 Data Sources and Variables

We simplified the patient care process to four stages: patient admission, surgery, ICU stay, and general ward stay (Figure 4.5). We used a diverse range of data sources for modelling patients in our study. Several Patient Data Management System (PDMS) were used to collect data. Admission information was recorded using a RedCAP [PB18] database clinical trial application. The main PDMS was I.S.H.Med [Gel+03], which is primarily used to collect stay information and laboratory results. The ICU vital signs were sourced from the COPRA system [24a], which was used to collect vital signs, scores, and other ICU data. The wearable data was collected using the Corsano CardioWatch (CW) 1 (CardioWatch 287-1B) and CW 2 (CardioWatch 287-2B), wrist-worn devices that collect various physiological signals: photoplethysmography (PPG), accelerometry, and skin temperature. The CW 2 also includes an oxygen saturation sensor. Note that, while functionality to perform manual measurements of e.g., "electrocardiogram (ECG)" – a simulated ECG measurement as this requires several chest sensors – is available, only the automatic signal measurement is used. One of the key features of these devices is the raw PPG output, which we can use to calculate different heart rate-related features [Mej+22]. Moreover, we measured the core temperature with the GreenTEG CORE (GTC) device. Table 4.5 provides an overview of the available aggregated features per device and their sampling frequencies for these devices. The wearable device data is collected using a Bluetooth device gateway mesh that transmits the data to a central server via the DeviceHub software [Bra+20]. Many features were renamed and aggregated after data exploration had determined that they represent the same mea-

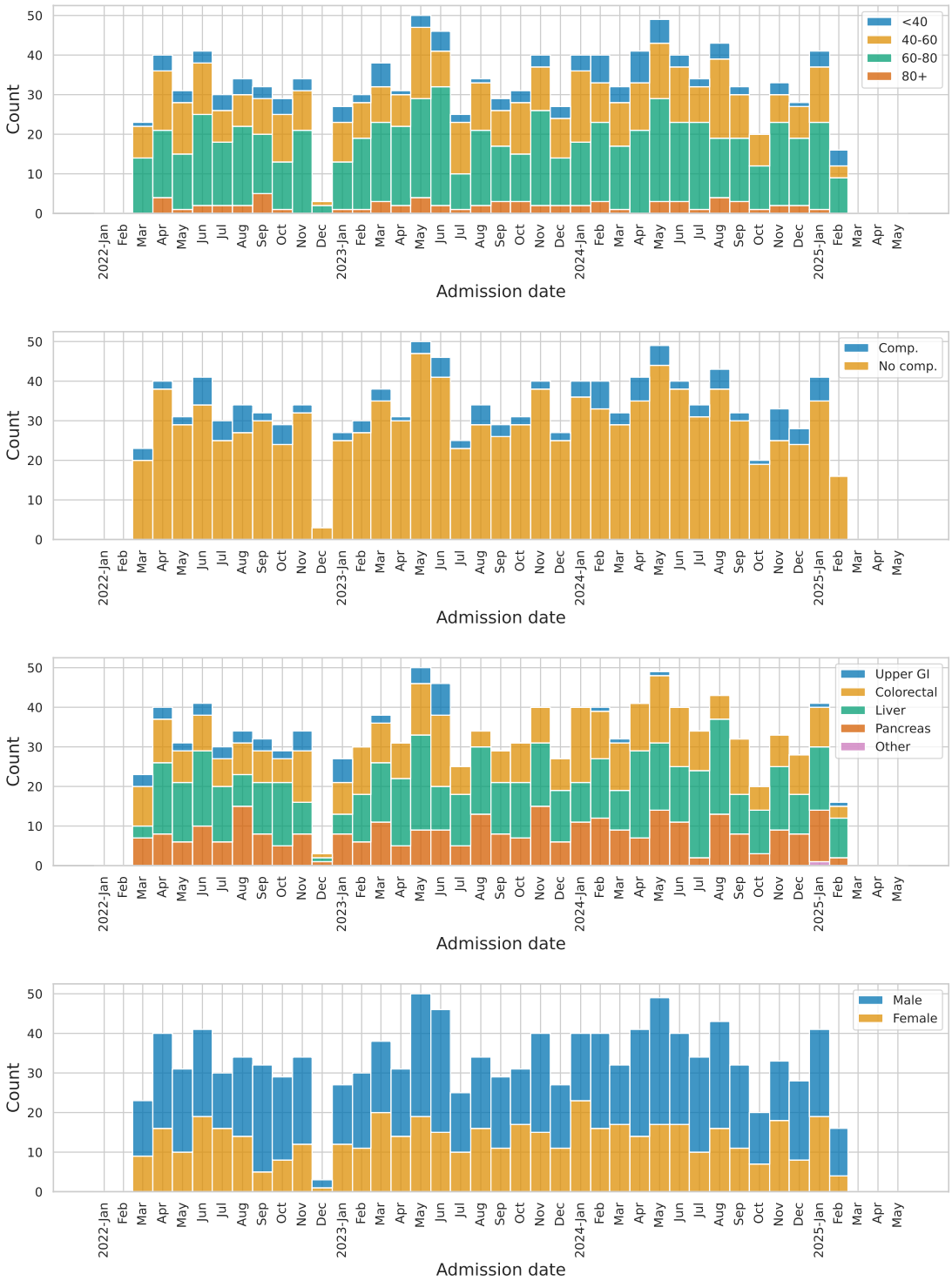


Figure 4.6: Distribution of subjects by surgical procedures over the study period binned monthly by the admission period. Histograms are stratified by age group (top), complication status (second), organ system (third), and sex (bottom).

Table 4.5: Overview of aggregated features and sampling frequency for GreenTEG CORE (GTC) chest patch, Corsano CardioWatch (CW) V1, and CW V2 wrist-worn devices. Note that aggregated features are derived from raw sensor data, including PPG signals, and have a lower sampling frequency. Information is extracted from the Corsano developer documentation (`https://developer.corsano.com/`) and GTC documentation (`https://help.corebodytemp.com/en/articles/10445532-viewing-and-analyzing-core-data`).

Feature	Description	CORE	Corsano V1	Corsano V2
Timestamp	Date and time of measurement	✓	✓	✓
TempQuality	Quality of skin temperature measurement	✓		
SkinTemp	Measured skin temperature value	✓	✓	✓
CoreTemp	Computed core body temperature	✓		✓
Activity	Activity classification type		✓	✓
EnergyExpenditure	Energy expenditure in kcal		✓	✓
HRFiltered	Filtered heart rate value		✓	✓
HRQuality	Quality of filtered heart rate		✓	✓
RespFiltered	Filtered respiration rate value		✓	✓
RespQuality	Quality of filtered respiration rate		✓	✓
RespRaw	Raw respiration rate value		✓	✓
Speed	Current speed of the wearer		✓	✓
StepCount	Number of steps counted		✓	✓
WearingStatus	Device wearing status		✓	✓
BatteryLevel	Battery charge level			✓
CadenceRaw	Raw cadence measurement			✓
CadenceQuality	Quality of raw cadence measurement			✓
CardioFitnessIndex	Quality estimate of cardio fitness index			✓
EnergyExRaw	Raw energy expenditure			✓
EnergyExIntermittent	Intermittent raw energy expenditure			✓
EnergyExIntermittentQuality	Quality of intermittent raw energy expenditure			✓
EnergyExQuality	Quality of raw energy expenditure			✓
HeartRateRaw	Raw heart rate value			✓
HeartRateRawQuality	Quality of raw heart rate			✓
RespirationRateRawQuality	Quality of raw respiration rate			✓
SpO2	Blood oxygen saturation			✓
VO2maxQuality	Quality estimate of VO2max measurement			✓
Totals		**4**	**14**	**27**
SamplingFreq	Aggregated feature sampling frequency	**1 Hz**	**6 Hz**	**0.2 Hz**
Subjects	Number of unique subjects collected	**1047**	**331**	**773**
PPGFreq	PPG sensor sampling frequency	–	**25 Hz**	**32 Hz**

surement. Values were translated from German to English and standardized to a common unit of measurement. The timestamps are harmonized from the source time zone to the UTC+0 (GMT) time zone. Moreover, all features are strictly filtered to their respective time periods within each segment to ensure that no data leakage occurs; all features extracted from any events that occur in the current segment occur either during the segment or before it.

4.4.3 General Study Cohort Description

Table 4.4 summarizes the time intervals spent in each part of the patient journey for all patients, as well as stratified by complication status. Figure 4.6 shows the distribution of surgical procedures over the study period stratified by relevant variables. The admissions are relatively evenly distributed with some outliers, such as in December 2022. Moreover, the upper GI procedures are more common in the first half of the study period. Figure 4.7 shows the age distribution of the study cohort stratified by sex. We observe slight differences in age distribution, with female subjects often being younger. Certain surgery groups are performed at older ages (e.g., upper GI surgeries). Most complications occur in the 60-70 year age group.

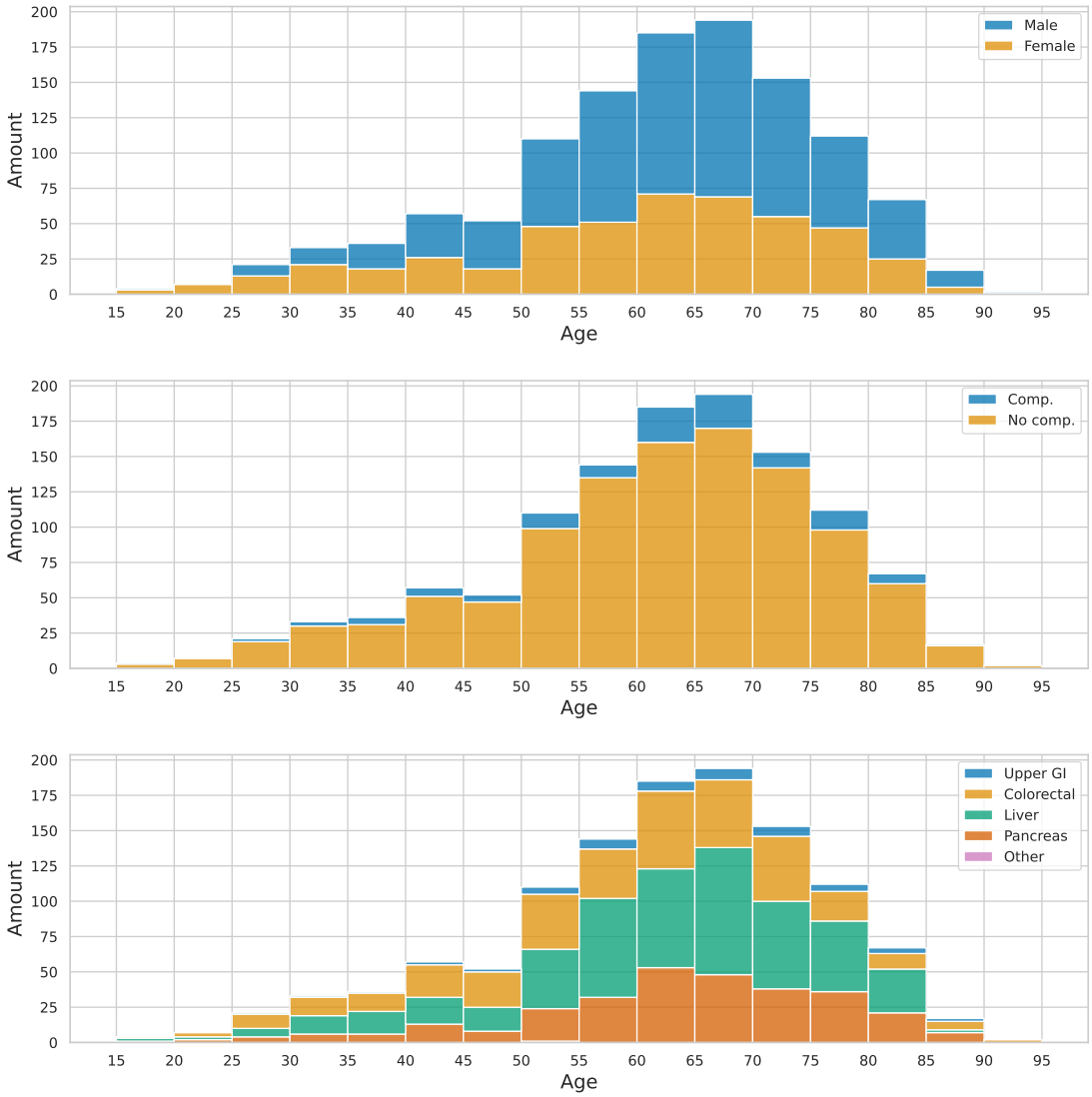


Figure 4.7: Age distribution of the study cohort subjects. Histograms are stratified by sex (top), complication group (middle), and target organ system (bottom).

4.5 Data Processing and ML Design

This section describes the data processing and machine learning design choices that make up the Clinical ASist Early Warning System (CASHEWS).

4.5.1 Domain-specific Processing and Fusion

We use early modality fusion [ZZG24] to combine low-level features after performing initial preprocessing, feature extraction, and (where applicable) embedding. The segmented combined features were input to the modelling pipeline.

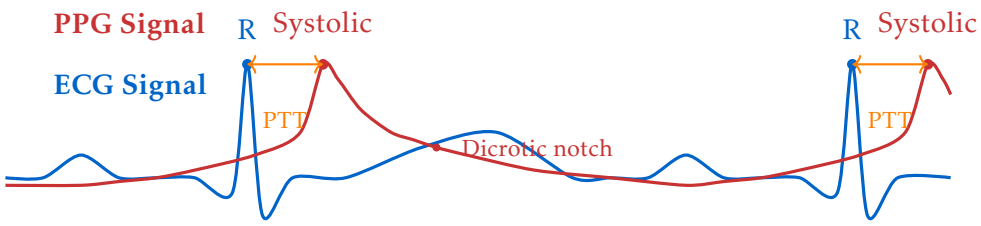


Figure 4.8: Comparison of the dynamics of electrocardiogram (ECG) and photoplethysmogram (PPG) signals. Upper trace (red): PPG signal showing sharp systolic peaks and dicrotic notches characteristic of arterial blood volume changes. Lower trace (blue): ECG signal showing sharp, biphasic QRS complexes representing electrical cardiac activity. Both signals are time-aligned but differ significantly in morphology, with the PPG reflecting hemodynamic (blood volume) changes and the ECG capturing electrical depolarization. The Pulse Transit Time (PTT) reflects the time it takes to detect the pulse at the PPG sensor, worn on the wrist.

Numerical ICU Monitoring and Lab Values

These features were treated uniformly. They are aggregated per subject into 30-minute segments after experimentally determining the right segment length. For each segment, we extract the following from each of the aggregated measurements: mean, minimum, maximum, standard deviation, count, skewness, entropy, median, and sum.

Wearable Data Feature Extraction

The preprocessing pipeline first ensures that all PPG traces are standardized to 32 Hz and normalized per patient (by subtracting the mean and dividing by the standard deviation). Then, it splits the traces into contiguous stay-aligned segments and fixed-length windows (default: 60 s, 1920 samples) with optional overlap; windows with a standard deviation of zero are skipped. For embedding, MiniRocket applies deterministic, extremely sparse 9-length convolutions with only three nonzero taps (84 unique 3-tap combinations) at multiple log-spaced dilations to capture patterns across temporal scales suitable for 32 Hz sampling. For each kernel/dilation pair, a set of scalar bias thresholds is assigned using a low-discrepancy quantile sequence computed from empirical convolution outputs, resulting in diverse binary detectors. At transformation time, each dilated convolutional response is filtered with the proportion of positive values (PPV) operator (1 if the response exceeds the bias, otherwise 0) and then pooled by computing the temporal mean (proportion positive), yielding compact, shift-robust statistics per detector. Implementations include a Numba-accelerated [Lam+25] CPU path that computes dilations, biases, and PPV pooling, and an optional output. Outputs are saved per segment as Parquet rows with feature columns named by index and segment length. This design emphasizes reproducibility, computational efficiency, and scale for large PPG cohorts sampled at 32 Hz.

The wearable data is processed in multiple ways to ensure that the high-resolution data can be used for the real-time alarm model. The raw PPG data is preprocessed with our pipeline:

1. Remove duplicates from the data and sort by datetime.
2. Segment data by identifying longer gaps (60s).
3. Resample signal to clean sampling rate (25/32Hz, depending on device).

- Fill in gaps by cubic spline interpolation.

4. Apply 4th Order Chebyshev Type II Filter (0.5-5Hz).
5. Detect motion artifacts (outside STD) from accelerometer data:
   - Set motion artifacts to NaN and interpolate.

Further preprocessing is then applied to efficiently feed the data into the segmentation pipeline:

1. Segmentation into 5-minute snippets.
2. Calculation of heart rate variability (HRV) features using NeuroKit2 [Mak+21]:
   - Time Domain.
   - Frequency Domain.
   - Non-linear Domain.
3. Reduction of 76 features down to 32 by aligning with features used by tools such as Kubios [Ele96; SG17].
4. Statistical aggregation in 30-minute, 1-hour, and 6-hour periods.

Table 4.6: Description of the collected wearable data. When applicable, variables are summarized by median [Interquartile Range (IQR)]. The gap sizes were chosen because of the nature of the signal; the CW outputs a high-resolution PPG signal, where small gaps can result in signal disturbances. The GTC outputs the precomputed temperature.

Category	Subcategory	Value
Corsano CardioWatch (CW)	Subjects	1100
Version	Corsano V1	331
	Corsano V2	773
Recording duration (h)	Per patient	91.96 [61.74, 145.32]
	Total	128,058.99
Missing data per gap size	< 1s	0.63 [0.03, 1.18]
	1s – 5s	10.21 [0.53, 18.40]
GreenTEG CORE (GTC)	Subjects	1047
Recording duration (h)	Per patient	110.08 [69.25, 165.00]
	Total	142,545.22

Medications and Clinical Note Features

Clinical notes were recorded by nurses and physicians as part of routine care and are written entirely in German, using medical domain language (including, e.g., domain-specific terms). We optimized the clinical texts by manually determining and replacing common abbreviations, anonymizing personal identifiers, and removing non-informative footers. Medications and fluid balance information from the ICU and the general ward were manually harmonized by clinical assistance: measurements and values considered

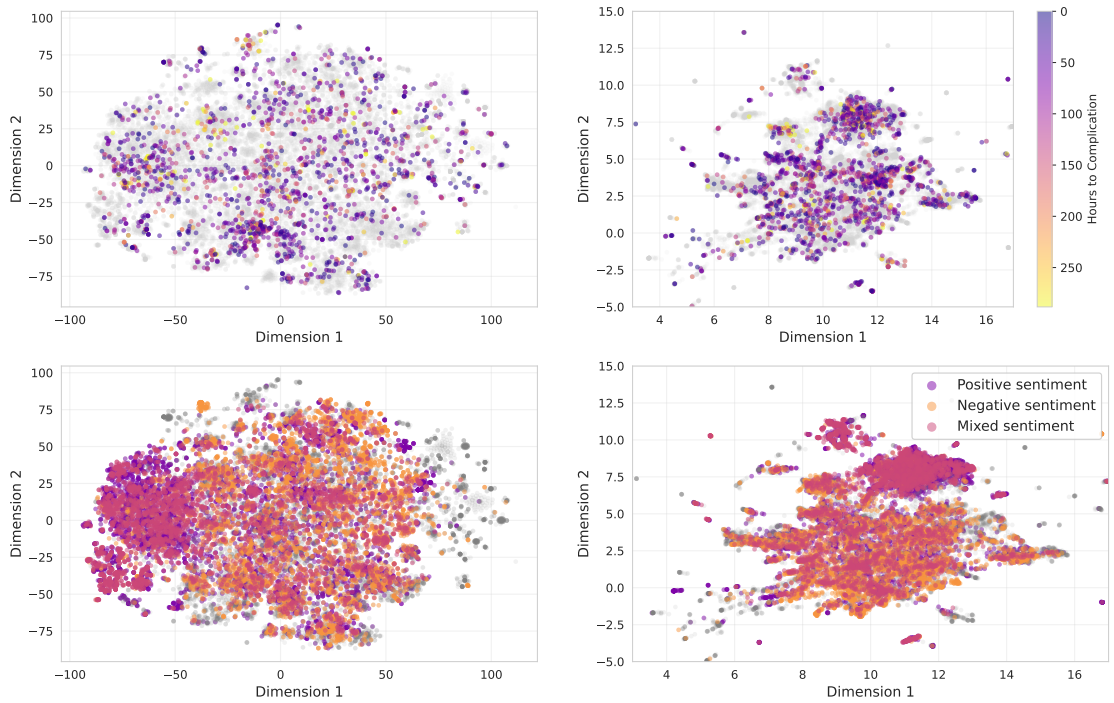


Figure 4.9: Clinical notes embedded and visualized by their proximity in hours to the endpoint after applying dimensionality reduction techniques. Left column: t-distributed Stochastic Neighbor Embedding (t-SNE) visualization of clinical notes embeddings. **Right column:** Uniform Manifold Approximation and Projection (UMAP) visualization of clinical note embeddings; the X- and Y-axis limits were reduced to exclude outliers. **Top row:** heatmap showing the proximity to the endpoint in hours. **Bottom:** showing notes with positive sentiment against negative sentiment (defined by manual regular expressions). Top visualizations suggest that the embeddings do not clearly separate patients who experienced the adverse event (higher risk) from those who did not (lower risk). Bottom visualizations indicate that notes with negative and positive sentiment demonstrate some clustering behavior. Sentiment is defined by German regular expressions and synonyms generated with LLMs; negative: *(schmerzen | schmerzhaft | schmerzhaftigkeit | beschwerden | leid | weh | wehtuend | klagt | klagen | klagte | übel | übelkeit | schlechtsein | unwohlsein | unpässlichkeit | nausea | brechreiz)*, positive: *(gut | wohl | besser | beschwerdearm | beschwerdefrei | schmerzfrei | schmerzzufrieden | schmerzarm | stabil | konstant | unauffällig | unauffaellig | unverändert | unveraendert | in\s+gutem\s+zustand | in\s+bestem\s+zustand)*, mixed (both at least one positive and at least one negative word): *"(?s)(?=.*negative_regex)(?=.*positive_regex).*"*.

the same or very similar were treated as the same feature. The clinical notes and free-text medications are fed into a flexible, pre-trained, general multi-lingual embedding model (jina-embeddings-v3 [Stu+24]) from the HuggingFace pre-trained model repository [Wol+20] to obtain contextualized embeddings. The clinical notes dimensions were set at 64 dimensions, and the medications at 128 dimensions, after observing improved performance from these settings. These values were selected based on the model's performance in preliminary experiments. The embeddings were then aggregated by taking the element-wise mean into the same segments as the numerical features. Table 4.6 provides an overview of the collected and recorded signals. Figure 4.9 shows the clinical notes embeddings visualized with t-distributed Stochastic Neighbor Embedding (t-SNE) [vdMH08]and Uniform Manifold Approximation and Projection (UMAP) [MHM20] dimensionality reduction techniques. The top row visualizations suggest that the embeddings do not clearly distinguish between patients who experienced the adverse event and those who did not. The bottom row visualizations indicate that notes with negative and positive sentiment demonstrate some clustering behavior.

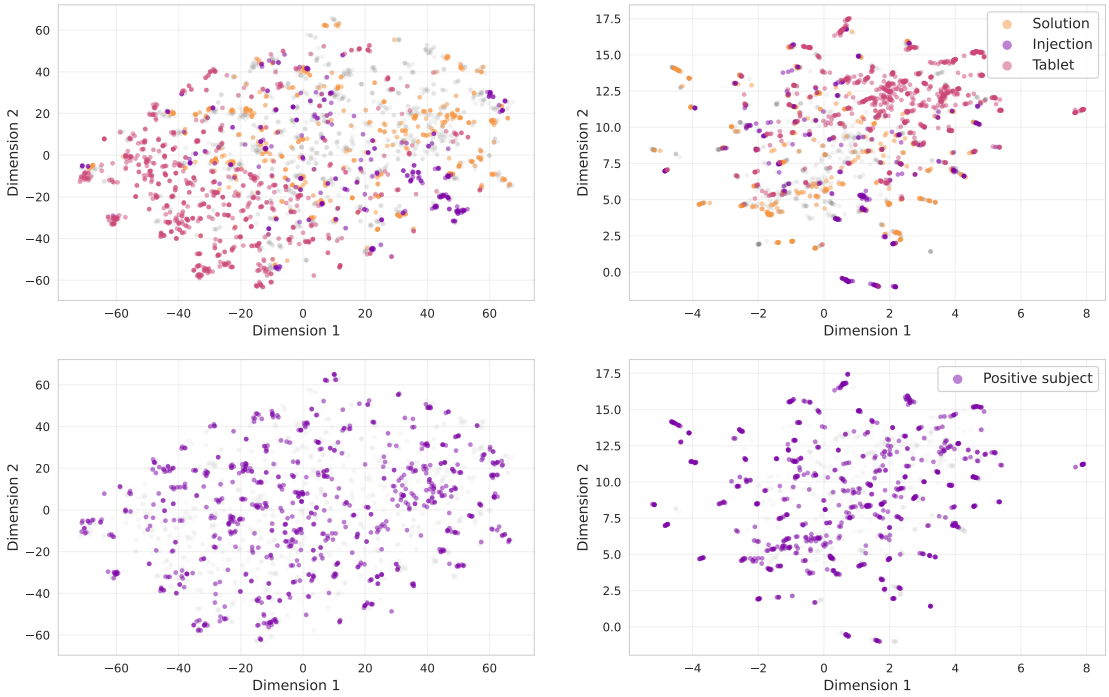


Figure 4.10: Medication texts embedded and visualized after applying dimensionality reduction techniques. **Left:** t-SNE visualization of medication embeddings [Stu+24]. **Right:** UMAP visualization of medication embeddings. **Top:** medication types visualized by color. **Bottom:** medications prescribed to who experienced the adverse event are marked; note, that this includes all medication that is taken by these subjects and includes routine medication (that all subjects receive).

Figure 4.10 shows the medication embeddings visualized with t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction techniques. The top row visualizations show medication types visualized by color, where there seems to be some clustering (as expected). The bottom row visualizations show medications prescribed to patients who experienced the adverse event, compared to those not prescribed to positive cases; there does not seem to be clear clustering behavior. We expect that these textual modalities will not significantly contribute to model performance. In some sense, this is desirable as it indicates that there is no obvious data leakage from the notes or medications due to incorrect timestamps (for example, when notes are registered or altered post-hoc) that would make the prediction task trivial.

Data Fusion and Segmentation

The data segmentation is a parallelized pipeline that segments the data into time intervals for monitoring across the patient's trajectory (ICU and general ward). It aggregates information for a patient within an adjustable time period, while also truncating the data after the adverse event to prevent data leakage and produce a dataset suitable for training an EWS. The feature extraction pipeline is designed to extract features from the segmented data. It encompasses a range of techniques, including statistical analysis, time-series decomposition, and domain-specific feature engineering, to capture the relevant information for each modality. Multi-horizon historical aggregation is per-

Table 4.8: The number of features in each of the most granular modalities.

Modality	N Features	Modality	N Features
Medications	187	Hour	12
Intra-OP Aggregated	972	Laboratory Values	1998
Comorbidities	274	PPG Embeddings	2017
Note	98	Wearable VS	576
ICU Fluids	36	Temperature	108
ICU Vital Signs	372	HRV Features	195
ICU Scores	432	**Total**	7280

formed to capture the temporal dynamics of the data, aggregating the data backwards at 30-minute, 1-hour, 6-hour, 12-hour, and 24-hour intervals.

We can flexibly add new modalities; adjust segmentation periods and add/process new patient timelines, extract a range of adjustable feature sets based on statistical features and retrospective, summarized time-series data; modify prediction horizons and exclude patients without certain modalities; aggregate data from periods before the monitoring starts (preoperative, operative, postoperative, ICU). Table 4.8 shows the resulting number of features per modality. To evaluate the temporal bin for our continuous analysis, data were aggregated into discrete time segments ranging from 7.5 minutes to 6 hours

4.5.2 Machine Learning Development and Analysis Pipeline

We expand an existing framework built for developing and benchmarking ICU models described in Chapter 3.

Preprocessing: In accordance with previous work, we performed runtime preprocessing steps to minimize data leakage between the training and test datasets.

The pipeline operates on three data segments: static patient characteristics, dynamic time-series measurements, and outcome labels. It processes training, validation, and test splits consistently using fitted transformations from the training data.

For static features, missing value indicators are generated across all static variables, followed by scaling and normalization procedures. Numeric variables undergo zero-fill imputation to address missing values, while categorical variables are processed through a two-step approach involving mode imputation to fill gaps, followed by label encoding to convert categories into numerical representations. The system creates missing value indicators for specified variables while respecting user-defined exclusions, then implements forward-fill imputation to propagate the last observed values forward in time. Any remaining missing values are addressed through zero-fill imputation. The pipeline concludes by generating comprehensive historical aggregate features, including minimum, maximum, count, and mean values computed over time windows to capture temporal patterns in the data. The system supports recipe caching for transfer learning scenarios, allowing preprocessing pipelines to be saved and reloaded. All transformations are fitted on training data only and consistently applied to validation and test sets. This preprocessing approach addresses common challenges in data, including irregular sampling, missing values, and the need to combine static patient characteristics with time-varying physiological measurements while maintaining temporal relationships and preventing data leakage.

Table 4.9: The modified REWS (Remote Early Warning Score) adapted from van der Stam et al. [van+23].

Vital sign	Range	Score
Heart rate (bpm)	<40 or >130	3
	40–50 or 111–130	2
	51–100	0
	101–110	1
Respiratory rate (breaths/min)	<8 or >30	3
	9–11 or 21–30	2
	12–20	0

ML Architectures

Several architectures were used: Deep learning (Transformers [Vas+17], GRU [Cho+14a], TCN [BKK18]), Gradient Boosted Trees (GBT, including XGBoost [CG16], LightGBM [Ke+17], and CatBoost [Pro+19]), Random Forest, and Logistic Regression. Adding more model architectures is easily possible due to the framework's flexibility.

We selected the final classifier after thorough benchmarking, which aligns with earlier work, indicating that boosted tree ensembles are a strong state-of-the-art classifier and often outperform Deep Learning models, such as Transformers [Bro+25; GOV22b], especially when domain-guided feature extraction is employed.

This system produces probabilities for a temporal window of a patient's stay. Each experiment was performed as a 5-times repeated 5-fold cross-validation to ensure the performance is balanced over the whole collected dataset. The test set was defined as 20% of subjects. The validation set was defined as 20% of the training subset. The patient timelines were split as a whole to ensure that data leakage between time windows could not occur (any patient window in the train set was not in the test set). Hyperparameter optimization was performed using predefined, publicly accessible configuration files. The traditional machine learning models were tested over 100 iterations of 5-fold cross-validation in a manner that the test set contained a unique subset of subject IDs each time. The Deep Learning models were instead trained with 50 iterations of tuning. Model calibration was performed, but needs to be redone when training the final deployment model for the population.

Analysis platform All analysis and model training were conducted on the central General Data Protection Regulation (GDPR) compliant Berlin Institute of Health's central HPC cluster (Charité-HPC). It has >5000 CPU cores, >80 GPUs, and >1.5PB high-performance data storage. However, all our model training finishes within 48 hours except for hyperparameter tuning. Spark, Python with a.o. PyTorch, NumPy, Polars, Scikit-Learn, XGBoost 3.0.0, CatBoost 1.2.5, LightGBM 4.6.0, ImbLearn, and Optuna are part of the experiment platform. More details can be found in Appendix D.4

4.5.3 Explainability

The explainer plot visualizes the importance and impact of features in the machine learning model using SHapley Additive exPlanations (SHAP) values for the CASHEWS+ model with 250 features ranked by the highest SHAP contribution. The main visualization is a violin plot that shows the distribution of SHAP values for the top features

across all predictions. Each feature is displayed with. The violin bodies represent the density distribution of SHAP values (wider sections indicate more predictions at that value). The individual data points, each comprising 1000 sampled values per violin, display the scaled magnitude of the feature value. In a SHAP plot, positive values indicate that the feature increased the prediction (higher risk), while negative values indicate that it decreased the prediction (lower risk). The features are ranked by their median absolute SHAP value, with the most influential features at the top.

To assess the contribution of different data modalities to model predictions, we aggregated SHAP values by their source modality. For each modality group, we computed the total absolute sum of SHAP values across all features and experiments to capture overall importance regardless of direction. To enable comparison across modalities with different feature numbers, we applied min-max scaling to transform aggregated values into a 0-100 range.

The plot includes an t-SNE visualization that projects the high-dimensional space of average feature values into 2D. This dimensionality reduction technique aims to demonstrate the distinctiveness of each feature.

4.5.4 Alarm System Design

Alarm generation was based on a set of parameters optimized through hyperparameter tuning across 500 thresholds. The final configuration was selected with clinical feasibility in mind, striking a balance between early detection and practical implementation constraints. An 8-hour blackout period was applied after transfer to suppress early postoperative alarms. *Persistence* required three consecutive activations above the lower alarm threshold within a six-hour period to prevent transient fluctuations from triggering false alarms. This feature emphasized rising trajectories rather than isolated peaks, helping to identify sustained risk increases likely to precede complications. An 8-hour *silencing period*, where the alarm will not be triggered, follows each alarm to reduce alarm fatigue. To generate the precision–recall curve (Figure 5), the *lower threshold* was systematically varied to obtain recall bins while other parameters remained constant.

4.5.5 Statistical Methods

Unless specified otherwise, the solid lines and performance metrics presented in all figures and tables represent results from the test set of the 'held out' test data split described previously. The episode-based precision-recall curve was calculated according to an earlier work that explicitly described this evaluation method [GYH22]. For box plots, the central line represents the median value, while the lower and upper boundaries correspond to the first quartile (Q1) and third quartile (Q3), respectively. The whiskers extend to show the first data point above Q1 – 1.5 × (Q3 – Q1) for the lower whisker, and the last data point below Q3 + 1.5 × (Q3 – Q1) for the upper whisker. Any points beyond these whiskers are marked as outliers. In violin plots, we used a bandwidth of 0.2 for the Gaussian kernel density estimation, ensuring the density distribution did not extend beyond the actual data range.

4.5.6 Adaptation to Different Clinical Settings

We adapted and thoroughly expanded an open-source ML benchmarking and development framework in active use by the Health AI community [Van+24a] to work with any hospital stay timelines, not just ICU data. This provides a modular framework that allows researchers to define reproducible and comparable clinical ML experiments. Combined with a transparent preprocessing pipeline and extensible training code for multiple ML and deep learning models, *YAIB* enables unified model development, transfer, and evaluation. While working with our data collection framework, we added numerous features during the experimentation and development of our multimodal dataset, including modality selection, explainability (via Shapley) capabilities, and outcome horizon adjustment. These resources are again made available for the community to use in developing and benchmarking their own models, datasets, and preprocessing approaches at the open-source repositories of *Yet Another ICU Benchmark* (*YAIB*)[3] and the Clinical ASsist Heuristic Early Warning System (*CASHEWS*)[4]. The comprehensive Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)-AI checklist for this study is provided in Appendix D.6.

4.6 Results

Table 4.10: Description of the collected data in the CASSANDRA study. Numeric variables are summarized by Numeric variables are summarized by median and percentiles: [25%, 75%]. Categorical variables are indicated as percentages of the entire cohort (%).

Years collected	2022-2025	**Surgery System**†	
Data points per patient	673,881 [455,829, 1,021,370]	Upper GI	50 (4.4%)
Number of patients	1,136		
Age at admission (years)	63.0 [53.0, 71.0]	Colorectal	336 (29.6%)
Female	455 (40.1%)	Liver	472 (41.5%)
Complications		Pancreas	277 (24.4%)
Deep Surgical Site Infections (SSI-III)	111 (9.8%)		
Of which occurred on GW	96 (86.5%)	**Wearable data**‡	
Time to SSI-III* (days)	6.9 [4.0, 8.0]	Wrist monitor	1104 (97.2%)
Hospital death	8 (0.7%)	Corsano Band 1	331 (29.1%)
Hospital length of stay (days)	9.9 [6.2, 18.3]	Corsano Band 2	773 (68.0%)
ICU length of stay	1.4 [0.9, 2.9]	Core Device	1047 (92.2%)
GW length of stay	6.0 [4.1, 10.3]		
No. of variables	7,280		

*From the moment of admission.
†Patients can undergo surgery for multiple systems; here, the primary surgery system is referred to.
‡Due to logistic issues in data collection, not every patient has usable wearable data; see discussion.

Table 4.10 describes characteristics of the included patient cohort. We will refer to all endpoints above as SSI-III for consistency with clinical practice standards and clarity.

Digital patient representations were synthesized, spanning the entire perioperative journey from preoperative patient baseline data to discharge or the first occurrence of the designated complication. Telemetric vital sign monitoring was implemented using a wrist-based PPG sensor, including an accelerometer, and body core

[3] https://github.com/rvandewater/YAIB
[4] https://github.com/rvandewater/CASHEWS

temperature sensors attached to the chest. The representations integrate heterogeneous data integrated using early multimodal fusion across multiple hospital information systems—including demographics and comorbidity profiles; anesthesia and operating-room records; procedural and intraoperative parameters; postoperative ICU vital sign-data; medication and fluid administrations; laboratory results; and free-text clinical notes—which were transformed into static covariates, high-frequency and intermittently sampled physiologic time-series, and unstructured text encoded using embeddings. Figure 4.11 describes the entire study and the model development pipeline for the *CASHEWS* and *CASHEWS+* models, which additionally incorporate wearable monitoring.

The study dataset included 13.8 billion PPG data points collected continuously at 32 Hz, 5.4 million intensive care unit (ICU) measurements, and 2.5 million lab values in total. We segmented the data at 30-minute intervals, resulting in 7,277 total extracted features per segment. Multiple model types were evaluated, including Deep learning (Transformers [Vas+17], GRU [Cho+14a], TCN [BKK18]), Gradient Boosted Trees (GBT: XGBoost [CG16], LightGBM [Ke+17], CatBoost [Pro+19]), Random Forest, and Logistic Regression.

Figure 4.12 illustrates the performance of various model architectures using segmented data. Several architectures are shown: Transformers [Vas+17], LSTM [HS97], GRU [Cho+14a], XGBoost [CG16], LightGBM [Ke+17], CatBoost [Pro+19], Logistic Regression, Random Forest, and Gradient Boosting [Ped+11b]. Adding more model architectures is possible due to the flexibility of our framework. We see that the differences between the Gradient Boosting Models are minor. These results confirm findings in earlier work, which indicated that boosted tree ensembles are a strong state-of-the-art classifier and often outperform Deep Learning models such as Transformers [Bro+25; GOV22a], especially given domain-guided feature extraction. LightGBM was chosen because it balanced a high Area Under the Receiver Operating Characteristic (AUROC) and Area Under the Precision Recall Curve (AUPRC), with fast training, which were our primary evaluation criteria.

Figure 4.17b shows the results of testing different amounts of top-k features selected based on SHAPley [LL17b] values of an XGBoost experiment. We observe that using feature selection improves performance; however, above $k = 250$, there is little difference in performance. Using all features leads to a slight drop in performance, which could be due to overfitting, given our relatively high amount of features as compared to individual subjects.

4.6.1 Dynamic risk stratification for general ward complications

Figure 4.13a (AUROC) and 4.13b (AUPRC) present the prediction performance of models built at successive stages along the perioperative patient trajectory. We compare our approach with an existing non-ML risk stratification score: the Remote Early Warning Score (REWS) [van+23]. Pre-operative information alone failed to predict postoperative complications, as indicated by an AUROC of around 0.50. Similarly, adding intraoperative data and subsequently ICU features did not improve predictability. By contrast, the integration of routine GW into a dynamic model (*CASHEWS*) demonstrated a considerable increase. The integration of continuous telemetric vital sign monitoring

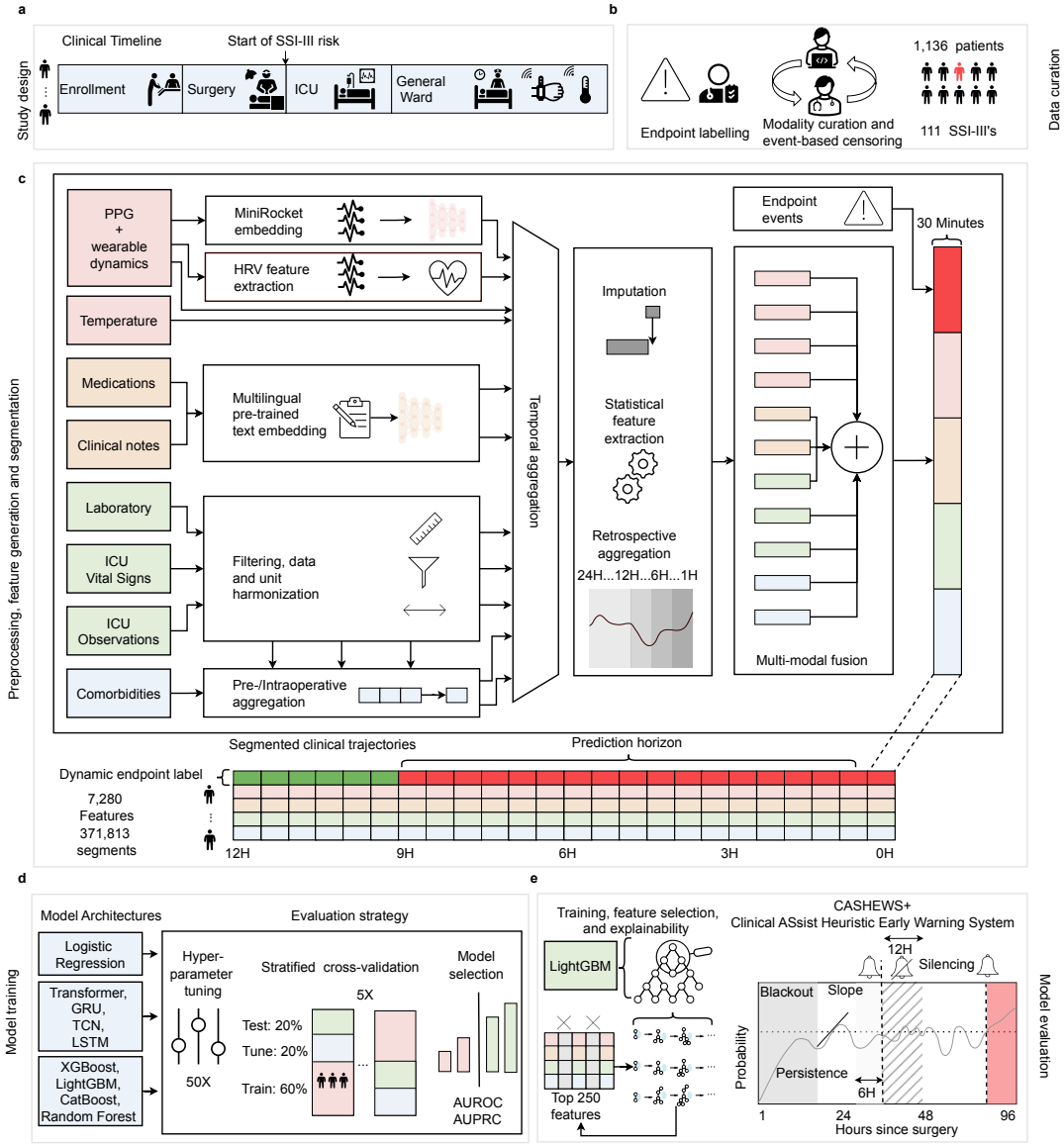


Figure 4.11: Our data segmentation and aggregation pipeline for training *CASHEWS+*. a: The patient journey of surgery patients. **b:** Data curation is done by precise endpoint labeling and extraction from 4 patient information systems. **c:** Segmented clinical trajectory creation: Data from the different modalities is processed in a proprietary manner. PPG data is fed into the time-efficient MINIROCKET [DSW21] embedding model, and Heart Rate Variability (HRV) features are extracted from time-domain and frequency-domain data. Clinical texts, medications, and fluids are embedded using a pre-trained, multi-lingual JinAI [Stu+24] model, and we extract common natural language processing (NLP) features. The clinical data is filtered, harmonized and cleaned. Curated comorbidities are also used, and all clinical features are aggregated to create data from before the monitoring period. Then, we aggregate all data within this time interval, perform missing value imputation, extract statistical features (skew, entropy, standard deviation, minimum, and maximum), and aggregate data from previous timesteps to ensure that we capture the data dynamics. Multimodal fusion and endpoint insertion result in the segmented clinical trajectories with a prediction horizon of 9 hours. **d:** Model training: we test a range of common architectures through tuning hyperparameters and performing cross-validation, after which we choose the best performing model in terms of predictive performance and speed: LightGBM. This model architecture is then evaluated for the best features through Shapley explanations. The resulting model serves as an alarm system, featuring mechanisms to prevent erroneous evaluations. Specifically, a "blackout" will prevent activations within the first 12 hours. The "slope" coefficient and "persistence" are used to quantify an increase in the probability of sounding an alarm. A 12-hour "silencing" period is implemented to prevent alarm fatigue.

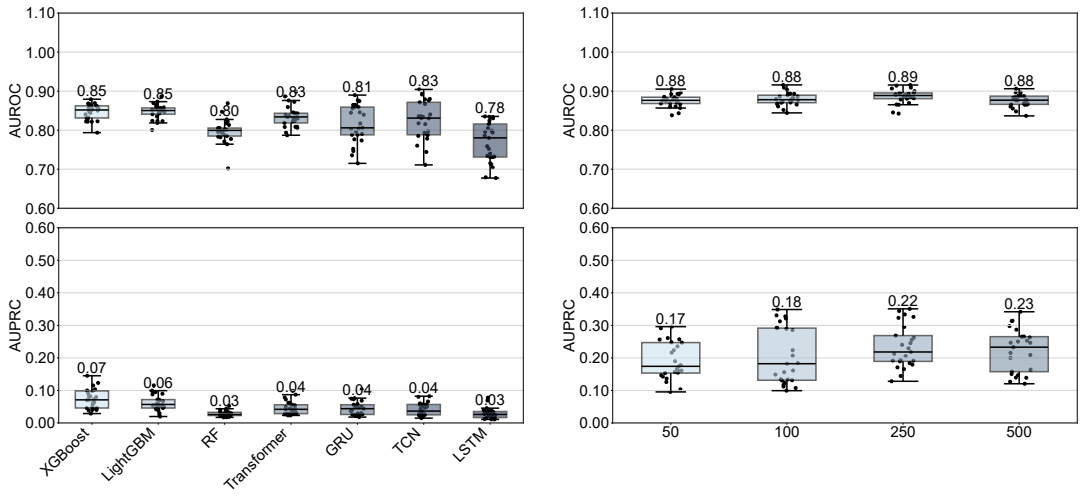


Figure 4.12: Experiments to determine model architecture and the amount of selected features. a) Comparison of different model architectures. For the non-gradient boosted models, a feature selection of the 500 most important was performed based on SHAPley [LL17b] values of an XGBoost experiment. Here, we evaluate the window-level AUPRC. b) Comparison of top-k feature selection on the number of features used in the LightGBM model using SHAPley values. AUPRC is reported at the episode level [GYH22].

(*CASHEWS+*) resulted in an 8 percent increase in AUROC and a 109 percent increase in AUPRC. Moreover, *CASHEWS+* significantly outperformed *CASHEWS* in all 25 cross-validation folds (Wilcoxon signed-rank test [Woo08]: W=0, p<0.0001, for both AUPRC and AUROC). To ensure this result was not an artefact of training set overlap, we applied the Nadeau-Bengio corrected paired t-test [NB03], see Code Listing D.1. The improvement in AUROC remained significant ($t = 4.67$, $p < 0.001$), as did the gain in AUPRC ($t = 3.40$, $p = 0.002$).

Applying the calculated telemetry-based REWS yielded substantially lower results (AUROC: 0.55; AUPRC: 0.02). Note that AUPRC results are driven by prevalence, which is on average 0.12 for the Pre-/Intra-/Postoperative models and 0.006 for the dynamic models. Evaluating the detection at several points in time prior to the event of interest, the highest detection performance was observed between an hour and three hours preceding the clinical recognition of the complication, with a detection rate of 0.90 ± 0.03 and 0.90 ± 0.01 (median ± standard deviation), respectively. Analysis of segments up to 24 hours before detection revealed a steady increase in accuracy leading up to the event. In this context, prediction at a 24-hour lead time still achieved an AUROC of 0.81 ± 0.03.

Finally, the bottom row demonstrates that predicting endpoints solely in the ICU, despite the lower prevalence, is considerably easier for *CASHEWS+*. The evaluation on only the GW segments still performs reasonably (AUROC 0.81 ± 0.03). This result suggests that we might benefit from even more continuous monitoring capabilities on the general ward and that detecting the endpoint is non-trivial in a mobile ward.

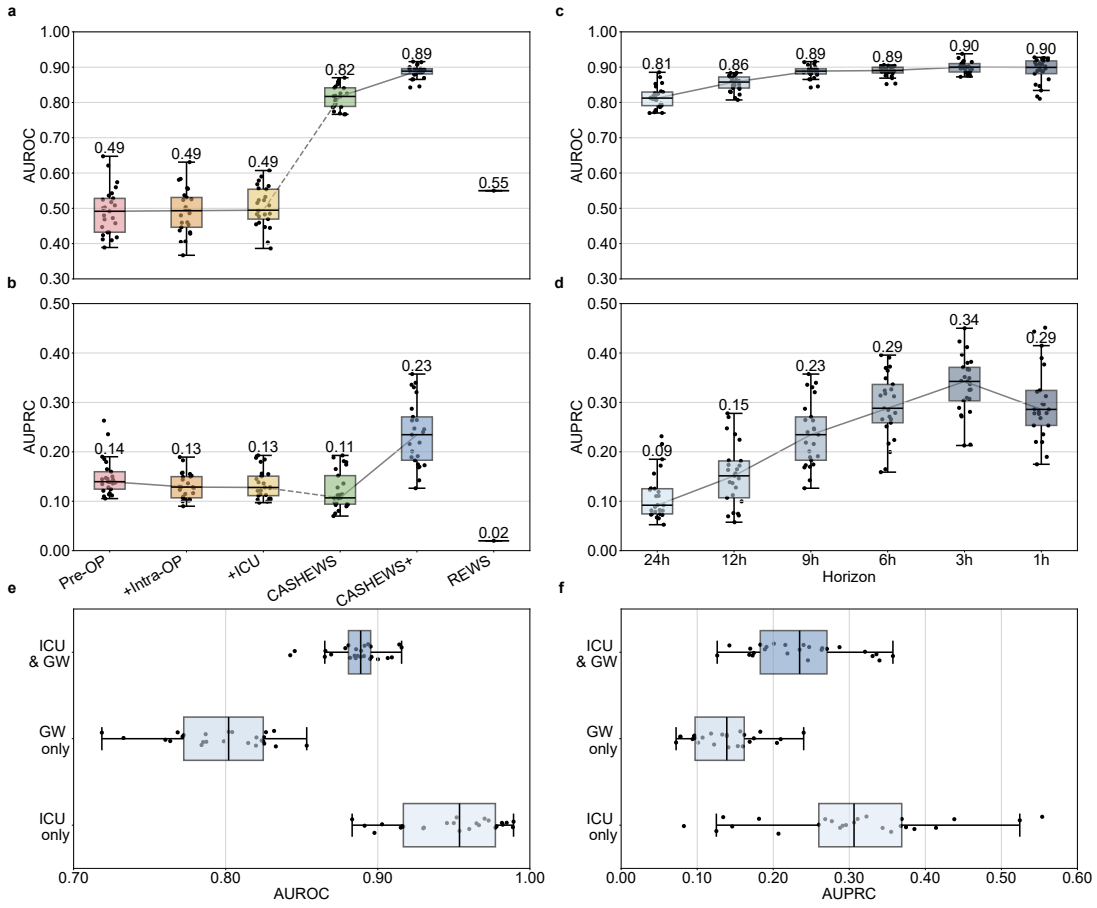


Figure 4.13: Model performance comparison: **a)** The Area under the receiver operating curve (AUROC) and **b)** patient-level area under the precision recall curve (AUPRC) model progression, from left to right: preoperative static variables (Pre-OP), variables until the end of the surgery (+Intra-OP), variables until the end of the ICU (+ICU). The second group of methods is dynamic (real-time)with a horizon of 9 hours: the dynamic (online prediction) model without wearable data (*CASHEWS*), the complete model with the 500 most important features selected (*CASHEWS+*), and the remote early warning score (REWS) [van+23]. The online prediction systems are calculated with snoozed episode-level precision-recall [GYH22]. **c), d)** Comparison of the time horizon of detection of our endpoints. Here, we set the *last hour* to be positive and remove all subsequent time windows from training and evaluation. **e), f)** AUROC and AUPRC, respectively, of the model experiment variations: using GW and ICU segments, only using general ward time segments for testing (GW only), and only using ICU segments for testing.

4.6.2 Modality ablations

The impact of each modality on prediction performance was evaluated by ablation of individual modalities, as shown in Figure 4.14. Relative to the base model with all 7,280 features, removing aggregated data, static comorbidities, and text embeddings resulted in minimal performance degradation. ICU data, temporal features, and laboratory values had a marginal impact on model performance. In contrast, removing wearable data had a major impact in terms of AUROC (9%) and AUPRC (50%). We evaluated the contribution of individual data modalities by comparing eight model configurations using the Friedman [Fri40] test, which indicated significant performance differences across modalities ($\chi^2 = 107.28$, $p < 0.0001$), especially for leaving out the *GW wearables*. Furthermore, this showed that leaving out *'Time of day'* and the *'Laboratory tests'*

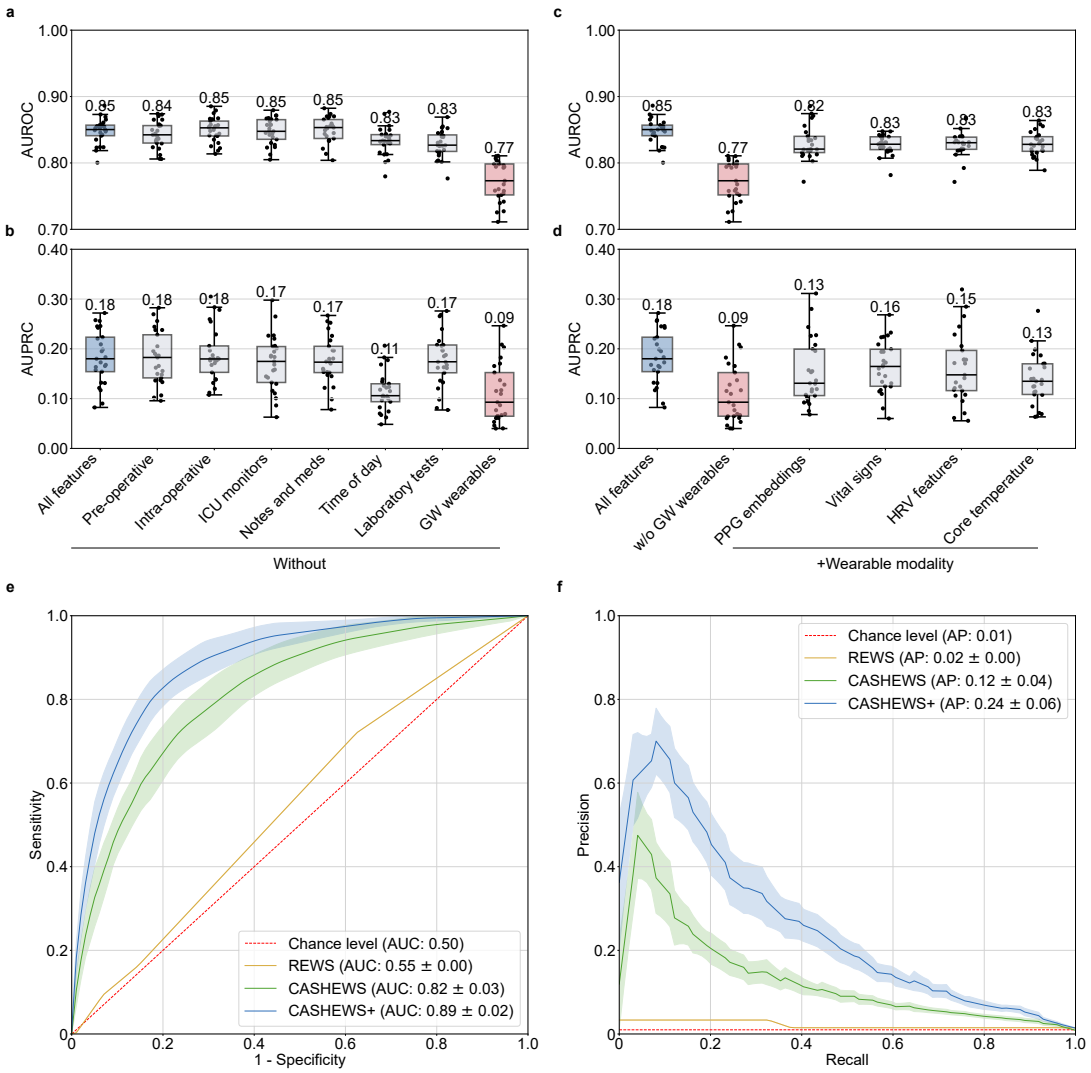


Figure 4.14: Modality ablation and prediction performance curves a) and b) AUROC and AUPRC for the models if we leave out particular modality subsets. c), d) seperating the performance impact of the individual wearable performance subcomponents. e), f) Receiver Operating Curve and Precision Recall Curve comparison; *CASHEWS*: Clinical ASsist Heuristic Early Warning System; *CASHEWS+*: dynamic prediction model with wearable monitoring. Baseline: Post-ICU static endpoint prediction based on ICU data, pre- and intraoperative features.

were different (significantly worse) from the baseline group but not from each other. Complete results can be found in Table D.17.

Moreover, to determine the contribution of individual wearable processing strategies, we performed additional ablations removing each signal-processing approach in all general-ward patients. Here, we observe that while adding any of the wearable modalities has an impact with varying interquartile ranges, the most significant aspect appears to be having at least one wearable modality. Moreover, the Precision-Recall and Receiver Operating Characteristic curves demonstrate the precise performance improvement of the wearable data: *CASHEWS+* over *CASHEWS*. The REWS [van+23] rule-based baseline performs especially poorly compared to our ML-driven system.

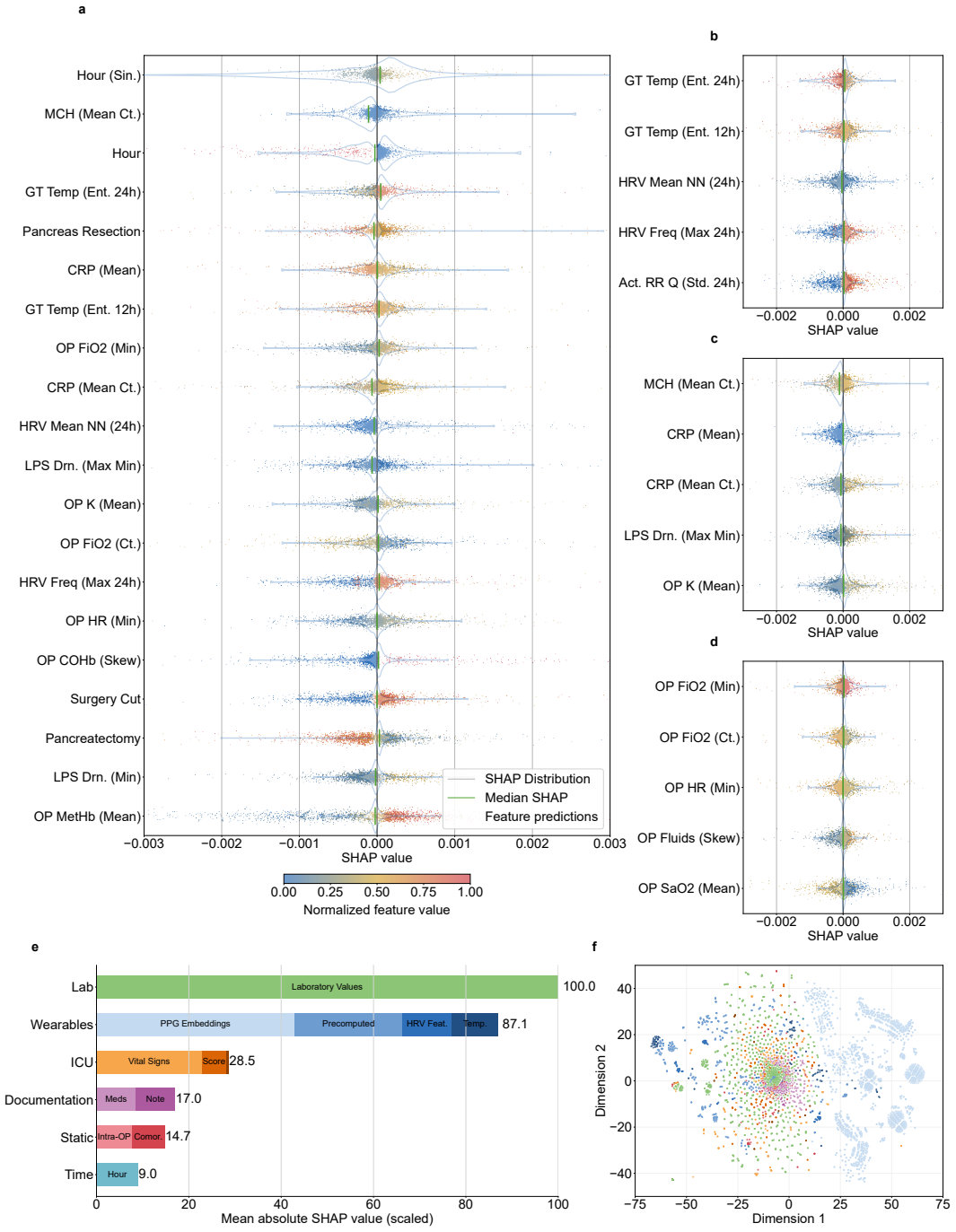


Figure 4.15: Explaining model feature contributions. a) A ranking of the top 20 features of the model. Each row shows a feature's SHAP value distribution as a violin plot (y-axis) against SHAP values (x-axis). Individual data points, consisting of 2,000 randomly sampled samples, are represented as colored dots, where the color intensity reflects the percentile rank of the feature value in the dataset, demonstrating the magnitude of the feature's impact and the relative values of the features. **b)** A ranking of the top 5 wearable features. **c)** A ranking of the top 5 lab features. **d)** A ranking of the top 5 ICU features. **e)**The aggregated feature importance by features in each group, scaled by the highest importance group. **f)** Dimensionality reduction (t-SNE) [vdMH08] plotted for each averaged feature value group shown in e).

4.6.3 Explanation analysis

The single feature weights of the complete model, as well as the most important modalities according to SHapley Additive exPlanations (SHAP) [LL17a] explainability analysis, are shown in Figure 4.15. Body-temperature entropy emerged as the most impactful individual physiologic feature, followed by respiration entropy and mean HRV (Figure 4.15b). Among laboratory results, Mean Corpuscular Hemoglobin, C-reactive protein, and Lipase were the most important (Figure 4.15c). Within ICU data, the fraction of inspired oxygen ranked highest (Figure 4.15d). Figure 4.15e shows that laboratory features contribute the largest SHAP values overall, closely followed by wearables. Finally, with the t-SNE [vdMH08] visualization of the full dataset as the mean of all feature values (Figure 4.15f), we demonstrate the (diss)-similarity of the modalities and individual features. The separate clusters suggest that wearable features and lab features contribute significantly to the diversity of the feature space.

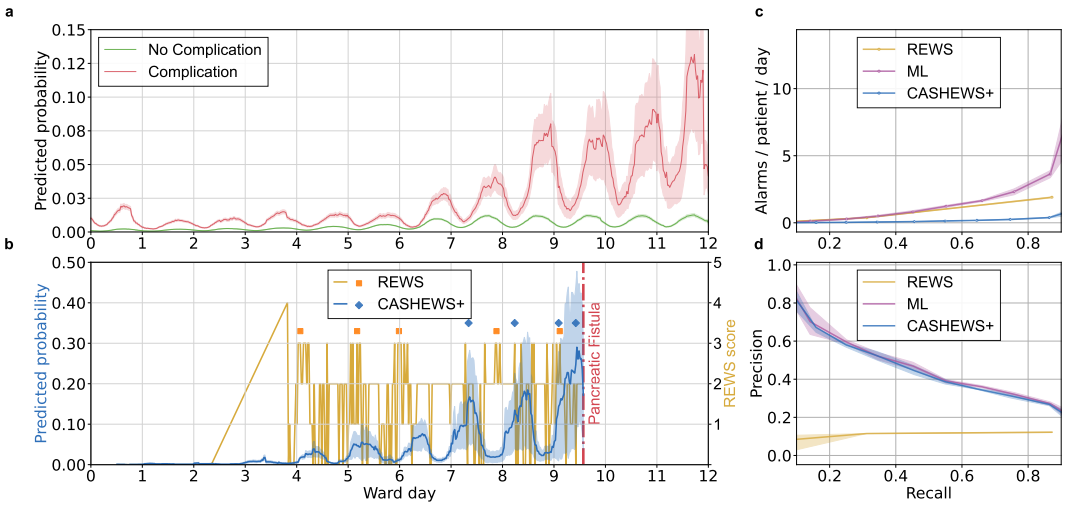


Figure 4.16: The implementation of *CASHEWS+*, a multimodal wearable-enabled early warning system for surgical wards. a) ML predictions stratified by the complication and non-complication groups. **b)** An example patient stay from the first ward day to the end of their stay in the ward. REWS: Remote Early Warning Score [van+23]. *CASHEWS+*: Clinical ASsist Heuristic Early Warning System + Wearables. Individual points indicate alarms for the respective strategy. We chose the highest precision in the same recall bin that still produced at least one alarm. **c)** Alarm burden (alarms per patient per day) for the raw ML predictions with static threshold, the alarm model with blackout periods, a dynamic threshold, and a REWS-based dynamic alarm system with the same alarm strategy. **d)** The precision-recall curve computed with the approach described in Gao et al. [GYH22] for the aforementioned alarming strategies.

4.6.4 Evaluating an ML-driven alarming strategy

To translate *CASHEWS+* predictions into actionable clinical alerts, we developed an alarm strategy addressing the critical challenge of alarm fatigue, for example, in the well-monitored ICU, where conventional threshold-based systems generate 100-1,000 alarms per patient per day, with up to 99.5% being non-actionable [Ngu+25]. We prioritized three desirable properties: (1) low alarm burden operational within nursing capacity [Pat+20], (2) similar precision at the same level of recall to maintain clinician trust, and (3) sufficient lead time for intervention, while not alarming before any

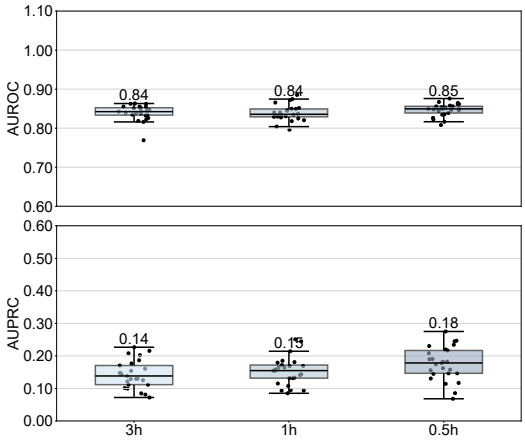


Figure 4.17: Ablation experiments for segment length. This figure shows the AUROC/AUPRC results of adjusting the segment length, i.e., how long the time windows are that we use to aggregate data. We see that a segment length of 30 minutes performs best, though the differences are small.

complication could have developed [Yad+24]. The probabilities output by the model are the source for the alarm system. As the basis of the system, we considered an 8h shift. We employed deactivating the alarm after a false positive alarm (*silencing*) for 8 hours [Hyl20]. The crossing of a threshold of alarms required in a sliding time window, known as *persistence*, was set to 3 times in the last 6 hours. The exact hyperparameters and methodology are described in the Methods section. The performance of the different alarm strategies was evaluated using a 24-hour prediction horizon, as we regarded this as an appropriate lead time. We employed *precision*, *recall*, *and the number of alarms per patient per day* as evaluation metrics. Figure 4.16 demonstrates that the predicted probability for patients in the complication group generally increased to higher risk probabilities later in the progression of the stay. Moreover, we compare a sample patient with the REWS-based alarm and demonstrate that the alarms per patient per day (alarm burden), reduced significantly, while maintaining a similar precision at each level of recall, e.g., if one desires an alarm burden of 2 because of operational capacity, we obtain a recall of 0.9 and a precision of 0.20.

4.6.5 Temporal adjustments

We investigated the impact of segment length, i.e., the amount of time aggregated into one time window. The results are found in Figure 4.17. We see that a segment length of 30 minutes has a slight advantage in terms of IQR, though the differences are slight. A longer segment length may be preferred in a clinical setting with limited resources, as it would reduce the number of predictions; this, in turn, could lower both the alarm burden and the computational burden. However, a shorter segment length could be preferred if one wants to detect deterioration as soon as possible and resources are less constrained.

4.7 Discussion

Postoperative complications have been termed a "hidden pandemic," [Lud22] reflecting their prevalence, lasting clinical impact, and the substantial economic strain on hospitals and health care systems. The CASSANDRA study introduces the concept of an AI-supported ward to identify and predict postoperative deep organ-space infections through continuous, real-time analysis of multimodal patient data. To our knowledge, this is the first investigation to develop an AI-based monitoring system for general surgical wards, integrating data with high-resolution telemetric vital sign acquisition for predictive analytics. Our findings demonstrate the feasibility of digital patient representations to enable the early detection of postoperative complications, as evidenced by a model AUROC of 0.89. Moreover, prediction horizons of 6 and even 12 hours ahead of the current standard of care achieves a reasonable AUROC of 0.90 and 0.86, respectively. These results suggest that multimodal, real-time models could enable AI-supported wards, support the timely detection of complications, and facilitate semi-autonomous clinical workflows.

4.7.1 Wearables in the General Ward

Previous work on AI-based adverse event detection has primarily been confined to intensive care units, where well-established continuous monitoring systems and the availability of large datasets facilitated retrospective analysis [Hyl20; Mey+18; Van+24a]. Most of these studies have targeted advanced events, such as circulatory failure [Hyl20] or sepsis [Nem+18], which typically represent only the terminal stage of a complication cascade, affecting fewer than 2% of patients. Moreover, there have been earlier attempts at predicting surgical complications [vBoe+24; Shi+23; Smi+22]. However, they usually integrate only a limited amount of modalities extracted from previously available EHR data and patient data collection systems.

In contrast, the majority of postoperative complications occur on regular wards, with early infectious events such as deep organ space infections affecting up to 20% of patients and prolonging hospitalization by approximately 16 days [Hea+24]. Redirecting predictive monitoring to regular wards, therefore, captures more frequent and actionable targets for enhancing perioperative safety. In this context, the implementation of telemedical sensors for continuous vital sign monitoring holds particular promise for both reducing nursing workload and enabling elaborate patient surveillance [Bec+23; Lee+22c; Møl+25; Wee+19].

However, previous studies on monitoring strategies in this emerging field have predominantly evaluated predefined vital sign thresholds, simple rule-based heuristics, or existing early warning scores [Haa+22; Ver+19; WFH18]. Moreover, these works mainly employed wearable devices as standard monitoring equipment, which limits their usability. For example, Zahradka et al. [Zah+22] notes that information overload is likely to occur for sensitive, non-ML-powered alarms. Moreover, Posthuma et al. [Pos+24] noted no significant difference between a cohort with wearables and without in normal general wards. van Rossum et al. [vRos+23] describes the development of a sensor-based EWS but, like most studies, does not employ an ML framework that can scale to multiple modalities.

By contrast, we investigated a variety of sensor-derived signal analysis methods, including precomputed features from manufacturer-processed activity data, heart rate variability, and entropy features derived from raw PPG signals, and deep embeddings from PPG waveforms. The optimal wearable configuration improved model performance by 50% in terms of AUPRC, indicating that high-resolution sensor data is one of the most important modalities. The optimal setup was achieved with hybrid signal processing, which combined precomputed features with raw-signal embeddings, suggesting that high-dimensional signal structures reflect complex physiological interactions that extend beyond the scope of single-parameter monitoring. This finding is consistent with prior work demonstrating that PPG dynamics facilitate early sepsis detection, likely through the capture of microcirculatory alterations [Lom+22; Pie+95; Ter+25] Moreover, HRV, temperature entropy, and respiration entropy were identified as the most substantial single feature contributors. In accordance, HRV has previously been linked to disease severity and prognosis in sepsis-associated states, as it reflects autonomic stress and neurocardiovascular dysregulation — both processes that typically precede clinical deterioration. Consistently, entropy measures reflect the "loss of physiological complexity" as an early marker of impending decompensation. These findings underscore that dynamic complexity encodes early signs of physiologic deterioration, making them usable for high-dimensional models. Embeddings derived from clinical documentation, including physicians' and nurses' free-text notes on examinations and bedside observations, contributed only marginally to model performance for the outcome of interest. To our knowledge, we have assembled the largest surgical dataset, which combines multiple clinical modalities with wearable monitoring. Leenen et al. [Lee+24] has a similar amount of total patients, but this includes controls, and they do not develop a multimodal dataset for ML-driven alarm prediction. We note that our dataset also contains considerable amounts of data points per subject.

4.7.2 Model architecture

The model architecture is a crucial choice for any AI-driven system, as it impacts numerous characteristics of the resulting system. We trained several machine learning model architectures using the engineered segments to ensure optimal performance in our predictive setup: Deep Learning (DL) (Transformer [Vas+17], GRU [Cho+14a], TCN [BKK18]), Gradient Boosted Trees (GBT) (XGBoost [CG16], LightGBM [Ke+17], CatBoost [Pro+19]), Random Forest, and Logistic Regression. GBTs are theoretically at a disadvantage as they do not inherently capture temporal information, like DL architectures such as Transformers or RNNs. While DL is a popular technique for handling massive amounts of unstructured sequential and time-series data, we found that this approach did not outperform GBT models in our experiments. This is likely due to our robust, clinician-guided feature engineering, which performs historical aggregation and statistical extraction. GBTs are an effective feature selector [Xu+14] that complements our multi-model, feature-rich data extraction approach. Moreover, recent ML literature [Che+24; GOV22b; McE+23] has cast doubt on the superiority of DL and Large Language Models (LLMs) [Bro+20] in many clinical prediction scenarios. Furthermore, using GBTs as prediction models has a key advantage: our setup runs on commodity hardware generally available in protected hospital environments, whereas DL is significantly more resource- and energy-intensive, as it requires the use of graph-

ics processing unit (GPU)s. Finally, model explanations are easier to extract than with deep learning, which is often a key evaluation criterion for certifying clinical models [Lun+20].

4.7.3 Translating Models to Alarms

Transforming predictive models into actionable alarm systems requires careful balancing of clinical utility with practical usability, which translates into a reasonable trade-off between sensitivity and the number of false alerts. Alarm fatigue is a well-known phenomenon in the ICU [Hu19]. Our alarm strategies are based on predefined probability thresholds, temporal trends, and the frequency with which these criteria were met within a defined time window. These findings are consistent with everyday clinical practice, as our chosen endpoint is characterized by a slow progression and subtle symptoms that may develop over an extended period. Nevertheless, patients profit from early detection and countermeasures. However, since our model can trigger alarms up to 48 hours in advance, further investigation is needed to determine the extent to which these early alarms can be verified through diagnostic measures. Complications detected this early may not yet be evident in, for example, a Computed Tomography (CT) scan. Our results further demonstrated that our ML models, combined with an optimized alarm strategy, substantially outperformed REWS. These findings suggest that structured, AI-generated alarms can detect subtle yet clinically significant deterioration early, while maintaining alarm frequency at a manageable level for clinical staff. An appropriate recall bin, number needed to evaluate, and prediction window need to be chosen by balancing the risk of excessive false alarms against the need to avoid missing complications, while also considering the available resources on a general ward.

4.7.4 Limitations

Generalizability This study has several limitations that may affect the generalizability of our findings. Despite being the largest reported cohort in surgical research using continuous vital sign monitoring, larger cohorts could enable more accurate and robust models. Moreover, as a single-center study conducted within our institution's Enhanced Recovery After Surgery (ERAS) program [JJL18], the results reflect the implementation of strict, standardized postoperative procedures and an institutional culture with exceptionally low failure-to-rescue rates. While these factors contribute to optimal patient outcomes, they may limit the applicability of our findings to centers with different postoperative care standards or complication management approaches. Additionally, the standardized nature of our protocols, though beneficial for internal validity, may not reflect the variability in clinical practice encountered across different healthcare settings.

Data Quality As Downey et al. [Dow+19] noted, less intrusive monitoring systems often suffer regarding the quality of individual data points (reliability). During the development of our framework, we noticed that incorrectly collected wearable data

can severely affect the accuracy of the model and EWS. Therefore, it is of utmost concern to ensure that wearables are securely attached and there is no interference in the monitored ward when deploying a wearable-based ML system. We suggest that clinical personnel benchmark the output of these devices against more robust equipment. Moreover, we expect newer devices to be equipped with better sensors. Additionally, using real-time, quality monitoring could enable a temporary switch to more robust, if less accurate, models. It is important to incorporate process improvements such as an increased capacity of fully-charged wearables, which ease the process of wearing wearable devices can lead to higher-quality sensor data and minimal loss of data.

Model Implementation and Updating Finally, clinical ML-specific problems, such as data drift [Rot+23], necessitate a continuously updated and validated EWS. We hope to see improvements in MLOps [Kri+22] and model evaluation processes [Fen+25] that address these needs; Chapter 5 will partially address this crucial efficiency improvement.

4.7.5 Future Work

Despite the complexities of different modalities collected and stored in legacy systems or through modern wearable devices, our work demonstrates that a multimodal framework can be developed to alert clinicians of post-surgical complications. We recognize a larger, coordinated study like van der Meijden et al. [van+25b], which includes a broad range of features and wearable data in a continuous setting, as future work. The PPG signal particularly lends itself well to various feature extraction techniques [Alm+22]. Moreover, more modalities can be included. Including embeddings from imaging modalities was considered, but the data for our cohort is scarce, especially considering that they are often collected after our primary endpoints have occurred for most patients.

We achieve this using a hybrid ward that combines EHR information with continuous monitoring of vital signs; we hypothesize that this approach can benefit ML in various clinical settings. More broadly, medical ML requires robust and specialized data collection efforts to bring more reliable systems to practice. We recommend that ML research focuses on problem modeling and the data pipeline — consisting of i.a. data collection, integration, cleaning, and preprocessing — to develop accurate, robust prediction systems. Moreover, medical research should acknowledge the capabilities and limitations of ML-powered alarm systems.

4.8 Conclusion

As in any experimental trial, while our system demonstrates promising technical performance, its true clinical utility can only be definitively established through rigorous evaluation of patient-centered outcomes such as length of stay, comprehensive complication index, readmission rates, and mortality statistics. These clinical endpoints represent the gold standard for assessing healthcare interventions, capturing the real-world impact on patient care quality and institutional efficiency beyond algorithmic

accuracy alone. Future research should prioritize comprehensive clinical trials that measure these outcomes to provide objective evidence of the system's meaningful improvement in healthcare delivery and justify its implementation in clinical practice.

In conclusion, the concept of an AI-supported, semi-autonomous surgical ward directly addresses persisting and even aggravating challenges in current postoperative care. Today, many general wards rely on intermittent bedside assessments, often performed only a few times per day, which makes the early recognition of subtle complications highly unreliable. This problem is exacerbated by chronic staffing shortages, increasing patient volumes, and the growing complexity of surgical populations, which together limit the capacity of nurses and physicians to provide continuous surveillance [Pag+24]. As a result, many complications are detected late, at a stage when interventions are less effective and associated with higher morbidity, costs, and mortality. A semi-autonomous ward equipped with continuous telemonitoring and AI-assisted analytics can provide uninterrupted surveillance, prioritize patients at highest risk, and issue manageable alerts that reduce alarm fatigue. Such systems not only enhance patient safety through earlier interventions but also alleviate workload, standardize surveillance independent of staffing variations, and optimize resource use. In doing so, they offer a scalable solution to bridge the gap between the demand for surgical care and the limited availability of human resources, marking a necessary evolution in ward organization in modern healthcare. Despite the promising technical performance of multimodal continuous models, their utility can only be conclusively established through the evaluation of patient-centered outcomes: length of stay, comprehensive complication index, readmission rates, and mortality statistics. Future research is key to providing objective evidence of the system's meaningful improvement in healthcare delivery and justifying its implementation in clinical practice.

As AI in medicine continues to scale up to create foundation models from retrospective data [Wax+25], we see evidence that prospectively changing hospital procedures to collect data to combine with existing data collection leads to real performance improvements, proving that one does not just need any data to create predictive systems; one needs the right data at the right moment.

5

Building a Modular Tooling Ecosystem for High-capacity Collaborative Clinical ML

This chapter is based on my interpretation and development of the work from the MEDS working group [Gro24; McD+24; McD+25d], of which I am a core member. My contributions include participation in discussions, ideation, implementation, revision, writing, and presentation. We also discuss the context of MEDS, which is currently under submission as a Datasets, Benchmarks, and Protocols paper at New England Journal of Medicine AI (NEJM AI) [McD+25c]. Moreover, this chapter contains original work describing the tooling ecosystem that I and others have built around MEDS. As a proof of concept, the MEDS ecosystem is used to train and benchmark health AI models across datasets and health systems, showcasing its potential for building clinical AI infrastructure at scale [Van25c]. Another work is in preparation on the practical use of EHR foundation models within the Mount Sinai Health System [Van+25a], which will expand on the results presented in this chapter. Here, 'I' refers to my personal contributions, while 'we' refers to the collective efforts of the MEDS working group and collaborators as this chapter is not based on a single published work but rather a synthesis of multiple (collaborative or personal) contributions.

This thesis has so far discussed various challenges in clinical machine learning, including benchmarking and developing intensive care unit (ICU) prediction models (Chapter 3) and multimodal data modeling in a concrete setting (Chapter 4). Addressing these challenges is crucial for advancing the field and ensuring that machine learning models can be effectively and safely deployed in clinical settings. While we transitioned from a relatively general challenge to a specific one, we now return to a more general challenge that underpins much of the work in this thesis: establishing the feasibility of clinical machine learning.

5.1 Motivation

A major challenge in machine learning for healthcare (ML4H) is ensuring reproducibility, transparency, and, by extension, correctness. Datasets and code are often not published, making it difficult to replicate cohort definitions, preprocessing pipelines, and training setups. *Yet Another ICU Benchmark* (*YAIB*) [Van+24a], a modular framework designed to support reproducible clinical ML experiments with ICU datasets, addresses this issue as shown in Chapter 3. However, outside of controlled and curated ICU Electronic Health Record (EHR) data environment, there has been little progress. At the same time, high-capacity foundation models FMs, designed to be adaptable to a wide range of tasks and domains, are emerging [Wor+23b; Kha+25] and have shown promising results [Pan+25a].

The result is that valuable ML4H knowledge, already gained by individuals in the community, is left unused. This impedes the development of AI-enabled tools and can lead to the deployment of faulty models, potentially compromising the quality of life. For ML and AI tools to be correctly deployed at scale, we need a comprehensive ecosystem of tooling and methodology. This should be developed and maintained by the scientific community and be scalable to entire health systems without compromising privacy.

In this chapter, we discuss an approach to solve these shortcomings: the Medical Event Data Standard (MEDS) [Gro24; McD+25b]. It is a flexible, low-level ML-first format that integrates with EHR processing and modeling tools. MEDS accelerates the training of predictive models on clinical tasks as an event-based standard with an expanding array of tools. We demonstrate several developed tools and show their viability in real-world ML applications.

Several personal contributions to support ML4H research and the development of the MEDS effort have been made: firstly, I contributed to the development of the MEDS format itself by sharing the accumulated knowledge to shape the schema. Secondly, I developed Extract, Transform, Load (ETL) and contributed to tools to convert EHR datasets of formats into MEDS: I spearheaded the conversion of five datasets into MEDS [van25c; van25b; van25e; Van25d] and for 2 data standards (Observational Medical Outcomes Partnership (OMOP) [van25d] and Fast Healthcare Interoperability Resources (FHIR) [Van25a]), and contributed to three others [vZM25; VRR25; MX25]. Moreover, I developed a tool for interactive data exploration in MEDS (MEDS-Inspect [Van25e]) and code harmonization (HarmonizEHR [vdVan25]). I contributed to a decentralized training framework [McD+24] that I extend with 9 new datasets in this work. Finally, I have trained and benchmarked multiple FMs on a multicenter health system dataset for 8 different tasks in different data regimes, evaluated dimensionality reduction, compared several aspects to other large datasets (Columbia University Medical Center (CUMC) and Stanford Medicine), and evaluated how they compare to natural language processing (NLP) [Van+25a] (unpublished, work in progress).

5.2 Background and Related Work

We distinguish related work in terms of data standards in health AI, libraries and tooling, open-access datasets, and foundation model (FM)s in healthcare. The next sections provide an overview of related work in these areas.

5.2.1 Data Standards in Health AI

Reproducible science in health AI has long been a challenge [Sem+25]. Individual works have tried to address this challenge through data harmonization and standardization efforts [Aro+23]. An important development is the Observational Health Data Sciences and Informatics (OHDSI) organization and its Observational Medical Outcomes Partnership (OMOP) format, as well as its connection to Common Data Model (CDM). Moreover, Patient-Centered Outcomes Research Network (PCORnet) is another large-scale initiative to standardize healthcare data for research purposes [Col+14]. Informatics for Integrating Biology and the Bedside (I2B2) is a widely used clinical data warehousing and analytics research platform [Mur+10]. HL7 Fast Healthcare Interoperability Resources (FHIR) is a standard for exchanging healthcare information electronically [BS13]. The MIMIC format is not necessarily a format but is often used for critical care research [Joh+16], for example, in the Northwestern ICU Database (NWICU) [Mou+24]. More broadly, Arora et al. [Aro+23] argue that the lack of standardization, difficulty in harmonizing and several methods of data capture and data-governance restrictions are significant barriers that can be addressed with the right standards.

Despite the large number of standards, there is still considerable friction in applying these standards to health AI research [Bor+23; Per+23; Aro+23; McD25; Sem+25]. This has been shown in practical [Bor+23; Per+23] and research settings [McD25; Sem+25]. As mentioned in Chapter 2, most data is Real-world Data (RWD) collected for diagnostic, billing, and administrative purposes, which may not always align with the needs of health AI. Partially this is due to the different types and goals of the collected data and health standardization efforts. For example, OMOP is primarily designed for observational research, while FHIR focuses on interoperability and data exchange. If we do narrow our scope to health AI, we note that Semmelrock et al. [Sem+25] identify four main challenges to reproducibility: 1) missing and vague methodological details, 2) inaccessible code, 3) limited access to data, 4) inherent non-determinism. Standards need to address these challenges to be effective in improving reproducibility in health AI research. However, it is still up to the community to adopt these standards and use them properly so they can tackle the challenges mentioned above. According to Donoho [Don24], the lack of 'frictionless reproducibility' is a major bottleneck in data science, which is further corroborated by McDermott [McD25]. This leaves a the gap that MEDS aims to fill by being an open-source, accessible data standard designed for health AI research. The next sections will discuss more related work to further contextualize MEDS where appropriate.

5.2.2 Libraries and Tooling

Data standards live or die by effective and efficient tooling aimed at the target audience of the standard. As discussed in Chapter 3, tooling often also introduces its own data formats. Harmonization efforts in Electronic Health Record (EHR) data described above partially overlap with the selection of libraries and tooling. Here, we will give a short summary of some relevant libraries and tools available at the time of writing.

Several data harmonization and benchmarking tools have been developed, mainly for more readily available ICU datasets. Wang et al. [Wan+20b] introduced a preprocessing pipeline for MIMIC-III, but it is limited to this dataset and lacks support for newer datasets like MIMIC-IV. Similarly, Adibuzzaman et al. [Adi+16] and Gupta et al. [Gup+22] focus on specific datasets without enabling interoperability or backward compatibility. Mandyam et al. [Man+21] provides tools for MIMIC-IV but lacks ML-benchmarking capabilities. Tang et al. [Tan+20] supports both eICU and MIMIC-III but does not include many modern model architectures. PyHealth [Zha+21] offers a library for model development but lacks dataset interoperability, imputation methods, and a customizable preprocessing pipeline. Finally, Jarrett et al. [Jar+21] provides an AutoML pipeline but does not offer an end-to-end benchmarking solution or extensible support for open-access datasets.

More tooling for research is emerging, such as the Common Longitudinal Intensive Care Unit data Format (CLIF) [Roj+24], which focuses on federated learning with intensive care unit (ICU) data. Moreover, Luo et al. [LL25] proposes a format accompanied by a pipeline that is used for a multi-institutional ICU collection effort in the USA.

Many of these frameworks are tailored to a limited set of ICU datasets with rigid schemas attained by aggressive preprocessing, which can be destructive. In contrast, MEDS is a light and versatile data standard designed to work with any EHR source, whether public or private, and does not require harmonization. This broad applicability makes MEDS more adaptable for general research purposes. Additionally, while these frameworks are often closed ecosystems reliant on Python-package-based interfaces, MEDS adopts a data-as-interface approach with an expanding open-source ecosystem. This design makes MEDS more flexible to researchers' needs and better suited to the evolving AI landscape. MEDS contrasts in this aspect with our own work on *YAIB* discussed in Chapter 3, which is extensible and can integrate with EHR, though for external validation and fine-tuning, it is limited by a rigid, potentially destructive, schema to attain harmonization across datasets; therefore, *YAIB* could provide more functionality at the cost of dataset generalizability and flexibility. We argue that both approaches can coexist and serve different purposes in health AI research.

5.2.3 EHR Coding Systems

Standardizing procedures, diagnoses, medications, and other clinical concepts is essential for observational research and interoperability between healthcare systems. This is typically achieved through the use of controlled vocabularies and coding systems. Common vocabularies include: the International Classification of Diseases (ICD) [Org92], the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) [Don06], Anatomical Therapeutic Chemical Classification System (ATC) [Ber06], Unified Code

for Units of Measure (UCUM) [Sch+99], Medical Subject Headings (MeSH) [oMed00], and the Logical Observation Identifiers Names and Codes (LOINC) [Huf+98].

Moreover, there are coding vocabularies primarily exclusive to the United States of America (USA), which is a major player in healthcare research and available healthcare data: the Current Procedural Terminology (CPT) [Les+16], National Drug Code (NDC), Normalized names for clinical drugs (RxNorm) [Nel+11], and Vaccine Administered Codes (CVX). This chapter describes and uses healthcare system datasets from the USA.

Unified Medical Language System (UMLS) [Bod04] is an effort to unify these vocabularies. It provides a comprehensive compendium of controlled vocabularies in the biomedical sciences, along with a mapping structure that connects these vocabularies. One of the main components of UMLS is the *Metathesaurus*, which contains information about biomedical and health-related concepts, their various names, and the relationships among them. UMLS is widely used in health informatics applications, including natural language processing, information retrieval, and clinical decision support systems. Note that OMOP uses a subset of the above clinical vocabularies as so-called 'standard concepts'.

However, often within a Patient Data Management System (PDMS), local coding systems are used that do not directly map to the above standard vocabularies; this creates challenges for data harmonization and interoperability which are important for research. There might be inherent reasons for this as local coding systems can be optimized rather for the operational functionalities of EHR systems rather than for secondary uses, such as health AI. Therefore, mapping local codes to standard vocabularies is often considered a crucial step in preparing EHR data for research and analysis [Bor+24]. Unified mappings are often not publicly available due to various reasons, including proprietary concerns, privacy issues, and the complexity of healthcare data. As a result, researchers and institutions often have to develop their own mappings or rely on third-party services to facilitate data harmonization [Bas+23]. In this Chapter, we address these some of these challenges with HarmonizEHR [vdVan25], a MEDS-compatible tool that can convert between different coding systems, and demonstrate the need for further work.

5.2.4 Open-access Datasets

Several open-access ICU datasets were discussed in Chapter 2 and 3. We provide a brief overview of the open-access datasets used in this chapter. Moreover, we refer to Table 5.1, which provides an overview of these known datasets with MEDS ETL support, including private datasets that are discussed in the next subsection.

AUMCdb: The Amsterdam University Medical Center Database (AUMCdb) [Tho+21] is the first freely accessible intensive care database from within the European Union containing de-identified health data related to tens of thousands of European intensive care unit admissions, including demographics, vital signs, laboratory tests, and medications.

EHRSHOT: EHRSHOT [Wor+23a] is a benchmark dataset of de-identified structured electronic health record data from Stanford Medicine (i.e., a subset) containing longitudinal health records for 6,739 patients with 41.6 million coded clinical obser-

Table 5.1: Overview of known datasets with MEDS Extract, Transform, Load (ETL) support. This includes both publicly available datasets and private datasets used within the MEDS working group. The 'My Role' column indicates my level of contribution to the development of the respective ETL pipeline.

Dataset	Stays	Version	Frequency	Origin	Year	License	MEDS ETL	My Role	Full Dataset Name
AUMCdb [Tho+21]	23,000	v1.0.2	up to 1 min	Netherlands	2019	Not specified	GitHub [VRR25]	Main Dev.	Amsterdam University Medical Center Database
eICU [Pol+18]	201,000	v2.0	5 minutes	USA	2017	PhysioNet	GitHub [VZM25]	Contributor	eICU Collaborative Research Database
EHRSHOT [Wor+23a]	6,739	v3.0	heterogeneous	USA	2023	Not specified	Pre-converted and OMOP [van25d]	-	EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models; a subset of the Stanford Medicine health system dataset.
HiRID [Hyl20]	34,000	v1.1.1	2/5 minutes	Switzerland	2020	PhysioNet	GitHub [Van25d]	Main Dev.	High-Resolution ICU Dataset
INSPIRE [Lim+24]	130,000	v1.2	Not specified	South Korea	2024	Korea Health Data	GitHub [van25b]	Main Dev.	INformative Surgical Patient dataset for Innovative Research Environment
MIMIC-IV [Joh+22]	73,000	v3.1	~1 hour	USA	2020	PhysioNet	GitHub	Contributor	Medical Information Mart for Intensive Care IV
NWICU [Mou+24]	25,000	v0.1.0	Not specified	USA	2023	PhysioNet	GitHub [van25c]	Main Dev.	Northwestern ICU Database
SICdb [Rod+23a]	27,350	v1.0.8	1 minute	Austria	2024	PhysioNet	GitHub [van25e]	Main Dev.	Salzburg Intensive Care Database
MSHS*	6.8M	OMOP 5.3	heterogeneous	USA	2025	Not specified	GitHub [van25d]	Main Dev.	Mount Sinai Health System
CUMC *	6.7M	OMOP 5.4	heterogeneous	USA	2025	Not specified	OMOP ETLs	-	Columbia University Medical Center
SM*	2.57M	OMOP 5.4	heterogeneous	USA	2023	Not specified	OMOP ETLs	-	Stanford Medicine
OMOP† [Vos+15]	-	5.x	heterogeneous	various	various	Various	GitHub [van25d]	Main Dev.	Observational Medical Outcomes Partnership (OMOP) datasets
FHIR† [BS13]	-	v4/v5	heterogeneous	-	-	Not specified	GitHub [van25a]	Main Dev.	Fast Healthcare Interoperability Resources (FHIR) datasets

* These datasets are not publicly available but are internally used with MEDS; my MEDS ETL pipeline [van25d] should facilitate their conversion as well. The Mount Sinai dataset updates continuously from ETL pipelines that feed into it from the Mount Sinai Health System. Metadata timestamps are provided. The ETL pipeline is designed to work with any Observational Medical Outcomes Partnership (OMOP)[Vos+15] 5.x dataset.
† These are generic ETL pipelines that can convert any dataset in the respective format to MEDS format. The OMOP pipeline supports any OMOP 5.x dataset, while the FHIR pipeline supports datasets in FHIR v4 and v5 formats.

vations and 921,499 visits, designed for few-shot evaluation of FMs on EHR data and available in Original, MEDS, and OMOP formats.

eICU: The eICU Collaborative Research Database (eICU) [Pol+18] is a multicenter database comprising de-identified health data associated with over 200,000 admissions to ICUs across the United States between 2014-2015, including vital sign measurements, care plan documentation, severity of illness measures, diagnosis information, and treatment information collected through the Philips eICU telehealth program.

HiRID: The High Time Resolution ICU Dataset (HiRID) [Hyl20; Yèc+22] is a freely accessible critical care dataset containing data relating to more than 33 thousand patient admissions to the Department of Intensive Care Medicine of the Bern University Hospital, Switzerland, developed in cooperation between ETH Zürich and the ICU, containing 712 routinely collected physiological variables, diagnostic test results and treatment parameters from January 2008 to June 2016 with a uniquely high time resolution of one entry every two minutes.

INSPIRE: The INformative Surgical Patient dataset for Innovative Research Environment (INSPIRE) is a publicly available perioperative medicine dataset including

approximately 130,000 surgical cases who underwent anesthesia at an academic institution in South Korea between 2011 and 2020, with comprehensive patient characteristics, vital signs from operating theatre, general wards, and ICUs, laboratory results from six months before admission to six months after discharge, and medication during hospitalization.

MIMIC-IV: The Medical Information Mart for Intensive Care IV (MIMIC-IV) [Joh+22] is a large, freely available database comprising de-identified health data associated with over 383,000 admissions to critical care units at the Beth Israel Deaconess Medical Center between 2008 and 2019, including detailed demographics, vital signs, laboratory tests, medications, and clinical notes.

NWICU: The Northwestern ICU Database (NWICU) [Mou+24] originates from Northwestern Medicine, a network of twelve hospitals in Chicago that migrated to the Epic EMR platform in 2018, with data made available through a relational Enterprise Data Warehouse stored on encrypted file servers.

SICdb: The Salzburg Intensive Care Database (SICdb) [Rod+23a] offers insights into over 27 thousand intensive care admissions, including therapies and data on preceding surgeries, collected between 2013 and 2021 from four different ICUs at the University Hospital Salzburg, containing de-identified case information, vital signs, laboratory results, and medication data with both aggregated once-per-hour and highly granular once-per-minute data.

5.2.5 Foundation Models in Healthcare

Foundation models [Bom+21] (FM) have shown promise in many applications and have drastically changed the field of natural language processing (NLP) and computer vision, and are increasingly being applied in many settings within computer science. Transformers [Vas+17] and the Bidirectional Encoder Representations from Transformers (BERT) architecture [Dev+19] have laid the groundwork for this. Then, Generative Pretrained Transformer (GPT)-3 [Bro+20] and its implementation in ChatGPT have demonstrated the power of LLMs in generating human-like text. These models leverage large-scale pretraining on diverse datasets, enabling them to generalize well to specific tasks with limited labeled data. Here, we address FMs specifically for EHR data. An overview of existing models is given in Table 5.2.

Within EHR research, the idea of foundation models (FMs) has been addressed in several studies and architectures even before the term was coined, specifically in the form of generative pre-trained architectures [Raj+18; Ste+21]. Before these developments, the field of clinical NLP dealt with similar issues.

Datasets A variety of datasets have been used to train FMs for EHR data. The MIMIC datasets (MIMIC-III [Joh+16], MIMIC-IV [Joh+22]) and eICU [Pol+18] are popular choices and can be accessed by all credentialed researchers, making these works, in theory, more reproducible. Other works use internal medical center datasets,"the Columbia University Medical Center (CUMC) Presbyterian [Pan+21; Pan+24; Pan+25a] (5 million patients), Stanford Medicine [Cal+23] (2.7 million patients). The Veterans Affairs

Table 5.2: Overview of foundation models and FM benchmarks for EHR data. We summarize the datasets that were used for evaluation, tokenization, and what type of pretraining loss they employed to train the models. The upper part of the table describes models whereas the bottom describes benchmarks. Adapted from Pang et al. [Pan+25a], used under CC BY 4.0 license, and expanded with more models and benchmarks.

Model	Dataset	Temporal Repr.	Tokenization	Pretrain Loss
BEHRT [Li+20]	Clinical Practice Research Datalink [Her+15]	PE & Age Embedding	Code-based	MLM
Med-BERT [Ras+21a]	Cerner Health Facts and Truven Health MarketScan	PE & Age Embedding	Code-based	MLM
CLMBR [Ste+21]	Stanford Medicine [Cal+23]	RoPE	Code-based	NTP
MedGPT [Kra+21]	London Hospital + MIMIC-III	RoPE	Text & Code based	NTP
CEHR-BERT [Pan+21]	CUMC-NYP	Time2vec +ATT	Code-based	MLM
GatorTron [Yan+22]	UF Health, Pubmed, Wikipedia + MIMIC III	None	Text-based	MLM & SOP
GenHPF [Hur+24]	MIMIC-III, IV + eICU	ALiBi	Text-based	SSL
Foresight [Kra+24]	London Hospital + MIMIC-III	Sinusoidal	Text & Code based	NTP
MOTOR [Ste+23]	Stanford Medicine + MERATIVE[18]	RoPE	Code-based	TTE
TransformEHR [Yan+23b]	VA Bedford	Sinusoidal	Code-based	Visit MLM
EBCL [Ouf+24b]	MIMIC-IV [Joh+22]	Custom	Code-based	EBCL
CEHR-GPT [Pan+24]	CUMC-NYP	ATT	Code-based	NTP
CORE-BEHRT [Odg+24]	Capital Region of Denmark	RoPE	Code-based	MLM
EHRMAMBA [Fal+24]	MIMIC-IV	Time2vec	Code-based	NTP
ETHOS [Ren+24]	MIMIC-IV	Age quantiles + ATT	Code-based	NTP
CEHR-XGPT [Pan+25b]	CUMC-NYP	ATT	Code-based	TD + TTE
Benchmark	**Dataset**	**Temporal Repr.**	**Tokenization**	**Pretrain Loss**
Long-context LLAMA, MAMBA [Wor+24]	Stanford Medicine	RoPE	Code-based	NTP
MEDS-DEV [McD+24]	Various*	Various*	Various*	Various*
FoMoH [Pan+25a]	CUMC-NYP	Various*	Code-based	Various*

MLM: Masked language modeling, NTP: Next token prediction, SOP: Sentence order prediction, SSL: Self-supervised learning, TTE: Time to event, EBCL: Event-based Contrastive Learning, ATT: Artificial Time Tokens, TD: Temporal Denoising Objective, RoPE: Rotary Position Embedding
*Benchmarks for foundation models for EHR data have various temporal representations and pretraining objectives.

(VA) healthcare system [Jus+06] is another source of EHR data, specifically the VA Bedford dataset [Yan+23b], which consists of 6.5 million patients. The Clinical Practice Research Datalink (CPRD) [Her+15] is a primary care database comprising approximately 11.2 million patients (as of 2015). Rasmy et al. [Ras+21a] uses the Cerner Health Facts database (68 million unique patients) and the Truven Health MarketScan® Research Database [18], a claims dataset commercial insurance, Medicare supplemental claims, and Medicaid claims for over 170 million individuals. Some models also utilize a combination of datasets to enhance their training data diversity [Kra+21; Kra+24; Ste+23; Yan+22]

Time representation A major difference between NLP FMs and EHRs is the temporal nature of EHR data; a variety of methods have been used to integrate time representation in these models. Positional embeddings (PE) and age embeddings have been used in BEHRT [Li+20] and Med-BERT [Ras+21a]. Rotary Position Embedding (RoPE) [Su+24], which multiple rotation matrices to indicate the order of words or events, in this case, have been employed in the Clinical Language Modeling based Representations (CLMBR) [Ste+21], MedGPT [Kra+21], MOTOR [Ste+23], and Long-context LLAMA, MAMBA [Wor+24] models. Time2vec [Kaz+19] was used in EHRMamba [Fal+24] and CEHRBERT, the latter also uses attention [Vas+17] mechanisms. Sinusoidal embeddings have been applied in Foresight [Kra+24] and TransformEHR [Yan+23b]. Custom time representations have been proposed in EBCL [Ouf+24b], while ETHOS [Ren+24] uses age quantiles and attention mechanisms.

Tokenization Most models use code-based tokenization, where each medical code (e.g., diagnosis, procedure, medication) is treated as a unique token [Li+20; Ras+21a; Pan+21; Ste+21; Kra+24; Ste+23; Yan+23b; Ouf+24b; Pan+24; Odg+24; Fal+24; Ren+24; Wor+24]. Some models incorporate text-based tokenization, processing unstructured clinical notes alongside structured codes [Kra+21; Yan+22; Kra+24].

Pretraining objectives Pretraining objectives are needed needed to be able to train FMs on large amounts of data in an unsupervised manner. Many models use Masked language modeling (MLM) objectives with Bidirectional Encoder Representations from Transformers (BERT) as pretraining tasks [Li+20; Ras+21a; Pan+21; ZLL22]. Moreover, several works have used Next token prediction (NTP) [Kra+21; Pan+24; Wor+24], known from the GPT style models [Bro+20]. Event-based Contrastive Learning (EBCL) has also been proposed as a pretraining objective [Ouf+24b] which leads to improved clustering performance. Finally, Time to event (TTE) has been used as a pretraining objective in the Many Outcome Time Oriented Representations (MOTOR) model [Ste+23].

Benchmarks Several benchmarks have been proposed to evaluate the performance of FMs on EHR data. FoMoH [Pan+25a] is a comprehensive benchmark that assesses models across multiple clinical tasks, including mortality prediction, readmission prediction, length of stay prediction, and phenotyping. MEDS-DEV [McD+24] provides a decentralized external validation framework for benchmarking models on private datasets using the MEDS format. Moreover, [Wor+24] can be seen as a benchmark for long-context LLM architectures on EHR data. [Zha+25a] explores scaling laws for FMs in healthcare, providing insights into model performance as a function of size and data.

Model size and architecture Model sizes are currently relatively small compared to those of large language models in NLP, often in the range of 100M parameters, primarily due to the limited availability of hardware. However, this has started to change with companies producing larger models [Wax+25]. The architectures used are predominantly transformer-based [Vas+17] and often adapted from existing NLP models such as BERT [Dev+19] and GPT [Bro+20] (2); however, newer architectures, such as

Mamba [DG24] are starting to be used as well [Fal+24; Wor+24]. Moreover, Waxler et al. [Wax+25] uses the Qwen-2 architecture [Yan+24]. Most pre-trained models are not openly available, i.e., their weights are not public (a notable exception is several Stanford Medicine models [Ste+21; Ste+23; Wor+24]); some did not publish code either at the time of writing [Wax+25]. The open-access models only partially generalize to other datasets after fine-tuning due to code tokenizer mismatch, where the input tokens differ from the original pretraining data, or distribution shifts [Guo+24].

Generalizability is an ultimate goal of all (ML) research; FMs, essentially claim to fulfill this promise, though many specific challenges to healthcare data and clinical implementation remain.

5.2.6 The Medical Event Data Standard (MEDS)

The Medical Event Data Standard (MEDS) is a novel framework designed to facilitate the representation and exchange of medical event data across diverse healthcare systems. By providing a unified, very simple, and flexible schema and set of guidelines, MEDS aims to enhance interoperability of high-capacity machine learning and AI applications in healthcare.

The schema consists of:

- `subject_id` (mandatory): The unique identifier for the subject; this is a 64-bit integer and it should not be null in the data.
- `time` (mandatory): The time of the event. This is a timestamp with microsecond precision. May be null in the data for static measurements.
- `code` (mandatory): The code for the event. This is a string (in an unspecified categorical vocabulary). Should not be null in the data.
- `numeric_value` (optional): The numeric value for the event. A 32-bit float. May be null in the data for measurements lacking a numeric value.
- `text_value` (optional): The text value for the event. A string. This is often used for both (a) measurements that have free-text descriptions, such as unstructured laboratory test results, and (b) free-text clinical notes.
- An arbitrary number of additional columns is allowed.

The data is sharded by `subject_id` into multiple parquet files. Each file contains data for a subset of patients. This enables efficient data loading and processing, as well as the easy distribution of data across multiple machines. Additional schemas are currently specified for code metadata, dataset metadata, labels, and subject splits (train/validation or 'tuning'/test or 'held-out').

5.2.7 Motivation

While Fast Healthcare Interoperability Resources (FHIR), Observational Medical Outcomes Partnership (OMOP), Informatics for Integrating Biology and the Bedside (I2B2),

```python
from pyarrow.schema import Schema as PyArrowSchema
import pyarrow as pa
class DataSchema(PyArrowSchema):
    subject_id: Required(pa.int64(), nullable=False)
    time: Required(pa.timestamp("us"), nullable=True)
    code: Required(pa.string(), nullable=False)
    numeric_value: Optional(pa.float32())
    text_value: Optional(pa.large_string())
```

Listing 5.1: The core MEDS schema as of version 0.4.0

and Patient-Centered Outcomes Research Network (PCORnet) prioritize interoperability and harmonized data exchange, they do not natively provide the structures and metadata that AI workflows require for model development and sharing. As a result, these standards often lack the necessary structure and metadata to support AI-specific tasks such as model training, evaluation, and deployment. MEDS aims to fill this gap by providing a standard specifically tailored to the needs of health AI. MEDS aims not to harmonize data, but instead provide a technical specification to make ETL pipelines easy to implement and maintain. Further harmonization can be done on top of MEDS, but it is not a requirement. In this way, MEDS complements efforts such as OMOP, FHIR, I2B2, and PCORnet, serving as a foundation for easily developing and benchmarking high-capacity machine learning.

5.3 Data Conversion to MEDS

To get started with MEDS, we first need to convert a dataset or database, which traditionally consists of a variety of tables with different schemas and keys, into an event stream format ordered by patient. The MEDS community has created several Extract, Transform, Load (ETL) pipelines and helper packages to convert various datasets to the MEDS format. I have expanded the list of available datasets by developing ETL pipelines to convert HiRID [Hyl20], SICdb [Rod+23b], NWICU [Mou+24], and INSPIRE [Lim+24] datasets. Additionally, I developed a publicly available ETL pipeline [van25d] for the Mount Sinai Health System (MSHS) OMOP dataset, which, itself, is not publicly available. Access to the dataset can be granted to researchers within the Mount Sinai Health System or Digital Health Partnership for specific projects. A description of the known compatible datasets can be found in the previous section and in Table 5.1.

Moreover, I developed the MEDS on FHIR ETL [Van25a], which can convert any FHIR R4 or R5-compliant dataset to MEDS. This format is used internally at hospitals worldwide for information exchange; however, there are currently no non-synthetic publicly available FHIR datasets, except for the MIMIC-IV Demo on FHIR [Ben+25].

Note that, while the ETLs are open source and anyone with the proper license and access to the data can run them, we inherently make some choices when converting the data to an event stream format. These choices can impact downstream model performance and generalizability. Moreover, some modalities present in the datasource,

such as waveform or high-resolution vital sign monitoring, clinical notes, or medical imaging, are not yet fully supported in MEDS.

5.3.1 The Mount Sinai Health System Dataset (MSHS)

The Mount Sinai Health System comprises a vast network of hospitals and healthcare facilities throughout New York City. This system maintains the Mount Sinai Data Warehouse (MSDW), where longitudinal clinical data are converted to OMOP and made accessible to researchers through the AI-ready Mount Sinai (AIR·MS) platform, which is a SAP-HANA database instance [Fär+12] with a subset of the data from the MSDW. The MSDW houses records for over 12 million patients, spanning more than 105 million clinical encounters, with regular updates ensuring the availability of current data. These records provide comprehensive longitudinal information, including diagnoses, laboratory results, prescriptions, hospitalizations, and procedures. The urban setting of Mount Sinai hospitals contributes to a diverse patient population that encompasses multiple ethnic and socioeconomic groups. Note that, though converted to the OMOP standard, a significant amount of concepts are *non-standard* concepts, which means that the coding vocabulary remains in the source (Epic EHR) vocabulary. The following hospitals contribute data, starting from the indicated year of the switch to an Epic EHR system, to the Mount Sinai Health System (MSHS) dataset: Mount Sinai Hospital (MSH) — 2011, Mount Sinai Queens — 2013, Mount Sinai West — 2018, Mount Sinai Morningside — 2018, Mount Sinai Brooklyn — 2019, Mount Sinai Beth Israel — 2020.

In this work, we restrict ourselves to data from these centers up to 2025-05-01 for all patient identifiers with at least one entry in the *visit_occurrence* table in the OMOP schema; there is a large number of patients that have been retrospectively imported from the Epic system, who only have demographic information in the OMOP `person` table, no recorded visits, and a small number of events in the OMOP observation table. This results in a dataset of *6,811,731* patients. Table E.1 in Appendix E.1 provides a detailed overview of the MSHS dataset, as provided by the MSDW team, in terms of OMOP tables, record types, distinct patients, and record counts.

Comparing the MSHS dataset to others: Table 5.3 shows the demographic distribution of the MSHS dataset compared to the Columbia University Medical Center (CUMC) [Pan+25a], the Stanford Medicine (SM) EHRSHOT and full SM 'pretraining' (used for training a publicly available *CLMBR-T-base* model [Ste+21]) datasets [Wor+23a]. The demographics of MSHS are broadly similar to those of CUMC, which can be expected due to their geographical proximity (New York City). Moreover, the EHRSHOT and SM datasets exhibit a different distribution, which may be more representative of its location (Palo Alto). We also observe that SM data seems to have less missing demographic characteristics compared to the other datasets. Note that the curated subset from the larger health system [Wor+23a] seems to have less even missing values for race. A discrepancy was observed in the way the *missing ethnicity* values for EHRSHOT were reported in Wornow et al. [Wor+23a] as compared to Pang et al. [Pan+25a]. A possible explanation for this could be that the missing values are subsumed in the 'non-Hispanic' category in Wornow et al. [Wor+23a] as no missing value was reported in that

work (but inferred from the total) and it appears unrealistic that SM reports no missing ethnicity values when we compare to MSHS and CUMC.

Table 5.3: Demographic distribution of the Mount Sinai Health System (MSHS), the Columbia Irving University Medical Center (CUMC), an open-access curated subset of Stanford Medicine EHR (EHRSHOT) data, and the Stanford Medicine (SM) data itself. Table values of the CUMC and EHRSHOT datasets are taken from Appendix Table 4 of Pang et al. [Pan+25a]. The entire Stanford Medicine values are taken from Appendix Table 5 of [Wor+23a]. Both tables are used under CC BY 4.0 license from their respective works. All datasets were collected in the United States of America. Shown are percentages and amounts in parentheses: % (n).

	MSHS	**CUMC**	**EHRSHOT**	**SM**
Location	New York City, NY	New York City, NY	Palo Alto, CA	Palo Alto, CA
Total	(6,811,731)	(6,716,938)	(6,739)	(3,668,615)
Sex				
Female	54.9 (3,737,081)	55.7 (3,744,185)	51.1 (3,441)	53.8 (1,972,089)
Male	45.0 (3,067,363)	44.0 (2,955,940)	48.9 (3,298)	46.2 (1,696,526)
Missing	0.1 (7,287)	0.3 (16,814)	0.0 (0)	0.0 (0)
Ethnicity				
Hispanic or Latino	8.0 (541,846)	8.3 (557,173)	15.4 (1,038)	12.7 (464,638)
Not Hispanic or Latino	33.0 (2,244,409)	16.6 (1,118,232)	77.7 (5,236)	87.3 (3,203,977)
Missing	59.1 (4,025,476)	75.1 (5,041,534)	6.9 (465)	0.0 (0)
Race				
Asian	3.9 (261,913)	1.2 (83,697)	15.5 (1,043)	14.5 (531,121)
Black or African American	9.8 (669,931)	5.9 (396,677)	4.4 (298)	3.3 (119,438)
Native Hawaiian or Other Pacific Islander	1.1 (73,282)	0.1 (6,534)	1.1 (74)	0.8 (29,744)
Native American or Alaskan Native	0.1 (6,789)	0.1 (7,443)	0.4 (25)	0.3 (10,254)
White	27.0 (1,837,612)	16.9 (1,133,364)	55.4 (3,736)	38.4 (1,410,367)
Other	1.2 (82,521)	14.1 (944,342)	0.0 (0)	0.0 (0)
Missing	57.0 (3,879,103)	61.7 (4,144,882)	23.2 (1,563)	42.7 (1,567,691)

5.3.2 Converting MSHS OMOP to MEDS

To work with the data, we first need to get it into a local data system. This is done by creating a split-wise data extract from the SAP HANA system with SQL queries. 'Flattening" the table structures of OMOP into event streams is non-trivial and requires several choices. I developed an ETL that can dynamically extract multiple events from several OMOP tables, which replaces the existing OMOP ETL (`https://github.com/Medical-Event-Data-Standard/meds_etl`). Some improvements are: the choice of using mapped OMOP concepts instead of only source concepts, the addition of the tables: *observation_period*, *specimen*, *dose_era*, *drug_era*, *and condition_era*, and the creation of separate 'start' and 'end' events for rows for several tables to minimize temporal leakage. Moreover, users can customize a YAML file to select which tables to extract and how to build the MEDS events from the available columns in the OMOP tables.

Complex events are created by taking the standardized concepts, i.e., concepts formulated in a standard vocabulary (e.g., SNOMED CT, LOINC, RxNorm) and combining them with additional information from the source tables. For example, a lab measurement event is created by combining the LOINC code with the measurement value and unit. Note that, due to administrative reasons, some event timestamps might be imprecise.

To expand upon the previous section, Observational Medical Outcomes Partnership (OMOP) is a CDM for healthcare data that standardizes the structure and content of

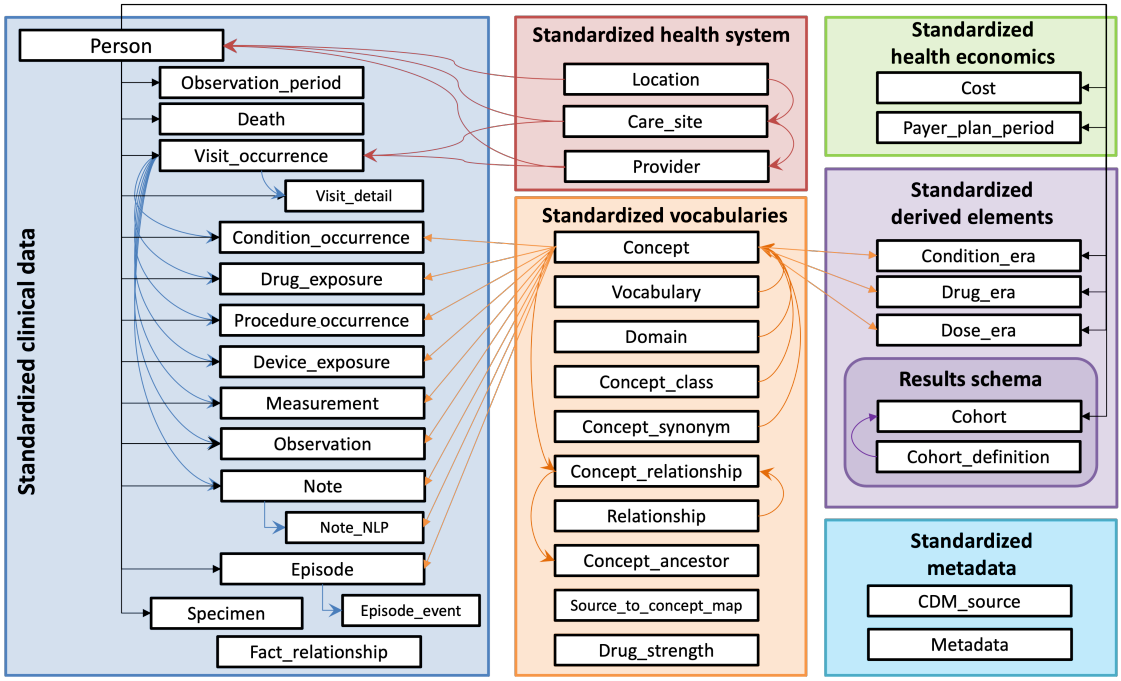


Figure 5.1: OMOP CDM v5.4 data schema [Rei+24]. The CDM is designed to support the conduct of research to identify and evaluate associations between interventions (drug exposure, procedures, healthcare policy changes, etc.) and outcomes caused by these interventions (condition occurrences, procedures, drug exposure, etc.). Content from OHDSI website (`https://ohdsi.github.io/`) used under Creative Commons CC BY-SA 4.0 license

EHR to facilitate research and analysis. It provides a consistent framework for representing clinical data, including patient demographics, diagnoses, procedures, medications, and laboratory results. The OMOP Common Data Model (CDM) is widely used in observational health research and supports the integration of data from diverse sources. Figure 5.1 shows the structure of the OMOP CDM v5.4, the latest supported version. MSHS data is stored in an OMOP 5.3 database in an SAP HANA database. The ETL, however, is designed and tested to work with any OMOP 5.x dataset.

We currently integrate information from the following OMOP tables from the original MSDW OMOP data warehouse: *person*, *observation*, *measurement*, *visit_occurrence*, *observation_period*, *drug_exposure*, *specimen*, *device_exposure*, *dose_era*, *condition_occurrence*, *drug_era*, *procedure_occurrence*, *condition_era*.

The random train, tuning, and held-out splits are created at the patient level with a 80% / 10% / 10% ratio. Table 5.4 shows the demographic distribution of the MSHS data across the different splits. The distributions are similar across the splits, indicating that the random splitting procedure was effective.

5.4 Data Exploration with MEDS-Inspect

A major part of working with clinical data is understanding its structure, completeness, and quality. Because modelling over an event stream format requires thinking about the modelling problem in a different way, we acknowledge the need for effective data exploration tools. I developed a tool that enables users to explore and visualize the converted MEDS data: MEDS-Inspect [Van25e]. It aims to answer common questions

Table 5.4: Demographic distribution of MSHS data across Train, Tuning, Held Out splits. Shown are percentages and counts in parentheses: % (n). Here, we use OHDSI standardized definitions for demographic variables [Rei+24]. Note that the splits are created at the patient level, so each patient only appears in one split.

	Train % (n)	Tuning % (n)	Held Out % (n)
Total	(5,449,385)	(681,173)	(683,173)
Sex			
Female	54.9 (2,989,964)	54.9 (373,717)	54.8 (373,400)
Male	45.0 (2,453,635)	45.0 (306,696)	45.1 (307,032)
Missing	0.1 (5,786)	0.1 (760)	0.1 (741)
Ethnicity			
Hispanic or Latino	8.0 (433,374)	8.0 (54,236)	8.0 (54,236)
Not Hispanic or Latino	33.0 (1,796,150)	32.9 (224,042)	32.9 (224,217)
Missing	59.1 (3,219,861)	59.1 (402,895)	59.1 (402,720)
Race			
Asian	3.9 (209,710)	3.9 (26,281)	3.8 (25,922)
Black or African American	9.8 (535,489)	9.8 (67,100)	9.9 (67,342)
Native Hawaiian or Other Pacific Islander	1.1 (58,767)	1.1 (7,255)	1.1 (7,260)
Native American or Alaskan Native	0.1 (5,408)	0.1 (706)	0.1 (675)
White	27.0 (1,470,338)	27.0 (183,634)	27.0 (183,640)
Other	1.2 (65,980)	1.2 (8,233)	1.2 (8,308)
Missing	57.0 (3,103,213)	57.0 (387,905)	57.0 (387,985)

that arise when working with the relatively new paradigm of clinical event stream data, such as:

- Which and how many codes does my dataset contain?
- How does my data compare to other MEDS datasets?
- What preprocessing steps are still needed in order to train models?

Also, it helps to identify potential data quality issues, such as:

- Are there unexpected gaps in the data?
- Is my ETL missing data?
- Are there unexpected distributions of certain event types?
- Are there unexpected patterns in the temporal distribution of events?

This improves the iterative process of mapping this dataset to an event-stream format for optimal downstream model training and evaluation. For example, for the MEDS on Fast Healthcare Interoperability Resources (FHIR) ETL[Van25a], we need to retrieve the correct identifiers in the nested FHIR resources to create meaningful events. It is built on a Dash framework [Plo24a] and provides interactive visualization through Plotly [Plo24b]. The memory-efficient and fast Polars [Vin+24] Lazy application programming interface (API) is used to dynamically cache visualizations that are expensive to compute. MEDS-Inspect is FOSS [1] and has a self-contained demonstration application that works with the MIMIC-IV Demo MEDS dataset [Joh+23b; Van+25d].

Figure 5.2 shows the interface elements of MEDS-Inspect. We follow the principles of progressive disclosure to prevent information overload [SW19].

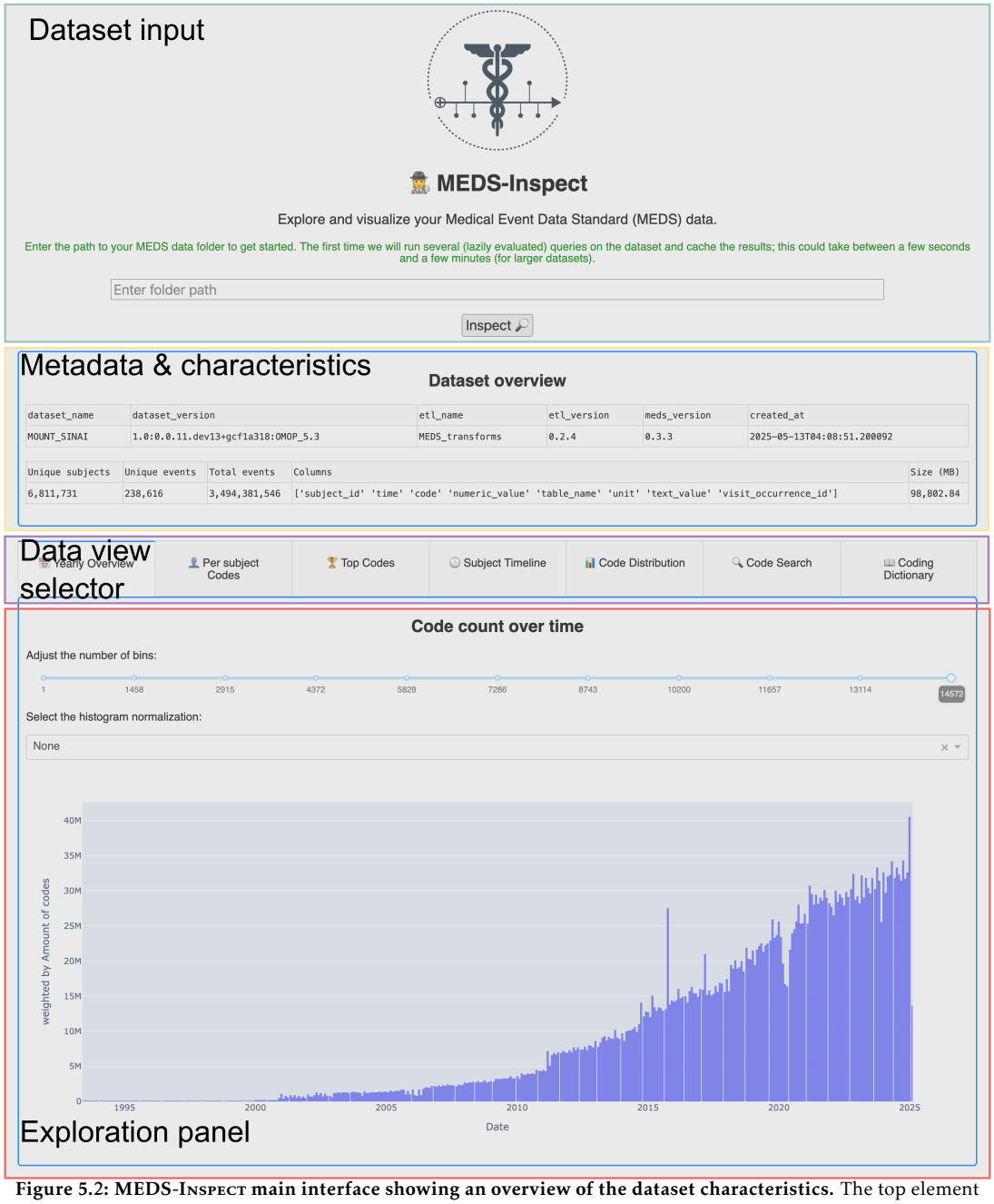


Figure 5.2: MEDS-Inspect main interface showing an overview of the dataset characteristics. The top element shows the dataset selector. Then, important metadata elements are displayed to provide an indication of the dataset's size. The view selector shows the currently implemented *data views*. Currently, the timeline view is displayed for the MSHS OMOP MEDS dataset, which illustrates the addition of new clinical centers and data sources over time. Each graph is interactive and can be zoomed and filtered.

MEDS-Inspect integrates with the task definition tool *ACES* [Xu+24] to help users define cohorts and labels based on the explored data. Figure 5.5 shows the subject timeline view with a defined in-ICU mortality task. Note that the prediction point is 24 hours after the ICU admission.

[1] https://github.com/rvandewater/MEDS-Inspect

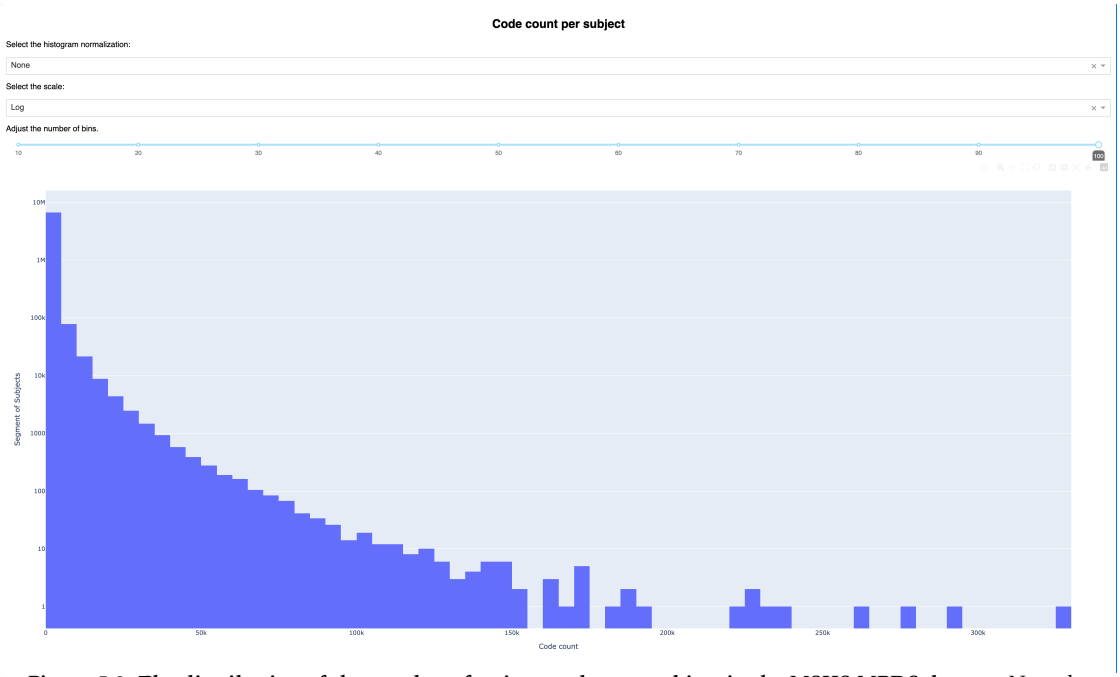


Figure 5.3: The distribution of the number of unique codes per subject in the MSHS MEDS dataset. Note the logarithmic scale on the y-axis.

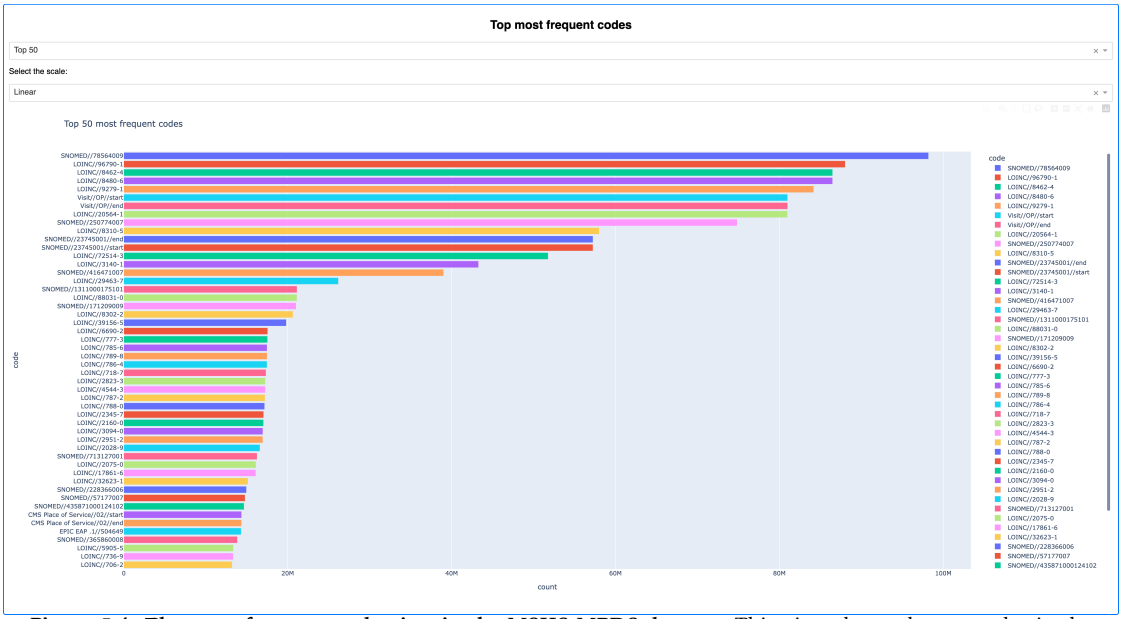


Figure 5.4: The most frequent code view in the MSHS MEDS dataset. This view shows the top codes in the dataset along with their frequencies. The codes can be dynamically filtered by selecting traces in the legend.

Figure 5.6 shows the distribution of diastolic blood pressure measurements in the MSHS MEDS dataset.



Figure 5.5: MEDS-Inspect subject timeline view for NWICU [Mou+24]. This allows users to, for example, explore and debug the labels for the corresponding subject. Note that we do not show the un-deidentified subjects from the MSHS dataset here for data privacy reasons.

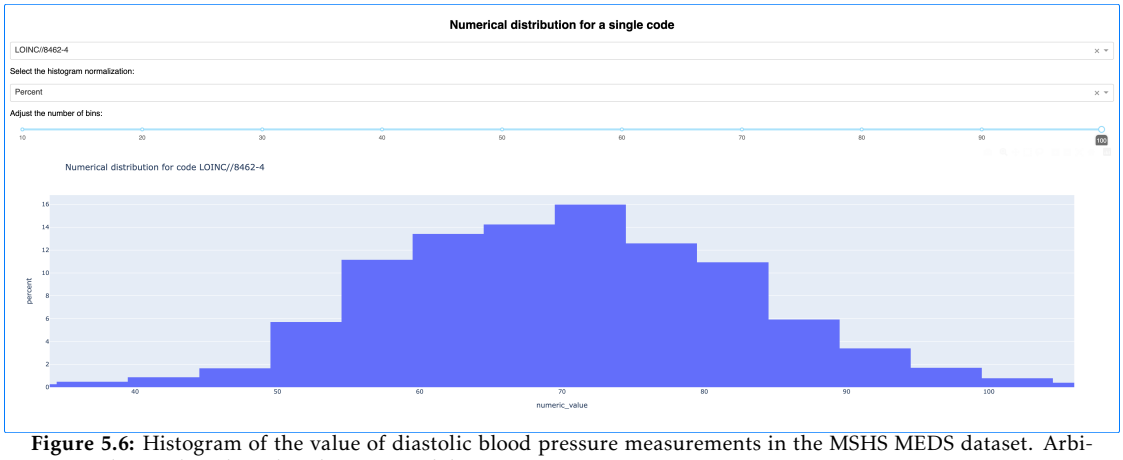
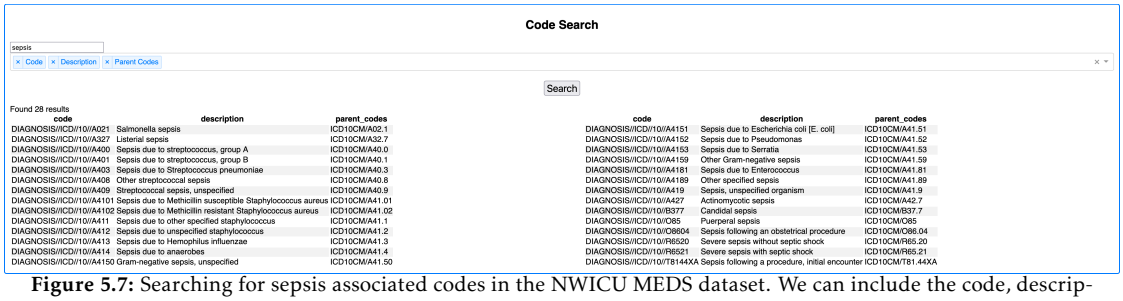


Figure 5.6: Histogram of the value of diastolic blood pressure measurements in the MSHS MEDS dataset. Arbitrary codes can be selected in the top search box.



code	description	parent_codes
DIAGNOSIS//ICD//10//A021	Salmonella sepsis	ICD10CM/A02.1
DIAGNOSIS//ICD//10//A327	Listerial sepsis	ICD10CM/A32.7
DIAGNOSIS//ICD//10//A400	Sepsis due to streptococcus, group A	ICD10CM/A40.0
DIAGNOSIS//ICD//10//A401	Sepsis due to streptococcus, group B	ICD10CM/A40.1
DIAGNOSIS//ICD//10//A403	Sepsis due to Streptococcus pneumoniae	ICD10CM/A40.3
DIAGNOSIS//ICD//10//A408	Other streptococcal sepsis	ICD10CM/A40.8
DIAGNOSIS//ICD//10//A409	Streptococcal sepsis, unspecified	ICD10CM/A40.9
DIAGNOSIS//ICD//10//A4101	Sepsis due to Methicillin susceptible Staphylococcus aureus	ICD10CM/A41.01
DIAGNOSIS//ICD//10//A4102	Sepsis due to Methicillin resistant Staphylococcus aureus	ICD10CM/A41.02
DIAGNOSIS//ICD//10//A411	Sepsis due to other specified staphylococcus	ICD10CM/A41.1
DIAGNOSIS//ICD//10//A412	Sepsis due to unspecified staphylococcus	ICD10CM/A41.2
DIAGNOSIS//ICD//10//A413	Sepsis due to Hemophilus influenzae	ICD10CM/A41.3
DIAGNOSIS//ICD//10//A414	Sepsis due to anaerobes	ICD10CM/A41.4
DIAGNOSIS//ICD//10//A4150	Gram-negative sepsis, unspecified	ICD10CM/A41.50
DIAGNOSIS//ICD//10//A4151	Sepsis due to Escherichia coli [E. coli]	ICD10CM/A41.51
DIAGNOSIS//ICD//10//A4152	Sepsis due to Pseudomonas	ICD10CM/A41.52
DIAGNOSIS//ICD//10//A4153	Sepsis due to Serratia	ICD10CM/A41.53
DIAGNOSIS//ICD//10//A4159	Other Gram-negative sepsis	ICD10CM/A41.59
DIAGNOSIS//ICD//10//A4181	Sepsis due to Enterococcus	ICD10CM/A41.81
DIAGNOSIS//ICD//10//A4189	Other specified sepsis	ICD10CM/A41.89
DIAGNOSIS//ICD//10//A419	Sepsis, unspecified organism	ICD10CM/A41.9
DIAGNOSIS//ICD//10//A427	Actinomycotic sepsis	ICD10CM/A42.7
DIAGNOSIS//ICD//10//B377	Candidal sepsis	ICD10CM/B37.7
DIAGNOSIS//ICD//10//O85	Puerperal sepsis	ICD10CM/O85
DIAGNOSIS//ICD//10//O8604	Sepsis following an obstetrical procedure	ICD10CM/O86.04
DIAGNOSIS//ICD//10//R6520	Severe sepsis without septic shock	ICD10CM/R65.20
DIAGNOSIS//ICD//10//R6521	Severe sepsis with septic shock	ICD10CM/R65.21
DIAGNOSIS//ICD//10//T814XA	Sepsis following a procedure, initial encounter	ICD10CM/T81.44XA

Figure 5.7: Searching for sepsis associated codes in the NWICU MEDS dataset. We can include the code, description, and parent code fields in the search.

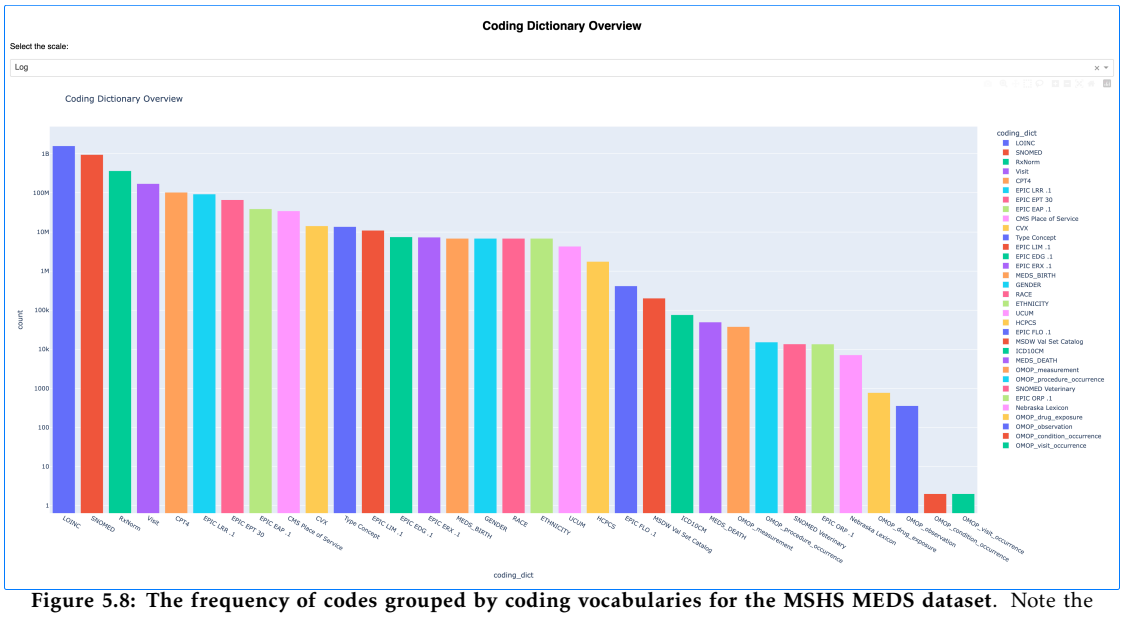


Figure 5.8: The frequency of codes grouped by coding vocabularies for the MSHS MEDS dataset. Note the logarithmic scale on the y-axis. We can dynamically filter the vocabulary by selecting traces in the legend.

5.5 Converting Clinical Coding Vocabularies with HarmonizEHR

As shown in Figure 5.9, we frequently encounter variations in coding vocabularies across datasets. While this is not necessarily an issue for training models within a single dataset or health system, many applications in the ML research domain, such as federated learning [Pfi+25] and transfer learning [Win+25], require harmonized features.

An example of adaption in the FM space is the work by Guo et al. [Guo+24], which uses an open-access model trained on Stanford Medicine data [Ste+21] and applies direct transfer learning, yielding decent results. This was done with the involvement of the original authors of the model. However, Pang et al. [Pan+25a] also uses the pretrained model on a major dataset and reports poor results due to differences in healthcare institutions and a 'severe lack of overlap" in codes, which meant that the tokenizer, and therefore the model, only uses a subset of codes. Moreover, defining cohorts and endpoints takes less time and could be less error-prone if we can use a common vocabulary across datasets: for example, defining a cohort of patients with diabetes could be done by using a common set of codes that represent diabetes across datasets, which is less error prone than manually querying the dataset for the correct codes, units, and thresholds. Observational Medical Outcomes Partnership (OMOP) is a primary example of an ecosystem that is adapted by many institutions worldwide to support clinical informatics research [Vos+15]. However, when systems adapt to OMOP, they often comply only with the CDM and only harmonize to standard vocabularies partially, leading to a variety of non-standard coding vocabularies; a good example of this is found in the MSHS dataset: Figure 5.8 shows that there is a large amount of codes that are still in the unmapped vocabulary. To provide an overview, the MSHS MEDS dataset comprises 14,415 unique codes, with more than 100 occurrences, totaling over 222 million occurrences. The most frequently occurring code occurs more than 143 million times in the dataset.

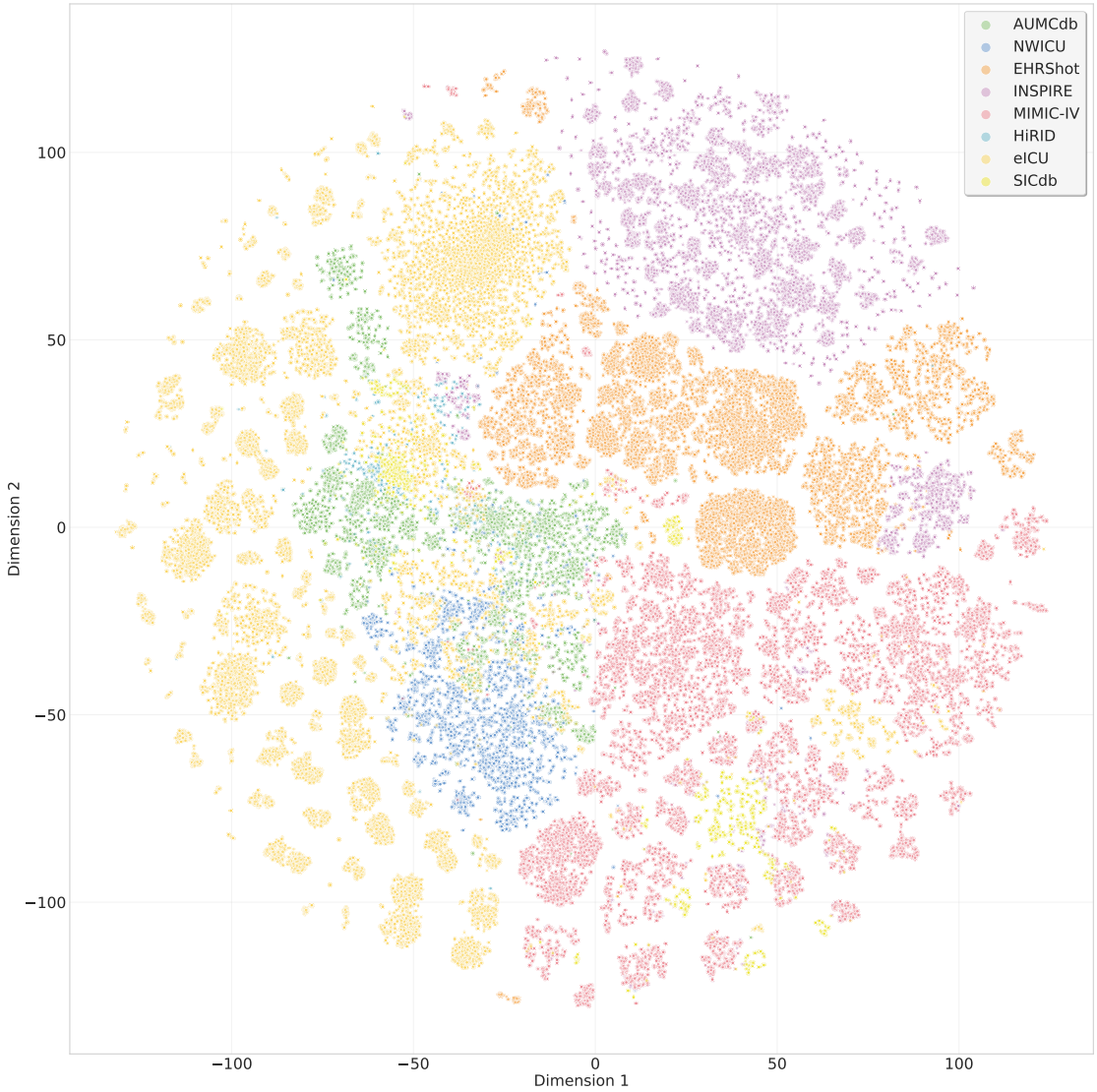


Figure 5.9: A t-SNE visualization of the unique codes in datasets converted to MEDS. Colors indicate the dataset of origin as shown in the legend. The clustering of codes from different datasets indicates that similar concepts are represented across datasets, despite the use of different coding vocabularies. This highlights the potential for HarmonizEHR to standardize clinical concepts across diverse datasets.

HarmonizEHR[2] is an experimental MEDS-compatible tool built on the existing NLP toolkit xMEN, a modular system for cross-lingual medical entity normalization [Bor+24] that can map to UMLS [Bod04] concepts. UMLS is a compendium of controlled vocabularies in the biomedical sciences and provides a mapping structure between these vocabularies. Common vocabularies include: ICD [Org92], SNOMED CT [Don06], and LOINC [Huf+98]. HarmonizEHR can normalize concepts and provide configuration files that allow for approachable editing and adjustment of the proposed mappings. Moreover, the underlying xMEN library employs an ensemble of Term Frequency-Inverse Document Frequency (TF-IDF) [Spa72] vectorizers and a SAPBert (Self- Alignment Pretraining BERT) model [Liu+21] on the raw text to create an embed-

[2] https://github.com/rvandewater/harmonizehr

ding space, which can then be used for medical entity normalization. HarmonizEHR extends xMEN for the more specialized task of harmonizing MEDS coding vocabularies. It uses the MEDS *codes* metadata to generate mappings from source codes and descriptions (and an arbitrary amount of extension columns) to target index vocabularies.

Returning to the issue of non-standard vocabularies in clinical datasets, I developed a benchmark dataset to map the original source codes, captured in OMOP, to UMLS concepts. We decided to use the LOINC codes as they are the most prevalent in the Mount Sinai Health System (MSHS) dataset. We used the `measurement` table, joining the `measurement_source_concept_id` columns with the `concept` table to obtain a natural language description (if available); this resulted in 14,901 values. Taking the predicted term with the highest confidence, we ended up with 5.2% correctly predicted codes. The low performance is likely due to the short and often ambiguous descriptions in the source data, as well as the 17.14 million unique LOINC codes. We expect an improvement if the mapping vocabulary is restricted to a smaller set of more frequently occurring codes. For example, when we convert codes from an MSHS subset to an SM or Columbia University Medical Center (CUMC) subset. Extensions with Human-in-the-Loop (HitL) functionalities for expert mappers or clinicians like in Van de Water et al. [Van+22b] are another future improvement. One could also consider using the embeddings directly, as in Kirchler et al. [Kir+25a]; however, this approach must be supported by the model training regime. To retain the clinical meaning of the codes and the unreliable performance of the mapping as benchmarked, we currently do not use HarmonizEHR for downstream model training with MSHS data.

5.6 Scaling Clinical Model Benchmarking across Datasets and Tasks with MEDS-DEV

After performing data conversion and exploration using the tools above, we proceed to model training and evaluation. Here, we can scale comparisons and experiments across datasets. Though we could encounter several challenges due to the heterogeneity in dataset characteristics related to how event-stream datasets are structured and recorded.

5.6.1 Comparing Event-Stream Characteristics in Open-Access Datasets

Figure 5.10 shows the two key characteristics of event-stream data: the sequence length (number of sequential events per subject) and the inter-event time (time between subsequent events for each subject). We can observe considerable variability in these characteristics across datasets, which may be attributed to several factors: the type of dataset (ICU vs. general hospital vs. surgical), the institution's clinical practices, the data collection and integration process, and any subsequent ETL processes. All these might impact the sequence length and time between events. For example, SICdb [Rod+23b] contains additional "waveform" data[3] with a frequency of 1-minute that is currently not integrated into the event-stream (just the hourly mean); integrating this data will likely require additional data transformations to make it manageable to train on for many

[3] `https://www.sicdb.com/Documentation/Signal_Data` [Accessed 2026-01-20]

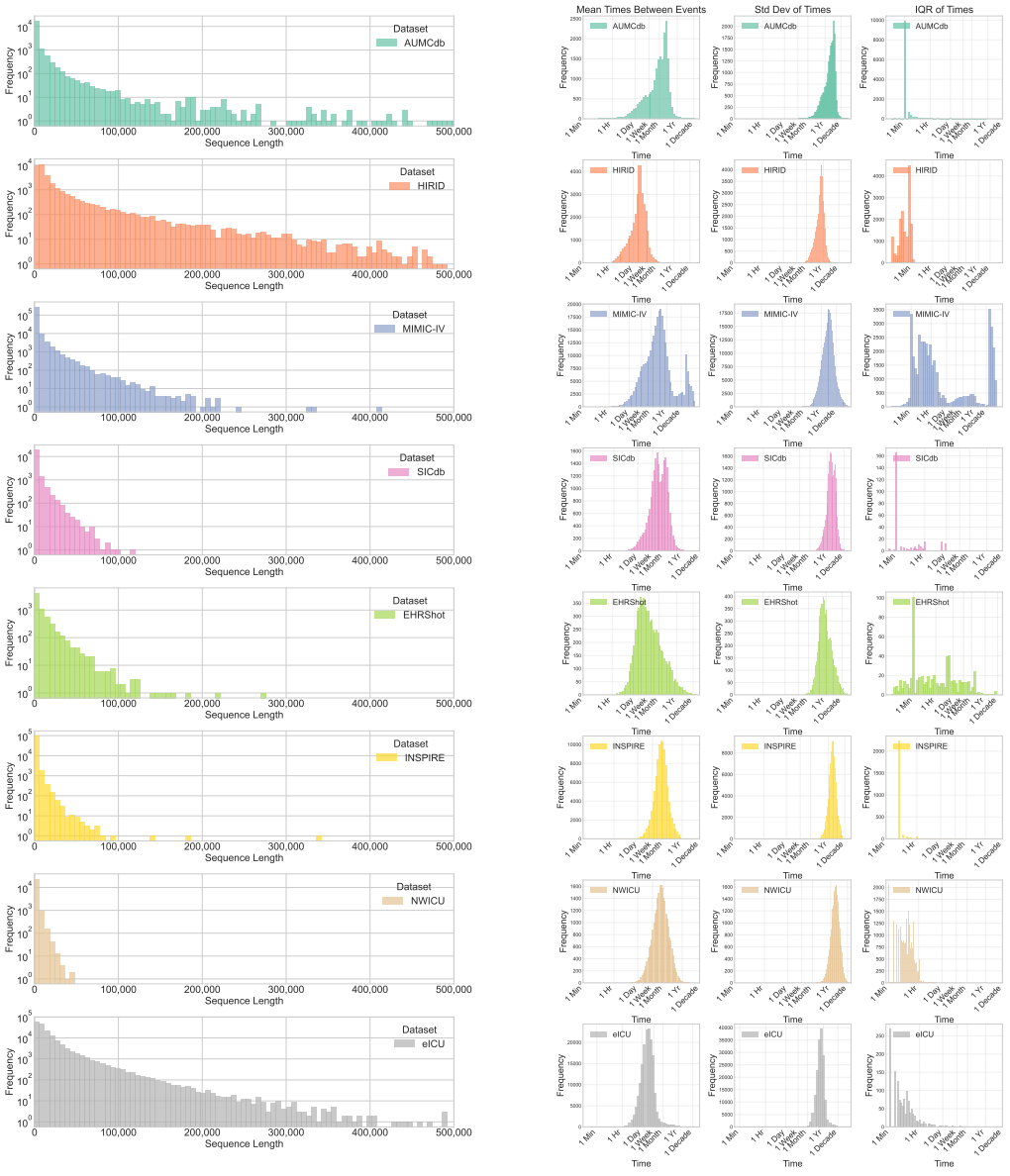


Figure 5.10: Data characteristics across MEDS datasets. Left: Distribution of sequence lengths (number of events per patient) across MEDS datasets. Right inter-event times across MEDS open-access datasets; from left to right: the mean, standard deviation, and inter-quartile range of inter-event times. Both plots were generated using a similar methodology to Wornow et al. [Wor+24], but with methods that work for all MEDS datasets (regardless of tokenizer). These characteristics highlight the variability in data density and temporal resolution across different clinical datasets, which can impact model training and evaluation.

models. This contrasts with HiRID [Hyl20], where I chose the "raw" minutes-interval on the basis that the aggregated data was too coarse and would miss many features; the differences can be seen both in the sequence length, where HiRID resembles the similar AUMCdb the closest, and in the inter-event times, which are skewed more towards the left. Notably, the longer sequence length has an impact on model training: models on HiRID take longer to train than similarly sized datasets with a lower resolution. This may impact model generalizability when it comes to transfer learning: models trained on datasets with longer sequences may not perform well on datasets with shorter sequences, and vice versa. Moreover, we observe that EHRSHOT [Wor+23a] has a more heterogeneous distribution of inter-event times, as expected in the general hospital setting; this contrasts with the ICU datasets, where patients are monitored more closely, and events are recorded more frequently. Furthermore, the only surgical dataset, INSPIRE [Lim+24], exhibits characteristics of a (heavily) preprocessed dataset with minimal inter-event time variation. These characteristics could impact the model's ability to learn temporal patterns in the data, as the variability in inter-event times may make it more challenging to capture long-term dependencies. We stress the importance of understanding and using these characteristics when designing modeling systems with clinical (event-stream) data.

5.6.2 Benchmarking

One of the key advantages of MEDS is the ability to use existing pipelines and tooling to easily reproduce the same or similar experiments across datasets. MEDS-DEV [McD+24] is a benchmarking framework that allows users to evaluate their models on various clinical tasks using the MEDS format. We intended it as a decentralized tool and combines the benefits of public datasets with the ability to benchmark on private datasets without sharing sensitive information. Configurations are shared publicly as a hierarchy of Yet Another Markup Language (YAML) files, and users can clone the project and run the experiments locally on their own data. Different tools make up MEDS-DEV. We have individual data transformation ETLs for each dataset (Table 5.1), the label or task generation tool [Xu+24], and several models [Hur+24; Ouf+24a; Pan+21].

5.6.3 Cohort Creation

Cohorts can be created using the Automated Cohort Extraction System for Event Streams (*ACES*) tool [Xu+24]. This tool allows users to define cohorts and labels using a simple domain-specific language formatted as YAML files. It supports various types of clinical endpoints and allows for the creation of predicates, such that tasks can, in theory, be defined for multiple datasets at the same time. We have created several cohorts for the MSHS dataset which can be found a software extension (fork) of MEDS Decentralized External Validation (*MEDS-DEV*)[4] and are pending integration in the main benchmark. Tasks defined in McDermott et al. [McD+24] have several categories: *Complete Blood Count (CBC)*, *Blood chemistry*, *Vital signs*, *Mortality prediction*, and *Readmission*. Below is a list of all the tasks currently defined in MEDS-DEV, including two new tasks added in this work (hyperkalemia and hypoglycemia): Complete Blood Count (CBC) tasks:

[4] `https://github.com/rvandewater/MEDS-DEV/tree/main`

Table 5.5: *Tasks defined across datasets over each split of each dataset.* Percentages (%) denote the positive case prevalence. Despite the heterogeneity in dataset type: ICU (AUMCdb, HiRID, MIMIC-IV, NWICU, SICdb, eICU), general health system or hospital (EHRSHOT, MSHS), and mixed operative (INSPIRE), we can define the same tasks across datasets. Note that some datasets lack certain tasks due to a lack of correct measurement or admission events. Tasks are benchmarked if their prevalence is above 0.01% in all splits. Tasks are marked in red if unsuitable for benchmarking.

Task	AUMCdb	EHRSHOT	HIRID	INSPIRE	MIMIC-IV	NWICU	SICdb	eICU	MSHS
30D Read.	11,278(11%)	706,483(89%)	–	13,005(3%)	330,753(23%)	30,854(40%)	2,458(4%)	–	17,784,788(45%)
train	9,020(11%)	556,984(89%)	–	10,370(3%)	265,147(23%)	24,454(40%)	1,957(4%)	–	14,236,503(45%)
held out	1,094(12%)	77,174(89%)	–	1,342(3%)	32,883(23%)	3,007(39%)	257(4%)	–	1,781,483(45%)
tuning	1,164(13%)	72,325(89%)	–	1,293(3%)	32,723(23%)	3,393(40%)	244(5%)	–	1,766,802(45%)
Anemia	150(45%)	387(32%)	1,206(57%)	152(48%)	7,718(66%)	4,374(35%)	1,225(47%)	24,471(46%)	118,039(27%)
train	114(46%)	293(31%)	970(56%)	128(47%)	6,211(66%)	3,437(35%)	981(46%)	19,634(46%)	94,624(27%)
held out	20(45%)	54(28%)	115(57%)	15(60%)	747(63%)	466(36%)	130(55%)	2,411(45%)	11,663(27%)
tuning	16(38%)	40(45%)	121(59%)	9(44%)	760(66%)	471(36%)	114(53%)	2,426(45%)	11,752(27%)
Creatinine	621(7%)	4,931(2%)	5,460(7%)	2,113(4%)	114,946(2%)	19,667(3%)	8,897(5%)	67,698(6%)	378,295(3%)
train	513(6%)	3,875(2%)	4,339(7%)	1,696(4%)	91,775(2%)	15,606(3%)	7,110(5%)	54,092(6%)	302,812(3%)
held out	51(14%)	535(2%)	543(7%)	208(4%)	11,593(2%)	2,005(3%)	878(6%)	6,852(5%)	37,384(3%)
tuning	57(7%)	521(1%)	578(6%)	209(6%)	11,578(2%)	2,056(2%)	909(5%)	6,754(6%)	38,099(3%)
Hyperkal.	12,154(0%)	167,807(0%)	14,289(2%)	5,784(2%)	202,673(2%)	37,291(2%)	–	134,347(2%)	420,268(4%)
train	9,682(1%)	134,095(0%)	11,438(2%)	4,633(2%)	162,101(2%)	29,387(2%)	–	107,498(2%)	336,313(4%)
held out	1,216(0%)	16,851(0%)	1,405(2%)	564(1%)	20,282(2%)	3,864(2%)	–	13,438(2%)	41,639(4%)
tuning	1,256(0%)	16,861(0%)	1,446(2%)	587(1%)	20,290(2%)	4,040(2%)	–	13,411(2%)	42,316(4%)
Hypoglyc.	1(100%)	14,001(0%)	13,317(0%)	–	34,864(2%)	36,990(2%)	–	123,991(2%)	561,100(3%)
train	1(100%)	11,123(0%)	10,629(0%)	–	28,020(2%)	29,221(2%)	–	99,159(2%)	448,767(3%)
held out	–	1,464(0%)	1,311(0%)	–	3,453(2%)	3,786(2%)	–	12,395(2%)	55,956(3%)
tuning	–	1,414(0%)	1,377(0%)	–	3,391(2%)	3,983(2%)	–	12,437(2%)	56,377(3%)
Hyponatr.	650(3%)	2,144(8%)	5,624(9%)	998(15%)	101,735(5%)	19,551(6%)	9,067(8%)	77,878(8%)	43,948(6%)
train	532(3%)	1,696(8%)	4,476(9%)	798(15%)	81,211(5%)	15,511(6%)	7,222(8%)	62,283(8%)	43,948(6%)
held out	58(5%)	204(9%)	544(10%)	101(18%)	10,376(5%)	2,031(6%)	913(9%)	7,828(8%)	–
tuning	60(5%)	244(9%)	604(9%)	99(12%)	10,148(5%)	2,009(5%)	932(7%)	7,767(7%)	–
Hypotens.	197(27%)	3,046(5%)	1,271(38%)	72(90%)	15,630(37%)	382(84%)	–	43,241(33%)	287,242(14%)
train	162(28%)	2,363(4%)	1,026(38%)	57(93%)	12,465(37%)	316(84%)	–	34,697(33%)	229,991(14%)
held out	17(35%)	370(7%)	118(47%)	5(100%)	1,596(36%)	32(88%)	–	4,301(32%)	28,736(14%)
tuning	18(17%)	313(4%)	127(35%)	10(70%)	1,569(39%)	34(79%)	–	4,243(33%)	28,515(14%)
ICU Mort.	8,077(16%)	232,601(0%)	9,150(10%)	3,936(5%)	34,960(9%)	12,866(10%)	21,583(18%)	77,805(4%)	–
train	6,458(16%)	183,482(0%)	7,319(10%)	3,125(5%)	28,071(9%)	10,190(10%)	17,284(18%)	62,306(4%)	–
held out	798(16%)	25,156(0%)	886(8%)	408(4%)	3,432(9%)	1,343(9%)	2,140(18%)	7,752(4%)	–
tuning	821(14%)	23,963(0%)	945(10%)	403(5%)	3,457(8%)	1,333(10%)	2,159(18%)	7,747(5%)	–
Leukocyt.	217(28%)	405(31%)	4,266(21%)	823(28%)	16,805(25%)	–	4,523(17%)	49,105(17%)	278,005(11%)
train	175(28%)	297(34%)	3,406(21%)	639(27%)	13,524(25%)	–	3,603(18%)	39,279(17%)	222,423(11%)
held out	22(36%)	50(20%)	407(19%)	88(34%)	1,653(26%)	–	451(14%)	4,958(17%)	27,724(12%)
tuning	20(15%)	58(26%)	453(23%)	96(30%)	1,628(25%)	–	469(18%)	4,868(17%)	27,858(11%)
MA	–	–	3,728(6%)	–	88,069(5%)	–	–	–	63,768(1%)
train	–	–	2,978(6%)	–	70,299(6%)	–	–	–	51,063(1%)
held out	–	–	380(7%)	–	8,831(5%)	–	–	–	6,313(1%)
tuning	–	–	370(6%)	–	8,939(5%)	–	–	–	6,392(1%)
Thrombocyt.	538(13%)	2,887(6%)	4,690(15%)	1,149(20%)	114,133(6%)	17,965(7%)	8,613(16%)	75,125(12%)	431,881(6%)
train	435(13%)	2,276(6%)	3,747(15%)	916(20%)	91,336(6%)	14,188(7%)	6,885(16%)	60,193(12%)	346,124(6%)
held out	52(15%)	292(5%)	448(16%)	124(17%)	11,288(6%)	1,845(7%)	859(14%)	7,481(11%)	42,795(6%)
tuning	51(14%)	319(5%)	495(14%)	109(21%)	11,509(5%)	1,932(7%)	869(13%)	7,451(12%)	42,962(6%)
Train	16,087	5,368	27,123	79,909	239,770	19,751	17,253	133,084	5,449,385
Held out	2,011	671	3,391	9,988	29,971	2,593	2,156	16,635	683,173
Tuning	2,011	671	3,390	9,989	29,971	2,592	2,157	16,636	681,173

Note: 30D Read.: 30-day readmission; Hyperkal.: hyperkalemia; Hypoglyc.: hypoglycemia; Hyponatr.: hyponatremia; Hypotens.: hypotension; ICU Mort.: ICU mortality; Leukocyt.: leukocytosis; MA: metabolic acidosis; Thrombocyt.: thrombocytopenia.

- Anemia: low hemoglobin levels indicating reduced oxygen-carrying capacity.
- Leukocytosis: high white blood cell count indicating potential infection or inflammation.
- Thrombocytopenia: low platelet count indicating potential bleeding risk.

Blood chemistry tasks:

- Elevated creatinine: indicating potential kidney dysfunction.
- Hyponatremia: low sodium levels indicating potential fluid imbalance.

- Hyperkalemia: high potassium levels indicating potential cardiac risk (added in this work).
- Metabolic acidosis: indicating potential metabolic disturbance.
- Hypoglycemia: low blood glucose levels indicating potential diabetic complications (added in this work).

Vital sign tasks:

- Hypotension: low blood pressure indicating potential shock or dehydration.

Mortality prediction tasks:

- ICU mortality: predicting the likelihood of death for patients admitted to the intensive care unit (ICU).

Readmission task:

- 30-day hospital readmission: predicting the likelihood of a patient being readmitted to the hospital within 30 days of discharge.

The tasks are all formulated as binary classification tasks, where the model predicts whether a patient will experience the defined outcome based on their entire clinical history up to the first 24 hours in their stay; the 30-day readmission task is an exception, which includes the data up to the point of discharge, which is also the time of prediction. There is a required "no-discharge" window of 24 hours after the moment of prediction, which filters out patients who are discharged soon after prediction. Then, if the respective *endpoint* occurs within the period 24 hours after the "input" period, the sample is considered positive. Additionally, the tasks do not contain patients under the age of 18 at admission. As an example, we provide the predicates configuration for AUMCdb in the appendix: Listing E.1. Moreover, we provide the YAML *ACES* definition file defined in *MEDS-DEV* for the hyperkalemia task: Listing E.2. Exact definitions of all tasks can be found in the *ACES* configuration files in the `tasks` subdirectory of the forked MEDS-DEV repository[5].

The initial tasks were developed on MIMIC-IV. However, by developing an ETL and providing the correct predicates in *ACES*, I have defined the same tasks on eight other datasets as well: AUMCdb, HiRID, INSPIRE, NWICU, SICdb, eICU, EHRSHOT, and MSHS. The tasks were based on existing benchmark tasks; they should be able to be performed in a decentralized manner on other datasets, such as SM and CUMC [Pan+25a; Wor+23a; McD+24]. Table 5.5 shows the prevalence of each task across datasets. We note that not all tasks are defined on all datasets due to a lack of the necessary measurements or admission events.

[5] `https://github.com/rvandewater/MEDS-DEV/tree/main/src/MEDS_DEV/tasks`

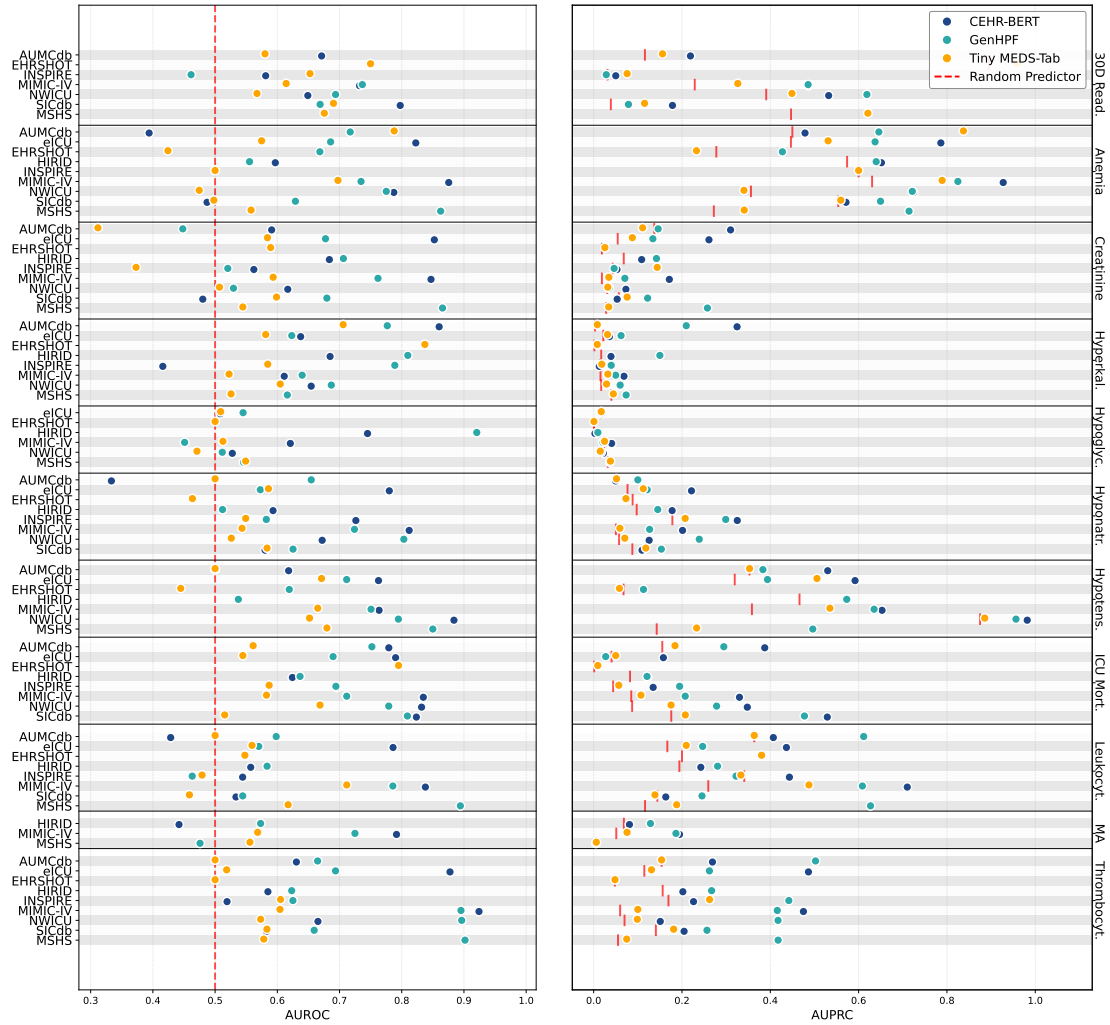


Figure 5.11: Performance metrics for the samples weighted benchmarking MEDS-DEV compatible models. We perform the benchmarks on 9 datasets (left) and the subset of compatible tasks (out of 11) for these datasets (right). Left column: Area Under the Receiver Operating Characteristic (AUROC); Right column Area Under the Precision Recall Curve (AUPRC). The vertical red line indicates the performance of a random predictor; this is 0.50 for AUROC and the prevalence in the test cohort for AUPRC.

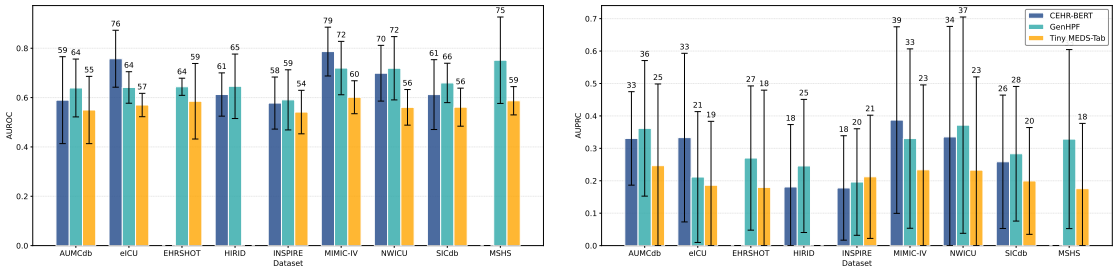


Figure 5.12: AUROC and AUPRC aggregated by dataset performance metrics for the samples weighted benchmarking *MEDS-DEV* compatible models. The error bars represent the standard deviation calculated over the individual samples. Mean values are shown on top of the error bars and rounded to two digits and multiplied by 100 for readability. Left: AUROC; Right AUPRC.

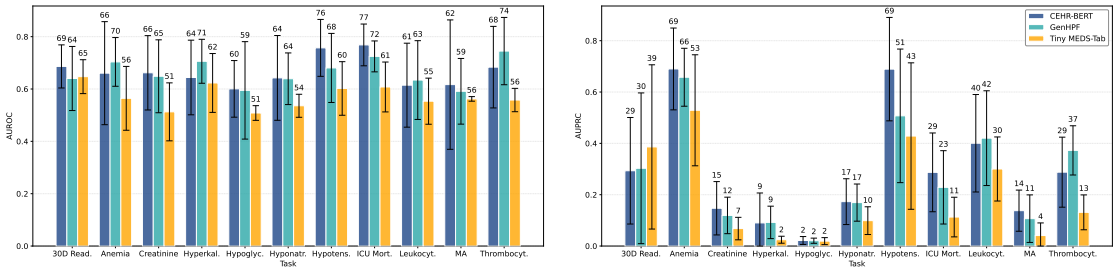


Figure 5.13: AUROC and AUPRC aggregated by task performance metrics for the samples weighted benchmarking *MEDS-DEV* compatible models. The error bars represent the standard deviation calculated over the individual samples. Mean values are shown on top of the error bars and rounded to two digits and multiplied by 100 for readability. Left: AUROC; Right AUPRC.

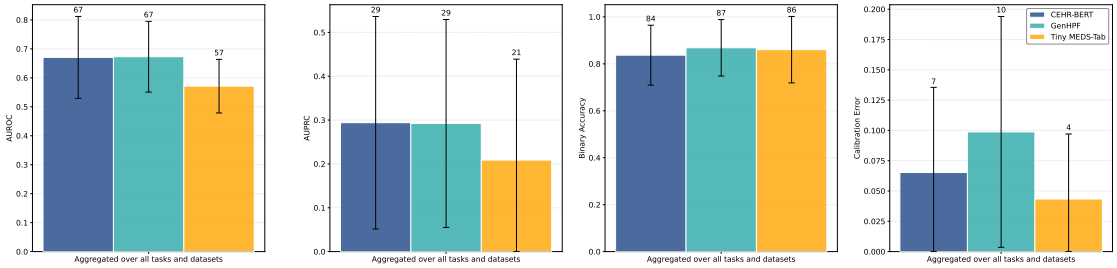


Figure 5.14: Aggregated overall performance metrics for the samples weighted benchmarking *MEDS-DEV* compatible models. The error bars represent the standard deviation calculated over the individual samples. Mean values are shown on top of the error bars and rounded to two digits and multiplied by 100 for readability. From left to right: AUROC, AUPRC, binary accuracy, calibration error.

5.6.4 Models

We use the existing defined models in *MEDS-DEV* to benchmark several models as these implementations are integrated into the framework by their original authors. Some bugs/incompatibilities had to be resolved to ensure correct functionality across all datasets and tasks. Other models could be easily integrated into the framework in future work. Hyperparameter choices are provided as code in the *MEDS-DEV* repository and in Appendix E.3.

We create a baseline with MEDS-Tab [Ouf+24a], a tool that automatically aggregates the event-stream data into different windows and trains a gradient-boosted tree model based on the resulting features. It has proven to be a strong baseline on several datasets [McD+24]. The 'tiny' version only aggregates data over 7 days and 30 days. MEDS-Tab is incompatible with HiRID due to its very high frequency data needing extensive memory.

Moreover, we benchmark GenHPF, which is a multi-task FM for healthcare data based on the transformer architecture [Hur+24]. GenHPF converts EHRs into a hierarchical textual representation while incorporating as many features as possible. It does this by embedding the event code and descriptions as text and aggregating them. This should result in a multi-task model that can be used for various downstream tasks. We, however, pre-train GenHPF on each task separately to ensure optimal performance. I have collaborated with the original authors to make several fixes to ensure GenHPF works correctly with the 9 datasets used in this benchmark.

Finally, we benchmark CEHR-BERT [Pan+21], a Bidirectional Encoder Representations from Transformers (BERT)-based model that incorporates temporal information

in specific tokens taken from the event stream data timestamps. It is fine-tuned for the respective downstream task due to its default unsupervised training method and high requirements on storage and compute. CEHR-BERT seems incompatible with MSHS, presumably due to the large size of the dataset. Moreover, CEHR-BERT is not robust to some events in EHRSHOT, causing it to crash during training. The issues have been reported to the original authors who have yet to address them.

5.6.5 Results

The results of all of the above tasks individually can be seen in Figure 5.11. Model training for the open-access datasets was done on an on-premise General Data Protection Regulation (GDPR) compliant cluster at the Hasso Plattner Institute, Potsdam, Germany. Model training for the Mount Sinai Health System dataset was done on the secure on-premise Health Insurance Portability and Accountability Act (HIPAA) compliant cluster at the Icahn School of Medicine at Mount Sinai, New York, United States of America. We are unable to report cross-validation results due to the the way the tasks are split from the complete dataset before model training, as discussed in the previous section. We report results both in terms of subject and sample weighted versions to correct for subjects that have multiple prediction samples. Here, we observe some individual experiments performing worse than random guessing (i.e., AUROC < 0.50), which could be due to several reasons, such as small sample sizes, class imbalance, or model instability.

To gain a better picture of the general trends, we aggregate results by dataset and task; these can be seen in Figures 5.12 and 5.13 for the samples weighted benchmarks. We see that both models generally outperform the MEDS-Tab baseline. Moreover, for some datasets, like AUMCdb, HiRID, INSPIRE, NWICU, and SICdb, GenHPF outperforms CEHR-BERT by a small margin (within confidence intervals); for eICU and MIMIC-IV CEHR-BERT outperforms GenHPF. A similar trend is seen in the task-aggregated results, where GenHPF outperforms CEHR-BERT on anemia, hyperkalemia, leukocytosis, and thrombocytopenia, while CEHR-BERT outperforms GenHPF on 30-day hospital readmission, elevated creatinine, hyponatremia, hypotension, ICU mortality, and metabolic acidosis. Notable is the large spread in performance across tasks. Overall, both models perform similarly when averaged across all datasets and tasks, as can be seen in Figure 5.14, where both models perform almost identically and outperform the MEDS-Tab baseline in most metrics. Interestingly, CEHR-BERT has a slightly better calibration error than GenHPF. All results for the subjects weighted benchmarks, which differ only slightly from the sample weighted results, and additional metrics can be found in Appendix E.4.

The results seem to suggest that both models are competitive across a wide range of datasets and tasks. However, there is no best model, as they both outperform each other on different datasets and tasks. This suggests that both models have their strengths and weaknesses; the choice of model should depend on the specific dataset and task at hand. Furthermore, a robust model training and evaluation framework, such as *MEDS-DEV* could be important to operationalize the use of these models in practice. Then, each institution could pick a model that outperforms the others on an experimental

basis using continuous evaluation. Moreover, ensemble methods could be explored to leverage the strengths of both models.

5.7 Foundation Model Training and Evaluation on the Mount Sinai Health System

The training and evaluation of FMs requires many components that have been partially discussed before. Figure 5.15 shows an overview of the entire process of FM training and evaluation on the MSHS dataset.

ETL The multiple centers that comprise MSHS contribute their EHR data to a centralized data warehouse, which is then converted to the OMOP CDM [Rei+24] and partially mapped. An ETL is then used to convert the OMOP CDM data into the flat MEDS, data-as-an-interface format. The data is then split into training, tuning, and held-out sets.

Cohorts Cohorts and labels are created using *ACES* [Xu+24]. This has advantages over earlier approaches (c.f., [Pan+25a] and [Wor+24]): *ACES* works with predicates and task configurations that have been developed for this benchmark, but can be 1) versioned easily, 2) extracted again as the database updates and 3) reused in a modular manner to construct more cohorts.

Model Training and Evaluation We include two state-of-the-art EHR FMs: CEHR-XGPT [Pan+25b] and Many Outcome Time Oriented Representations (MOTOR) [Ste+23]. This decision was based on Pang et al. [Pan+25a], as it compares multiple FMs on CUMC data, which is similar in size and demographics as we established above. This benchmark compares an earlier version of CEHR-XGPT (CEHR-GPT [Pan+24]) with several other EHR FMs; CEHR-GPT and MOTOR outperform most other models on a wide range of clinical tasks in terms of calibration and discrimination [Pan+25b]. This is in both small and large data regimes with *linear probing* — here, Pang et al. [Pan+25a] artificially limit the training dataset to smaller subsets (data regimes) and train linear classifiers with the embeddings of this subset. We will also use to evaluate the defined tasks. Therefore, we select these two models for developing EHR FMs. The used hyperparameters can be found in the Appendix in Table E.6 and E.7.

Models are trained and evaluated using the MEDS tooling usually benefit from a predefined benchmarking framework, for example, MEDS-DEV [McD+24], which is discussed above. However, the models we benchmark in this section, CEHR-XGPT [Pan+25b] and MOTOR [Ste+23], require a large number of steps, which could necessitate custom code and parameters. They are not yet integrated into MEDS-DEV but could be in the future. We use FM embeddings of timelines, limited to the prediction point, for linear probing and training a model with all data for each task. Baselines are trained with the MEDS data itself.

Transferability At other sites, Stanford Medicine [Wor+24] and Columbia Medicine [Pan+25a], the same or similar pipelines can be defined depending on the infrastruc-

Figure 5.15: Foundation model training in the context of the Mount Sinai Health System. Hospital data is aggregated into one health system database in the OMOP 5.3 format and filtered to include only persons with visits. The OMOP-MEDS ETL "flattens" the data into MEDS while preserving key event occurrences. We split the data into training, tuning, and held-out sets beforehand, as the lengthy pretraining time would otherwise prevent us from performing cross-validation on the entire dataset for different tasks. The open-access datasets can undergo the same process due to the defined ETLs, predicates, and the existing software infrastructure built around MEDS. Then, cohorts and labels are created using *ACES* from each split. Models are trained and evaluated using the MEDS tooling. Baselines are trained with the MEDS data itself. Finally, we use FM embeddings of timelines limited to the prediction point to do linear probing and train a model with all data for each task. At other sites, Stanford Medicine [Wor+24] and Columbia Medicine [Pan+25a] have defined similar systems. All health systems can benefit from a standardized data format, ETL, cohort definition, and model training and evaluation framework.

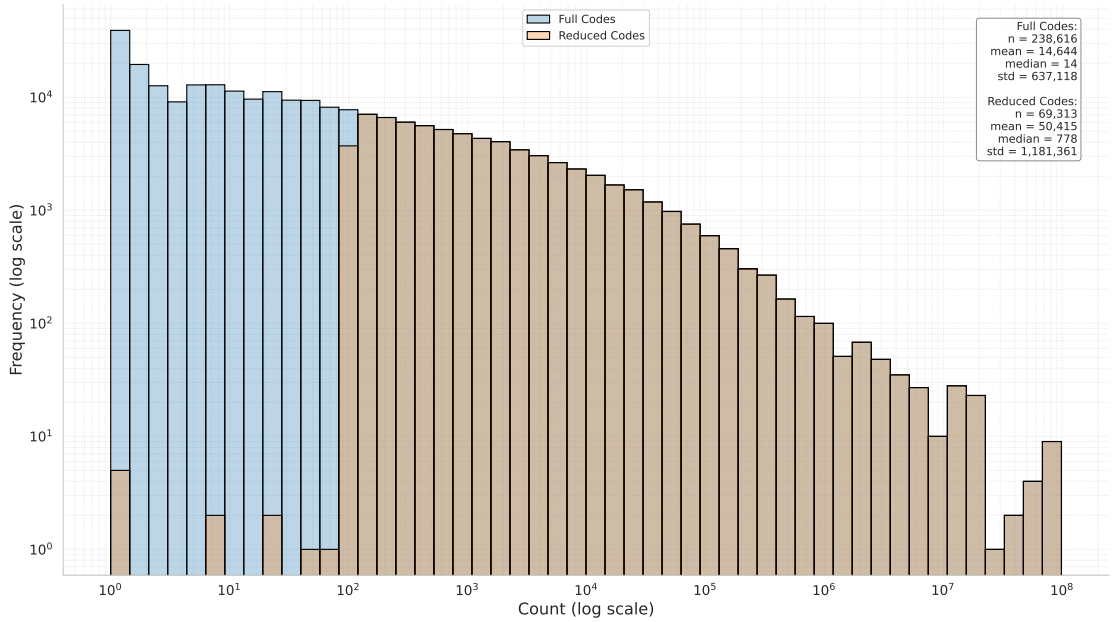


Figure 5.16: Unique code count distribution for the MEDS MSHS dataset. The code reduction adds three bins for rare codes (1-10, 11-50, 51-100 occurrences), which does not impact the overall distribution for more common codes but still reduces vocabulary size while maintaining event occurrence.

ture. This demonstrates that health AI infrastructures can be versioned and shared across health systems, which could accelerate open research and integrate methods that have been developed elsewhere. We have evidence to suggest this could be realized as the tasks above were defined across eight datasets and one health system dataset. We expand upon the code repository of Pang et al. [Pan+25a] to define the same tasks that we defined for *MEDS-DEV* benchmarking using *ACES*, which should make this benchmark more interoperable. Moreover, we touch on transferability below, where we compare tokenizer vocabularies. Note that we do not necessarily mean that we develop and employ a single model across the different health systems as they could have very different underlying EHR collection techniques and Patient Data Management Systems; future research needs to explore further what kind of FMs could be helpful in this context.

5.7.1 Tokenizer Development

Tokenization is an important consideration when training models on EHR data [Att+25]. The tokenizer converts MEDS event streams into text sequences (tokens) that are then used to train, evaluate, or fine-tune models; these models can be similar to LLM architectures as demonstrated in [Ste+21]. An existing adjustable tokenizer and LLM fine-tuning library was used to experiment with MEDS MSHS tokenization [Wor+24]. Moreover, for many MEDS-compatible models, the MEDS data is first converted into MEDS-reader databases [Ste+24], which creates a structured event-stream per subject. Then, each model has its own tokenizer that converts the event stream into tokens. Often, extra tokens are added to represent temporal information, such as the gaps between events or between individual events [Pan+21]. Some works use tokenizers that use

knowledge graphs to be able to use hierarchical relationships between codes [Hur+24; Su+25; Ste+23]. More variations have been described above in Section 5.2.5.

As can be seen in Figure 5.16, there is a large number of unique codes with fewer than 100 occurrences, with a few codes being common and many codes being rare. This is a common characteristic of clinical data, posing challenges for model training, as models need to generalize from limited examples of rare codes. To mitigate this, we categorize very rare codes (i.e., those occurring in fewer than 100 patients) into a single 'rare' code per unique vocabulary. This reduces the vocabulary size and helps the model to focus on learning patterns from more common codes, while maintaining a training dataset with as many usable codes as possible. Rare codes are converted into three separate bins: 1-10, 11-50, and 51-100 occurrences; this reduces the number of codes from 238,616 unique codes to 69,313 unique codes (a 71% reduction). Note that here, we trade off some granularity for a more manageable vocabulary size, which is beneficial for model training, though some information is inevitably lost. Future research could explore more advanced methods for handling rare codes based on text descriptions, which might lead to other challenges [Kir+25a].

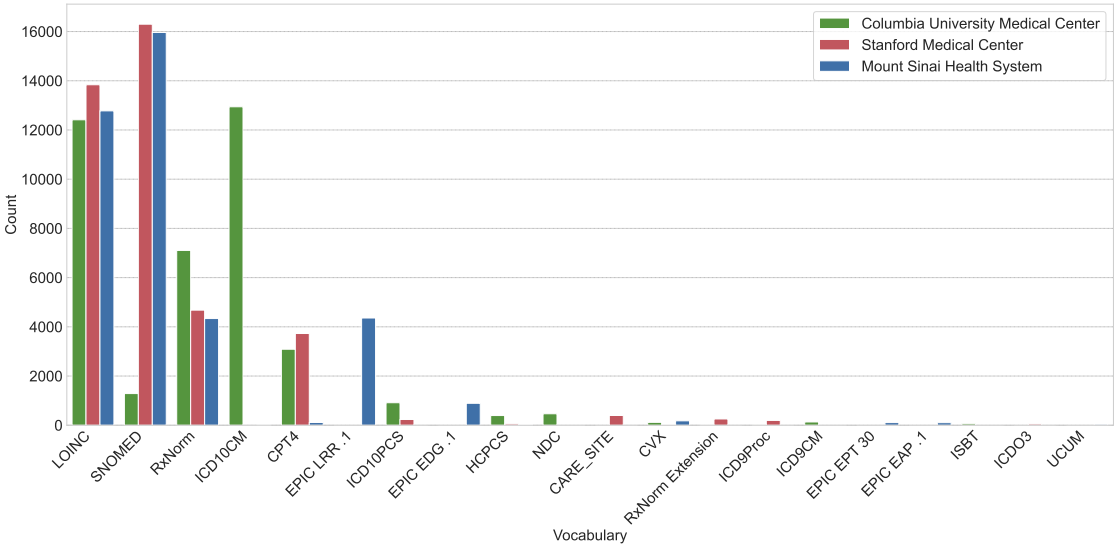
Tokenizer Vocabulary across Medical Centers

We compare the tokenizer vocabulary and individual token counts across medical centers for the models presented in Wornow et al. [Wor+24]: Figures 5.17a and 5.17b. Here, we can see that there are some differences in the coding systems used across medical centers, which is expected due to different EHR systems and coding practices. Moreover, there is likely a difference in the code mapping from the local EHR to standardized vocabularies. We could speculate that SM and CUMC have more standardized codes since no Epic codes show up in the tokenizer vocabulary. This might have advantages for transferability and generalization on the source dataset, since usually codes can be aggregated more easily. However, it is hard to make conclusive statements since we only have access to the source data from MSHS and do not know what transformation steps have been taken before training these tokenizers. This complicates the use of FMs across medical centers, though one could simply use the subset of codes that are available at the evaluation site [Guo+24; Ale+25]. Furthermore, for many models in this category, there is an inherent limit to the tokenizer vocabulary size due to model architecture constraints. Therefore, it is important to analyze the tokenizer vocabulary and individual token counts across medical centers to understand the impact of these differences on model performance. Future work could explore methods to harmonize tokenizers across medical centers by collaborating across institutional boundaries and develop more flexible tokenization strategies that can adapt to different coding systems.

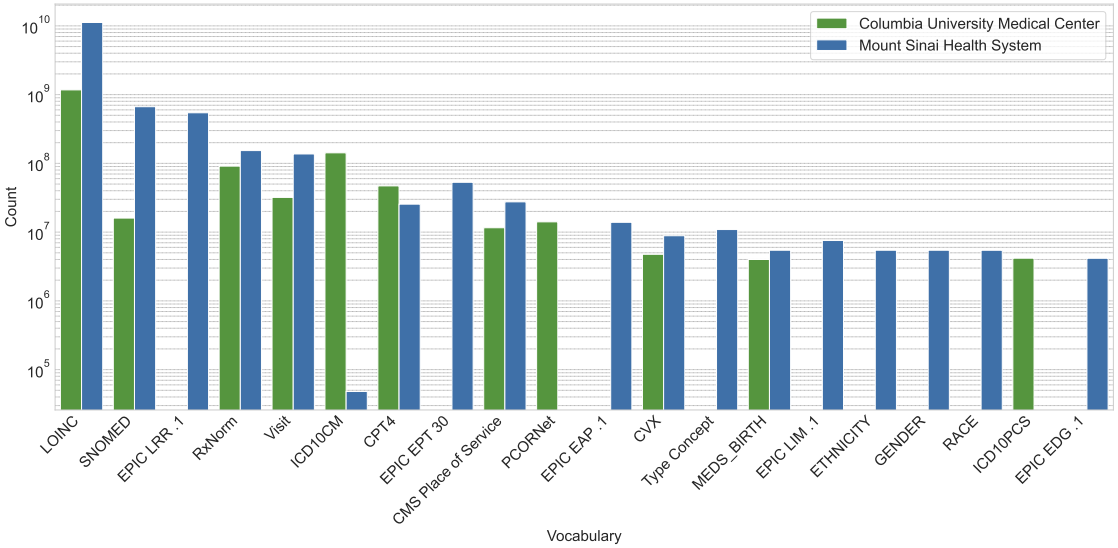
5.7.2 Sequence Length

Moreover, we analyze the sequence length distribution (Figure 5.18a) across splits. Here, we can see that the sequence length is remarkably similar across splits, with a median length of 75 tokens. We also observe that the 99th percentile is approximately 6,500 tokens, indicating that most models should be able to handle the majority of sequences without truncation. The code count distribution shows a long tail, with many

Figure 5.17: Comparing FMs at multiple centers based on the same tokenization procedure for the same tokenizer size. We limit the size to the top 39,000 tokens by occurrence.



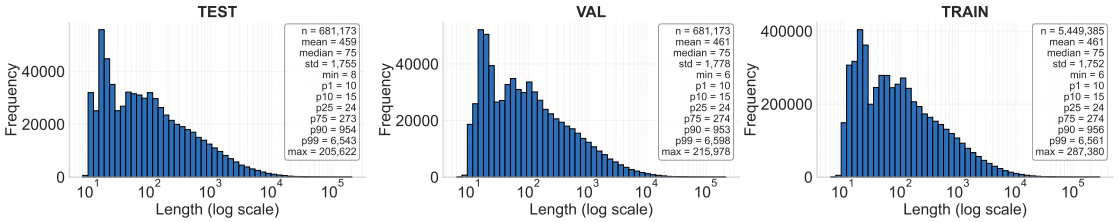
(a) Tokenizer vocabulary coding for the HuggingFace adapted auto-regressive models [Wor+24] compared across medical centers. Note that some coding vocabularies might be purposely converted or excluded due to licensing issues (e.g., Epic codes at Stanford Medicine) and to increase generalizability.



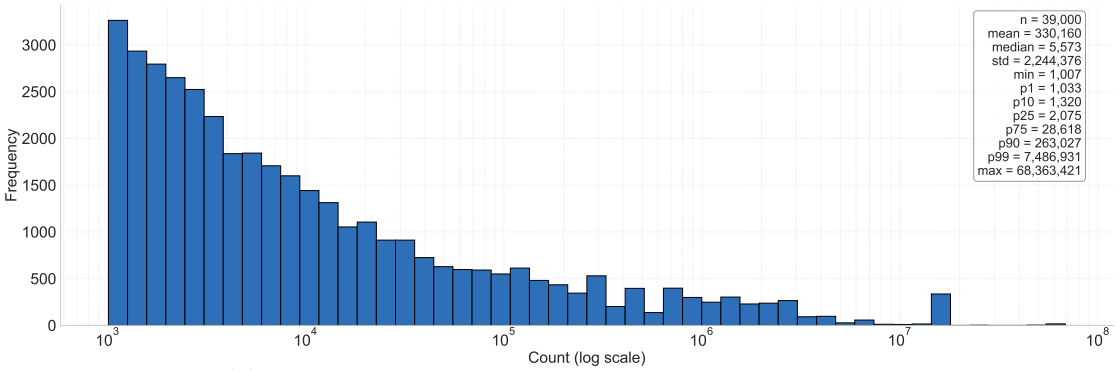
(b) Individual token amounts aggregated by vocabulary and compared across medical centers. Stanford Medicine is excluded due to this information not being available in the tokenizer configuration.

codes occurring only a few times and a few codes occurring very frequently. This is a common characteristic of clinical data, posing challenges for model training, as models need to generalize from limited examples of rare codes. The code count distribution is shown in Figure 5.18b. We observe that there are many codes with low counts, as well as some very frequent codes. This long-tail distribution is typical in EHRdata [Wor+24] and requires careful consideration during model training to ensure that rare but clinically significant codes are adequately represented.

Figure 5.18: Tokenized dataset specifics for the MSHS dataset.



(a) Sequence length distribution of MSHS dataset for each split.



(b) Aggregated individual code count distribution for the tokenizer.

5.7.3 Linear probing

We perform linear probing on the FM embeddings for each task defined in the MSHS dataset. Linear probing [TS24] involves training a simple linear classifier (logistic regression) on top of the frozen embeddings generated by the FM. This enables us to evaluate the quality of the learned representations and their capacity to capture relevant information for downstream tasks. One could also use fine-tuning as shown with CEHR-BERT [Pan+21] above. This has two main downsides: (1) it is computationally and store-wise more expensive, as the entire model needs to be updated during training, and (2) it can lead to overfitting, especially when the downstream task has limited labeled data. Linear probing offers a more efficient method for evaluating the usefulness of the embeddings, minimizing the risk of overfitting.

We use a tabular baseline for comparison from the Framework for Electronic Medical Records (FEMR) library [25] to create a tabular baseline for linear probing. FEMR automatically aggregates the event-stream data up until the prediction point. Then, we train a logistic regression and LightGBM [Ke+17] (denoted as GBM, which abbreviates Gradient Boosting Machine) using these embeddings as input features and the task labels as targets. Moreover, we perform hyperparameter tuning using the tuning set to optimize the performance of the linear classifier. A similar procedure is followed for the Foundation Models: we extract the embeddings from the FM for each prediction sample in the cohort. We essentially do the same as with the baseline, but we use frozen embeddings without fine-tuning. We refer to the models trained on the embeddings as " adapter models". The performance of the linear classifier is evaluated using domain standard metrics: Area Under the Receiver Operating Characteristic (AUROC), Area Under the Precision Recall Curve (AUPRC), Calibration, and Brier Score.

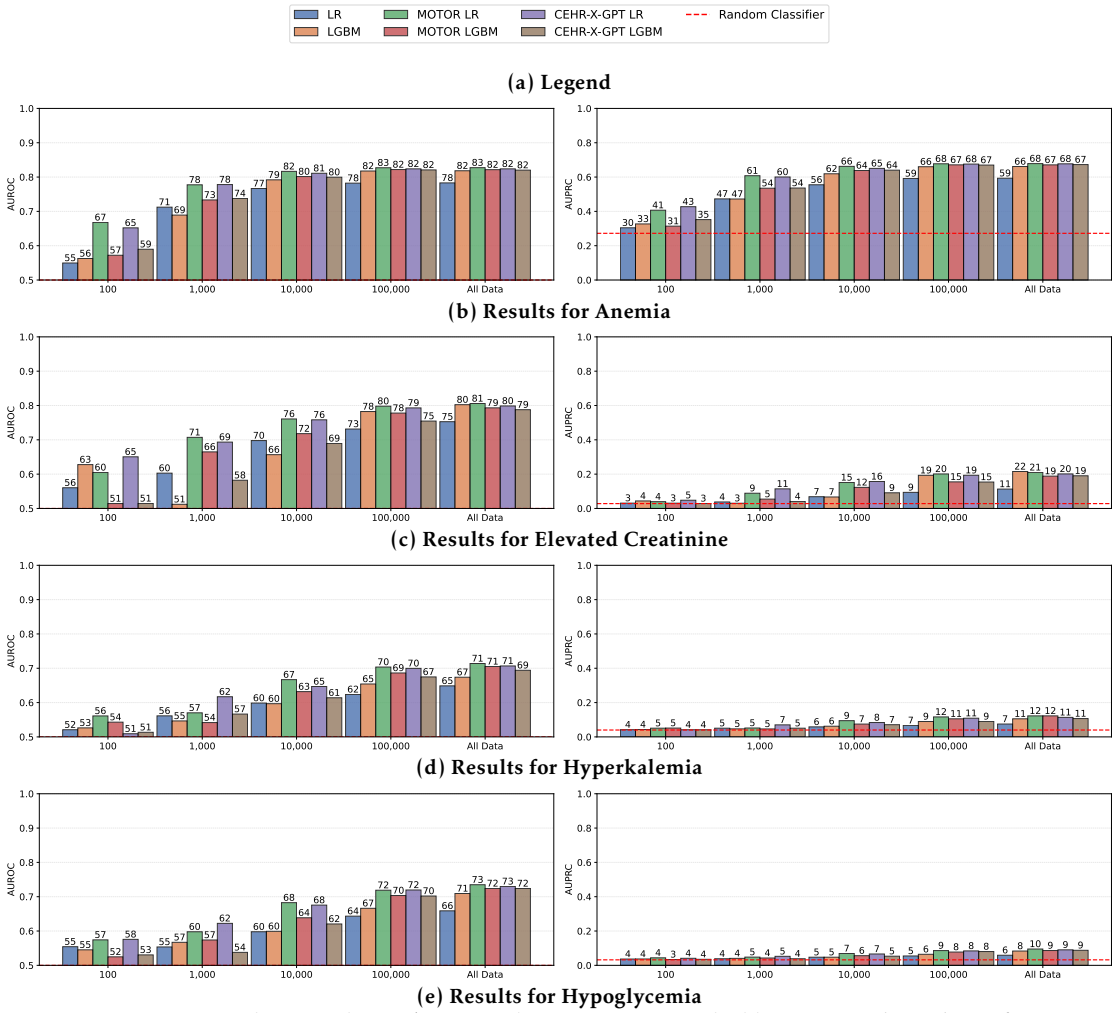
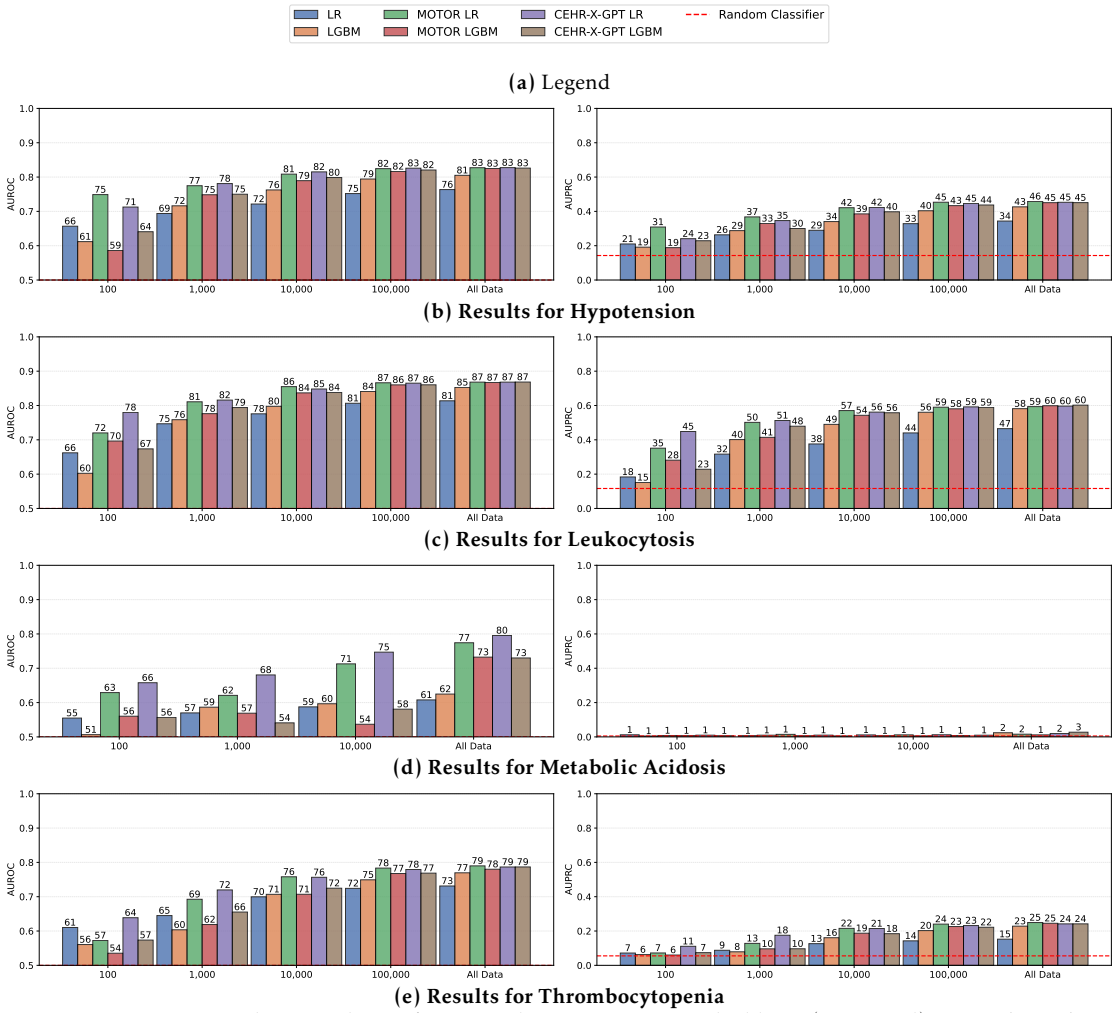


Figure 5.19: Linear probing results on the MSHS dataset using FM embeddings. We evaluate the performance of linear classifiers trained on frozen FM embeddings for various clinical tasks defined in the MSHS dataset. The results are presented in terms of AUROC and AUPRC metrics for each task.

Per-task results Figure 5.19 and 5.20 show the results of linear probing on various clinical tasks defined in the MSHS dataset. The results are presented in terms of AUROC and AUPRC metrics for each task. Unfortunately, the 30 day hospital readmission embeddings could not be extracted from MOTOR, which is why we leave it from the comparison. We can see that the FM embeddings perform well across a range of tasks, indicating that the learned representations capture relevant clinical information. MOTOR and CEHR-XGPT generally outperform the tabular baseline and demonstrate comparable performance to each other. For some tasks, such as Leukocytosis and Metabolic Acidosis, CEHR-XGPT shows a slight advantage over MOTOR, while for others, such as Hyperkalemia and Anemia, MOTOR performs better. This suggests that different FMs may be better at capturing specific clinical features or patterns relevant to certain tasks. The most prominent differences are seen, as can be expected, in low-data regimes (100 samples in particular). CEHR-XGPT performs particularly poorly in the lower data regimes of hyperkalemia compared to the baselines and MOTOR.



(a) Legend

(b) Results for Hypotension

(c) Results for Leukocytosis

(d) Results for Metabolic Acidosis

(e) Results for Thrombocytopenia

Figure 5.20: Linear probing results on the MSHS dataset using FM embeddings (continued). We evaluate the performance of linear classifiers trained on frozen FM embeddings for various clinical tasks defined in the MSHS dataset. The results are presented in terms of AUROC and AUPRC metrics for each task.

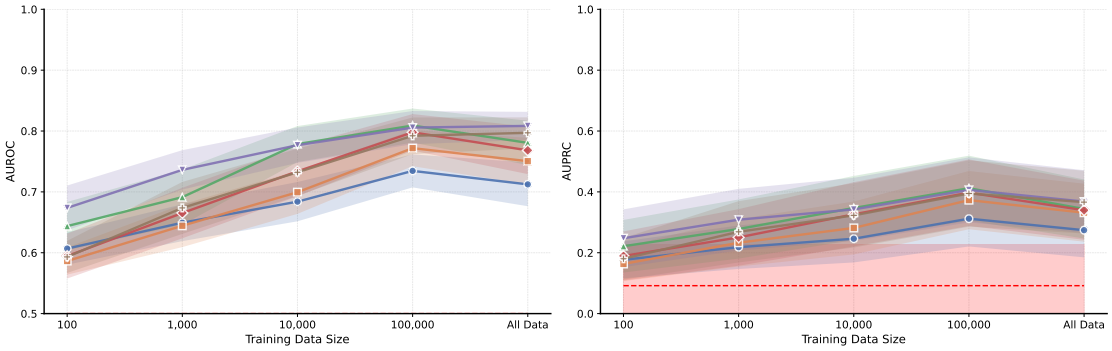
Aggregated results Figure 5.21 shows the aggregated results of linear probing across all tasks. Here, we can see that both CEHR-XGPT and MOTOR outperform the tabular baseline, indicating that the FM embeddings contain useful information for downstream tasks. This effect is more pronounced in low-data regimes, which is in line with prior work [Pan+25a].

CEHR-XGPT generally performs better than MOTOR in terms of discrimination, suggesting that it may be more effective at capturing relevant clinical features in its embeddings. Moreover, we observe that CEHR-XGPT is more sample-efficient than MOTOR, as it performs better in low-data regimes (i.e., with fewer training samples). This is an important consideration for clinical applications, where labeled data may be scarce. The calibration error for the logistic regression models is significantly higher.

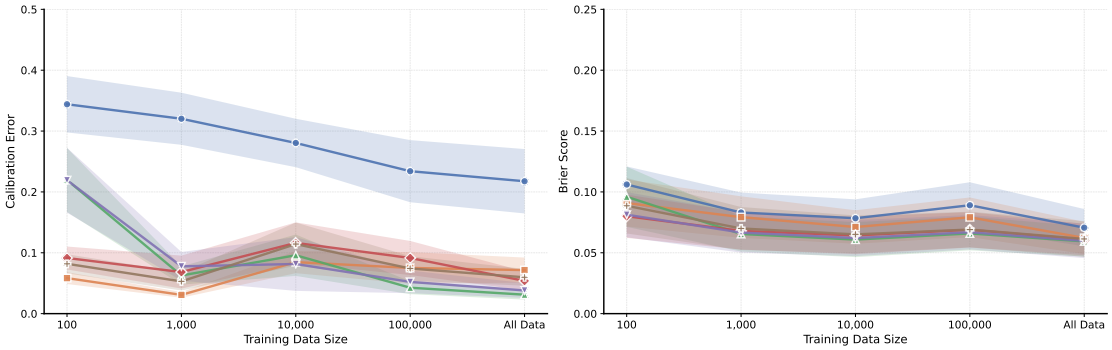
Furthermore, the adapter model logistic regression models tend to perform better than the LightGBM models in terms of discrimination metrics. This suggests that the FMs already capture non-linear relationships in the data, making a simple linear clas-



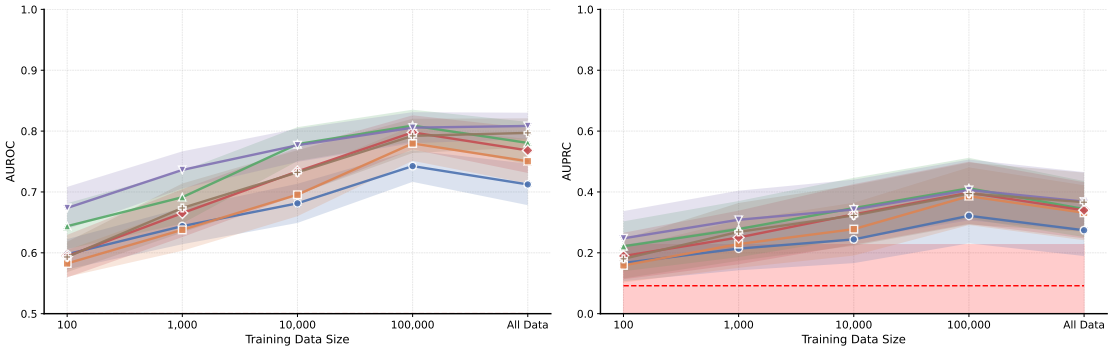
(a) Legend



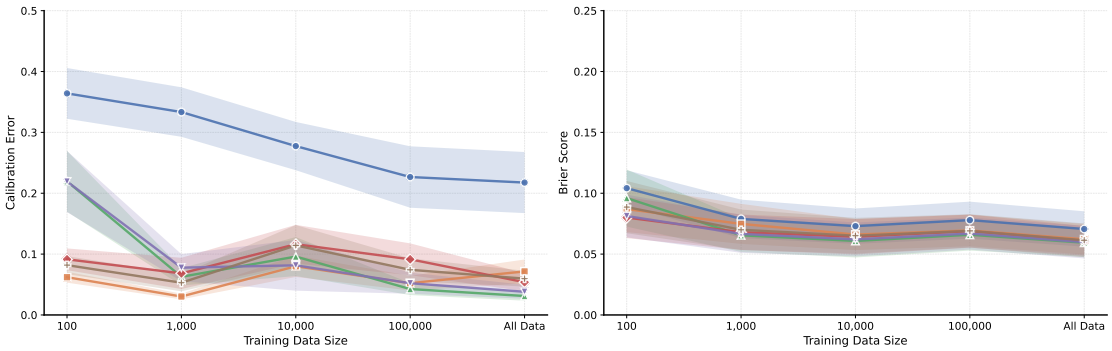
(b) Linear probing aggregated results for AUROC and AUPRC with tasks equally weighted (macro average).



(c) Linear probing aggregated results for calibration error and Brier score with tasks equally weighted (macro average).

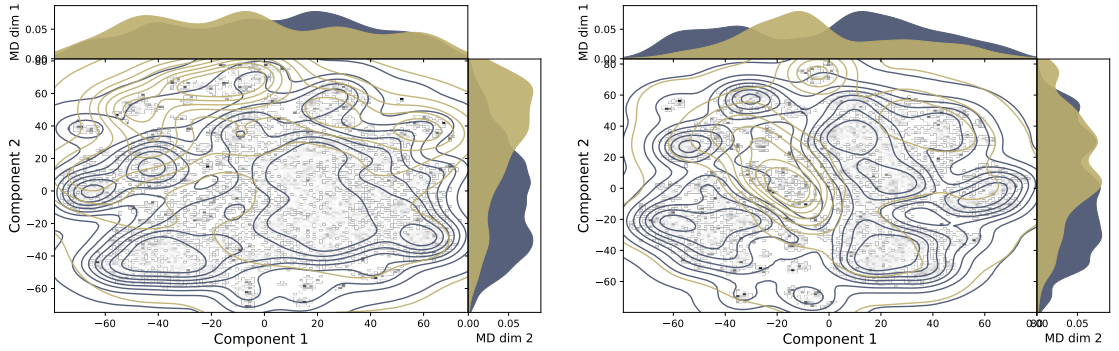


(d) Linear probing aggregated results for AUROC and AUPRC with tasks weighted according to cohort size (weighted average).



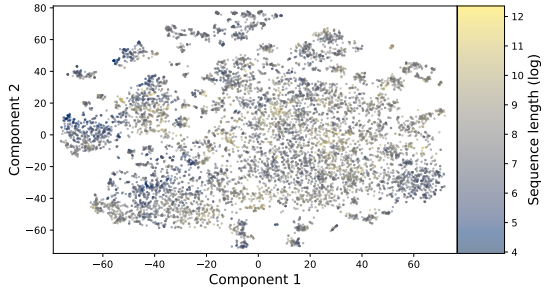
(e) Linear probing aggregated results for calibration error and Brier score with tasks weighted according to cohort size (weighted average).

Figure 5.21: Aggregated results for linear probing with FMs. We evaluate the performance of linear classifiers trained on frozen FM embeddings for various clinical tasks defined in the MSHS dataset. The results are presented in terms of AUROC, AUPRC, Calibration Error, and Brier Score metrics, aggregated using both macro and weighted averages across tasks.

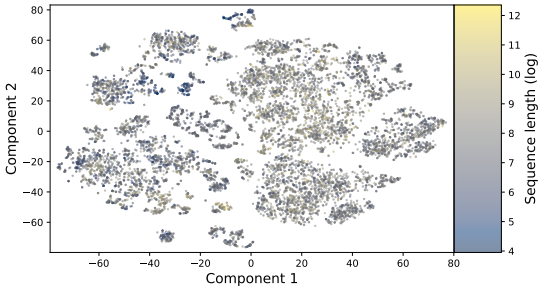


(a) MOTOR embeddings colored by task labels (positive/negative).

(b) CEHR-XGPT embeddings colored by task labels (positive/negative).



(c) MOTOR embeddings colored by sequence length.



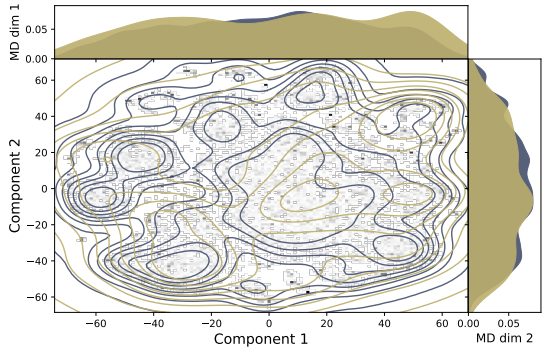
(d) CEHR-XGPT embeddings colored by sequence length.

Figure 5.22: t-SNE plots of FM embeddings for the Leukocytosis task. The top two plots show the embeddings colored by task label (positive/negative) as contour plot with a point density heatmap; a marginal density plot for the two dimensions grouped by label has also been added. The bottom two plots show 10,000 randomly sampled embedding samples colored by the sequence length. The left column corresponds to MOTOR embeddings, and the right column corresponds to CEHR-XGPT embeddings. The colormap gradient was chosen to be as colorblind-friendly as possible [K25].

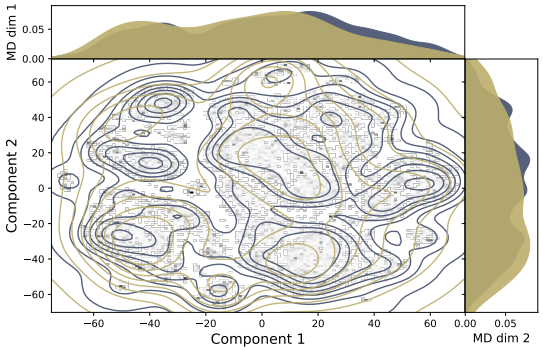
sifier sufficient, whereas the LightGBM models may overfit to the training data. However, the calibration error is lower for the LGBM models in the lowest data regime (100 samples) which may be caused by the regularization techniques in LightGBM for a low number of samples with high dimensionality (768 for both FMs); this suggests we could improve performance further by adding regularization or feature selection to the logistic regression models. Additionally, the LightGBM demonstrates familiar behavior in the lowest data regime (100 samples), where the amount of samples is much lower than the feature dimensionality (768 for both FMs), leading to especially poor generalization [McE+23]. As the number of samples increases, the LightGBM models are able to capture the underlying patterns in the data better, leading to improved performance.

5.7.4 Dimensionality reduction of FM embeddings

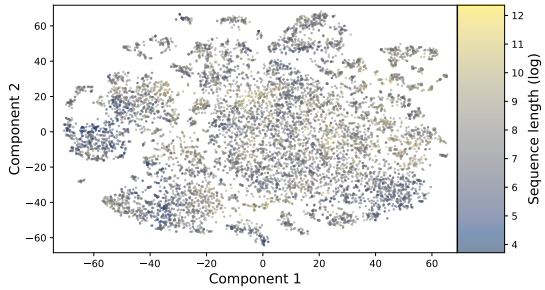
To better understand the structure and quality of the learned representations, we visualize the high-dimensional FM embeddings using dimensionality reduction techniques. Dimensionality reduction could provide further qualitative insights into how the models organize clinical information in their representation space. These visualizations serve multiple purposes: they reveal whether embeddings exhibit meaningful clustering patterns aligned with task labels, indicating that the models have learned clinically relevant features, they expose potential confounding factors that may influence model



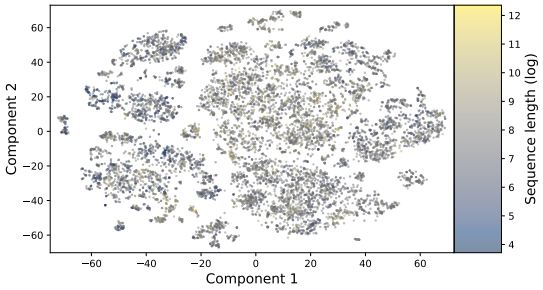
(a) MOTOR embeddings colored by task labels (positive/negative).



(b) CEHR-XGPT embeddings colored by task labels (positive/negative).



(c) MOTOR embeddings colored by sequence length.



(d) CEHR-XGPT embeddings colored by sequence length.

Figure 5.23: t-SNE plots of FM embeddings for the Hyperkalemia task. The top two plots show the embeddings colored by task label (positive/negative) as contour plot with a point density heatmap; a marginal density plot for the two dimensions grouped by label has also been added. The bottom two plots show 10,000 randomly sampled embedding samples colored by the sequence length.

behavior, and they help explain differences in linear probing performance across tasks and models by showing the degree of class separation in the embedding space. We apply t-distributed Stochastic Neighbor Embedding (t-SNE) [vdMH08], Uniform Manifold Approximation and Projection (UMAP) [MHM20], and Principal Component Analysis (PCA) to project the embeddings into two-dimensional space, enabling visual inspection of the representation structure. This qualitative analysis complements the quantitative linear probing results and provides interpretable insights into what patterns the FMs have learned from the clinical event sequences.

Figure 5.22 and 5.23 show t-distributed Stochastic Neighbor Embedding (t-SNE) [vdMH08] plots of the FM embeddings for the Leukocytosis and Hyperkalemia tasks, respectively. The tasks have been chosen because they represent the tasks with the highest and lowest AUROC, respectively. The plots show the embeddings colored by the task labels (positive/negative) and sequence length. We can see that both MOTOR and CEHR-XGPT embeddings exhibit some clustering based on the task labels, indicating that the embeddings capture relevant information for distinguishing between positive and negative cases. Moreover, there seems to be some clustering on sequence length, suggesting that the models are sensitive to the amount of information in the input sequences. However, we cannot conclusively say this seems to be a confounding factor with respect to the task labels. Appendix E.5 contains t-SNE, Uniform Manifold Approximation and Projection (UMAP) [MHM20], and Principal Component Analysis (PCA) plots for all tasks and models. Generally, we observe more separation in tasks

where the linear probing performance is higher. Moreover, there is clear clustering based on sequence length in most tasks, indicating that the models are sensitive to the amount of information in the input sequences, especially for the linear PCA plots. 3-dimensional reductions were also explored but did not yield additional insights. Future work includes exploring clustering techniques to quantify the observed patterns and investigating their downstream applications to improve clinical predictions; i.e., one could decide diagnostic procedures based on the current cluster adherence at hospital admission.

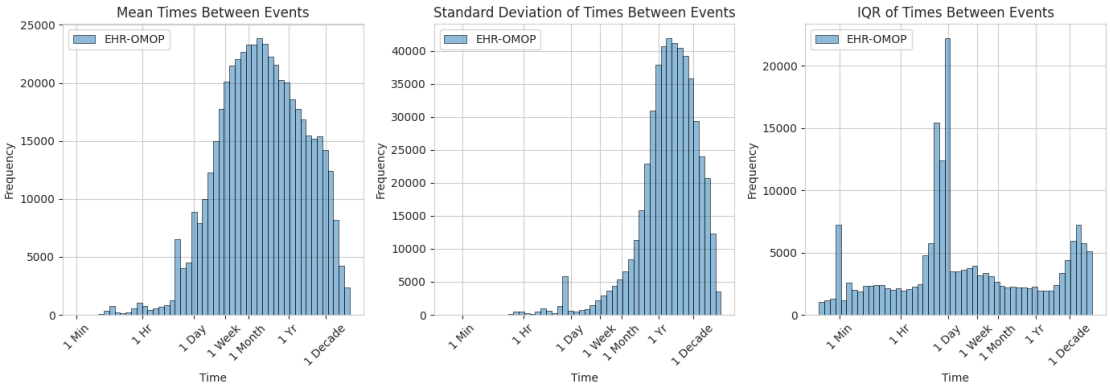
5.7.5 Similarities between EHR data and natural language

As we are using LLM architectures for EHR data, there is a benefit in investigating how the EHR we train on may differ from natural language. Steinberg et al. [Ste+21] was one of the first to note the applicability of natural language processing (NLP) as generalized FMs. Moreover, we draw on the work of Wornow et al. [Wor+24] — which claim that longer context windows can improve model performance for EHR just like it does for NLP — and closely follow their experimental procedure to explore the similarities between EHR data and natural language on a bigger dataset in a different area: New York City, New York, USA, instead of Palo Alto, California, USA. The goal is to see whether the findings by Wornow et al. [Wor+24] generalize to other health systems and datasets and to inspire new approaches to address the differences between the two domains.

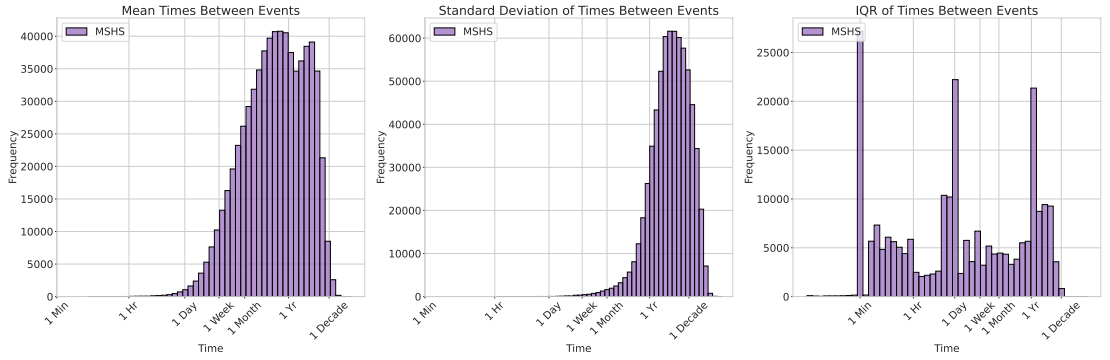
Event irregularity Firstly, we note that EHR has an inherent temporal aspect to it. As discussed earlier, models accommodate this aspect in different ways. By analyzing the event irregularity [McD+23], i.e., the variability in time between events, we can see how the events are distributed across each patient's timeline. Figure 5.24 compares the irregularity between two health systems: SM and MSHS. We can see that SM has a different temporal profile than MSHS: the latter shows two peaks in mean inter-event times, which are approximately 6 months and 5 years. Moreover, we observe that the interquartile ranges also exhibit three distinct peaks in the MSHS dataset, compared to a more concentrated peak in the SM data.

N-gram repetition rates We analyze the n-gram repetition rates [Wor+24] for both datasets in Figure 5.25. N-grams are contiguous sequences of n items from a given sample of text or speech. In this case, we analyze the sequences of clinical events in the EHR data. By comparing the n-gram repetition rates between EHR data and high-quality natural language text (Wikipedia articles), we can assess how similar the structure of EHR data is to natural language. This is likely caused by the billing and documentation practices in healthcare [WL14] as noted by Wornow et al. [Wor+24].

Unique word occurences Zipf's law is an empirical principle that states that when a set of measured values is arranged in decreasing order, the value of the n-th entry is often approximately inversely proportional to n. A well-known application of Zipf's Law [Pia14] is that the frequency distribution of words within a text or a corpus of natural language adheres to this principle: $\text{word frequency} \propto \frac{1}{\text{word rank}}$. We show Zipf's



(a) Event irregularity distribution for the 0.5M tuning split in the SM data. From Wornow et al. [Wor+24] used under CC BY 4.0 license (`https://creativecommons.org/licenses/by/4.0/`).



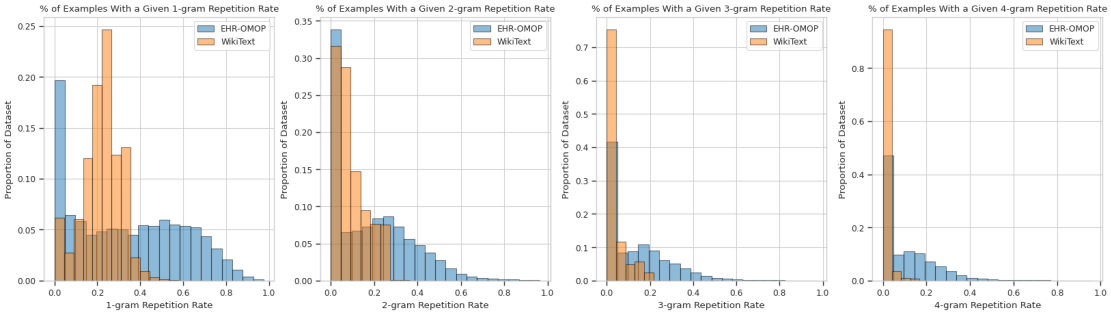
(b) Event irregularity distribution for the 0.68M tuning MSHS dataset.

Figure 5.24: Event irregularity distribution for the SM and MSHS datasets. The event irregularity differs between the two health systems. From left to right: the mean time between events, the standard deviation of time between events, and the interquartile ranges of times between events.

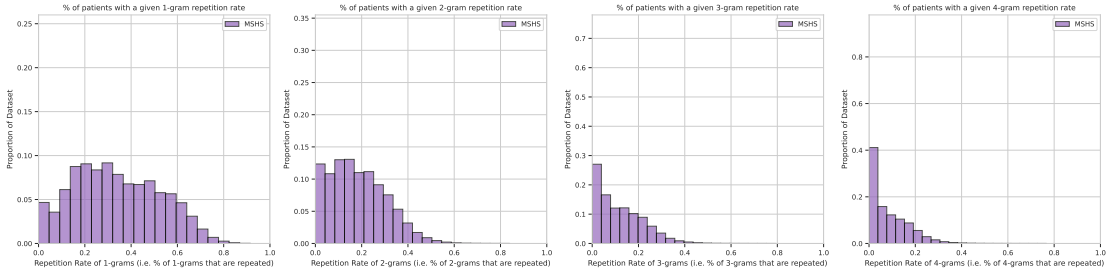
Table 5.6: Fit statistics for Zipf's law analysis in Figure 5.26. Fit statistics, including the power-law exponent (α) and R^2 values, are displayed for each dataset. Here, a clinical code is treated as a "word"."

Source	α	R^2	Unique "words"	Total "words"
MSHS	2.817	0.946	238,616	3,494,381,546
The Pile Uncopyrighted	1.229	0.964	3,155,979	193,652,926
WikiText-103	1.710	0.981	525,727	69,239,169
Homer's Odyssey	1.242	0.970	7,248	89,099
James Joyce's Ulysses	1.009	0.966	29,834	213,212

law adherence in Figure 5.26, where we compare natural language in several forms to MSHS. We plot the law for two books, Homer's Odyssey (translated to English) [Hom15] and James Joyce's Ulysses [Joy22] and two datasets: WikiText-103 [Mer+16] (also described in Wornow et al. [Wor+24]), containing a large amount of English wikipedia articles, and a subset of the Pile Uncopyrighted dataset [Gao+20], which contains a large amount of mainly academic and professional texts. We observe that a sequence of codes occurs very frequently for MSHS events, making the distribution "flatter" but falling quickly towards the lower ranks. Moreover, the fit is worse, in terms of R^2, as seen in Table 5.6 than the natural language resources. This reinforces the higher repetition rate observed from the comparison of n-gram repetition rates.



(a) N-gram repetition rates for the SM dataset and Wikitext. From Wornow et al. [Wor+24] used under CC BY 4.0 license (`https://creativecommons.org/licenses/by/4.0/`).



(b) N-gram repetition rates for the MSHS dataset. Limits have been adjusted to match the SM figure for easier comparison. Note that the y-axes differ per n-gram.

Figure 5.25: N-gram repetition rates for EHR data compared to high-quality natural language text (Wikitext [Mer+16]). The EHR data clearly shows higher repetition rates across all n-grams. MSHS exhibits a similar trend to SM, but with a greater spread of repetition rates.

Adapting to EHR characteristics We observe that EHR data exhibits distinct characteristics compared to natural language, including higher event irregularity, increased n-gram repetition rates, and a different adherence to Zipf's law. These differences highlight the need for specialized modeling approaches when applying LLM architectures to EHR data. Future work could explore techniques to better accommodate these unique properties. For example, Huang et al. [Hua+25] proposes multi-gram tokenization for LLMs to effectively tokenize common repeating patterns; this might especially benefit these NTP FMs. Moreover, the temporal aspect of EHR can be addressed in a variety of ways and past benchmarks seem to suggest this could make a big difference [Yu+25; Ste+23]. Finally, to better capture the long-tail of rare events in EHR data, one could explore techniques to enhance the representation of infrequent codes, such as with HarmonizEHR or embedding methods [Kir+25b; Heg+25].

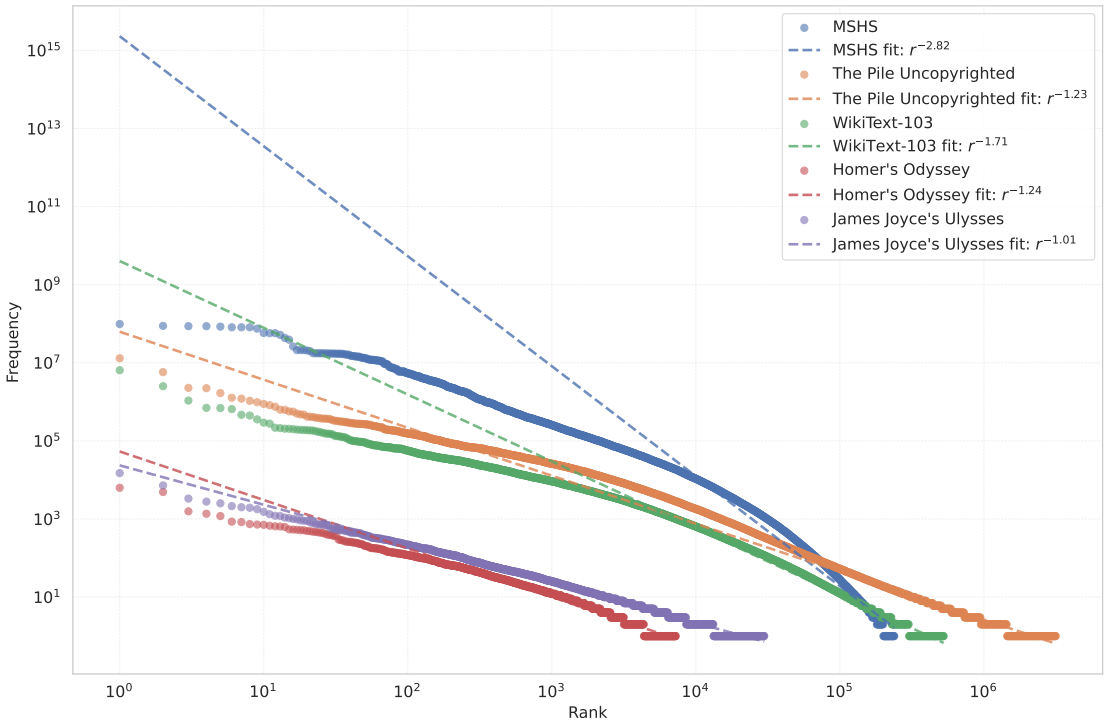


Figure 5.26: Zipf's law: comparison of medical codes in the complete MSHS dataset (6.8M subjects) to natural language corpora; this illustrates the rank-frequency relationship for four datasets: MSHS medical codes, the first 300,000 documents from the Pile Uncopyrighted dataset [Gao+20], WikiText-103 [Mer+16] corpus, Homer's Odyssey, and James Joyce's Ulysses. The log-log scale highlights the power-law behavior, with fitted lines demonstrating the approximate inverse proportionality between rank and frequency. Subsampling was applied to reduce the number of plotted points while preserving the shape of the distribution.

5.8 Discussion

In this chapter, we used MEDS as a vehicle to build modular tooling that addresses several challenges in clinical machine learning. We demonstrated that MEDS can be used for thorough data exploration purposes and trained several models on the data, ranging in complexity levels; here, it enables efficient model training and evaluation across multiple datasets and health systems. We defined eleven cohorts across eight open-access datasets, spanning three continents. Of these tasks, seven were adapted from MEDS-DEV. We defined tasks on the Mount Sinai Health System (MSHS) dataset and demonstrated that these tasks can be defined, with minor adaptations, on other datasets in a standardized manner. Moreover, we used MEDS to train two state-of-the-art EHR foundation models, one of the largest multicenter health system EHR datasets to date (c.f. [Wax+25]) — MSHS — and evaluated these models on multiple clinical prediction tasks across multiple datasets. Below, we discuss the implications of our work, its limitations, and future directions for open tooling ecosystems and EHR foundation models in particular. We also briefly discuss initial steps towards building a clinical modelling infrastructure for an entire health system.

5.8.1 Adoption of MEDS

We describe MEDS as adapted from the most recent survey described in the McDermott et al. [McD+25c] manuscript. Standards live and die by their adoption by the community; this means they must prove useful over time. MEDS presents substantial opportunities for advancing health AI research through its emphasis on simplicity and algorithm transportability. The ecosystem has already demonstrated remarkable efficiency gains, achieving computational speedups ranging from 1.9x to approximately 40,000x compared to existing preprocessing solutions [Ste+24]. At the same time, MEDS-Tab reduces CPU memory and runtime by factors between 7x and 300x [Ouf+24a]. These performance improvements enable researchers to iterate more rapidly and work with larger datasets that were previously computationally prohibitive. Furthermore, MEDS has proven useful in developing frontier foundation models for EHR data, including CEHR-XGPT [Pan+25b], EHRMamba [Fal+24], and ETHOS-ARES [Ren+24], demonstrating its capacity to support cutting-edge research at scale.

The Free and Open Source Software (FOSS) ecosystem emerging around MEDS offers another significant opportunity. With adoption across 19 institutions, 17 supported datasets — out of which 8 with open-source ETLs [VRR25; van25d; Van25d; Van25b; vdWM25; van25e; MX25; VZM25], and 11 published AI algorithms [Gro24], MEDS enables collaborative development and knowledge sharing across the health AI community. The data-as-interface design philosophy ensures accessibility for new contributors without requiring adoption of proprietary frameworks, while specialized tools for task specification [Xu+24], high-performance training [Ste+24; McD+25a], and model evaluation, provide a comprehensive infrastructure for the entire modeling pipeline. This ecosystem approach mirrors successful paradigms in general AI, such as Hugging Face [Wol+20] and TorchVision [MR10], potentially accelerating scientific progress through shared engineering efforts and reproducible benchmarking initiatives like MEDS-DEV [McD+24].

Despite these opportunities, MEDS as a standard faces several important limitations. The standard currently provides limited support for additional data modalities beyond structured EHR events, potentially restricting its applicability as AI systems increasingly integrate multimodal information from clinical notes, imaging, and waveform data. While MEDS emphasizes algorithm transportability and remains agnostic to data harmonization, this design choice means that researchers must still address the substantial challenge of semantic interoperability when comparing results across different vocabularies and coding systems. Although metadata files can link codes to external ontologies, the burden of harmonization ultimately falls on individual research teams.

The minimal schema, while promoting simplicity, may prove insufficient for capturing specific contextual nuances important for clinical applications. MEDS provides limited mechanisms to distinguish between different providers, care settings, or institutional contexts within datasets, information that may prove crucial for modeling care delivery patterns or understanding practice variation. Additionally, the current specification could benefit from enhanced support for enriching dataset metadata through external ontologies, particularly as the field moves toward more sophisticated knowledge representation and reasoning capabilities. As MEDS continues to mature, addressing these limitations through careful community-driven evolution will be essential to

maintaining its relevance as health AI research advances into increasingly complex and integrated applications.

5.8.2 EHR foundation models

In this chapter, we introduced MEDS as a vehicle for building foundation models on EHR data. Next-token-prediction models show promising results on several clinical prediction tasks across multiple datasets. They seem to perform better than most baselines for probing sizes larger than 100. However, we believe that the primary value lies in the ability to serve as a foundation for more specialized models that cater to specific clinical needs. One could think of fine-tuning these models for specific clinical tasks, such as predicting sepsis onset in the ICU. Moreover, a general risk-prediction stratification could be made at the time of admission, which could then be used to guide clinical decision-making; however, careful analysis of the biases in the data and model is necessary before deployment in a clinical setting. The compromises on tokenizer size and the aggregation of rare events could lead to overgeneralization on minority groups. The right level of granularity is necessary to ensure that the system that uses these models provides actionable, non-obvious insights to clinicians.

Moreover, models may need to be adapted to be more flexible, allowing us to utilize datasets that combine multiple health systems. Then, the tokenizer needs to be able to handle different coding systems, vocabularies, and prevalence rates. An option is to use free-text embeddings of medical concepts [Ras+21b], provided by biomedical language models, to represent medical events in a continuous space [Kat+24]. This could allow for better generalization across different coding systems and vocabularies. This has been done before, however, not with next token prediction [Kir+25a; Heg+25]. This approach also presents its own challenges, as the model must be able to output these embeddings, and they must be translated back into human-readable concepts in the health system or clinic-specific coding style for downstream use.

Once one or multiple reliable EHR foundation models are trained for a health system, they can be used as a part of a clinical prediction system. The model can also be fine-tuned for specific tasks relevant to a particular clinical setting.

Model Performance Evaluation From the results presented in Section 5.7.3, it is clear that the application of EHR foundation models holds promise in lower data regimes. This indicates that these models can be used for smaller cohorts where high-quality labelled data is scarce. While CEHR-XGPT demonstrated the strongest performance overall, this was not consistent across all tasks. This further underscores the importance of (continual) evaluation of foundation models for specific clinical tasks. Ensemble methods could be employed to gain robustness across tasks at the cost of computational and storage efficiency.

Dataset biases The MSHS dataset used likely contains several biases that may affect the results. We have tried to map this with the socioethnic data. However, there are likely other biases that we have not accounted for. For example, the dataset may contain biases related to the types of patients that are more likely to be admitted to the hospitals

in the health system, or the types of treatments that are more commonly used. These biases could affect the generalizability of the models trained on this data. Future work should investigate these biases further and develop methods to mitigate their effects.

5.8.3 Extending EHR Foundation Models to Other Modalities

Integrating genetic information, imaging data, and more free-text notes could improve the performance of auto-regressive EHR foundation models. The type of modality fusion has to be carefully considered based on data quality, training efficiency, and clinical relevance. Future work should investigate the best way to integrate these modalities into the MEDS framework and evaluate their impact on model performance. There are already datasets that contain multiple modalities. For example, the SICdb dataset [Rod+23b] contains high-resolution waveform data alongside structured EHR data. Future work could investigate how to flexibly integrate these modalities into MEDS models, allowing researchers and engineers to make modeling choices.

5.8.4 Future Work

Future work for MEDS includes the incorporation of more modalities, such as imaging and free-text notes. Moreover, the tooling ecosystem needs to be expanded to better meet the needs of the clinical AI research community, which includes a variety of community-contributed tasks, models, and evaluation metrics, as well as improved documentation and support.

5.8.5 Building Clinical Modelling Infrastructure for an Entire Health System

Several “building blocks" are needed to prepare a hospital and health information system that leads to improved patient outcomes and cost efficiency through the use of AI.

- An integrated hospital information system in which historical electronic health records are found. Optionally, for many decision systems, real-time streaming data that consists of i.a., vital signs, ward movements, and imaging, are necessary.
- Compute infrastructure that can handle data-processing, model-training, and model inference needs as newer methods become available
- A software layer that enables machine learning specialists, medical data scientists, and, in some cases, even clinicians to get from collected data to a deployed clinical support tool.
- Involvement of clinicians to deploy the AI solutions that fit their needs.
- Effective model evaluation and maintenance infrastructure to ensure that models remain safe and effective over time and adapt or update them as needed.

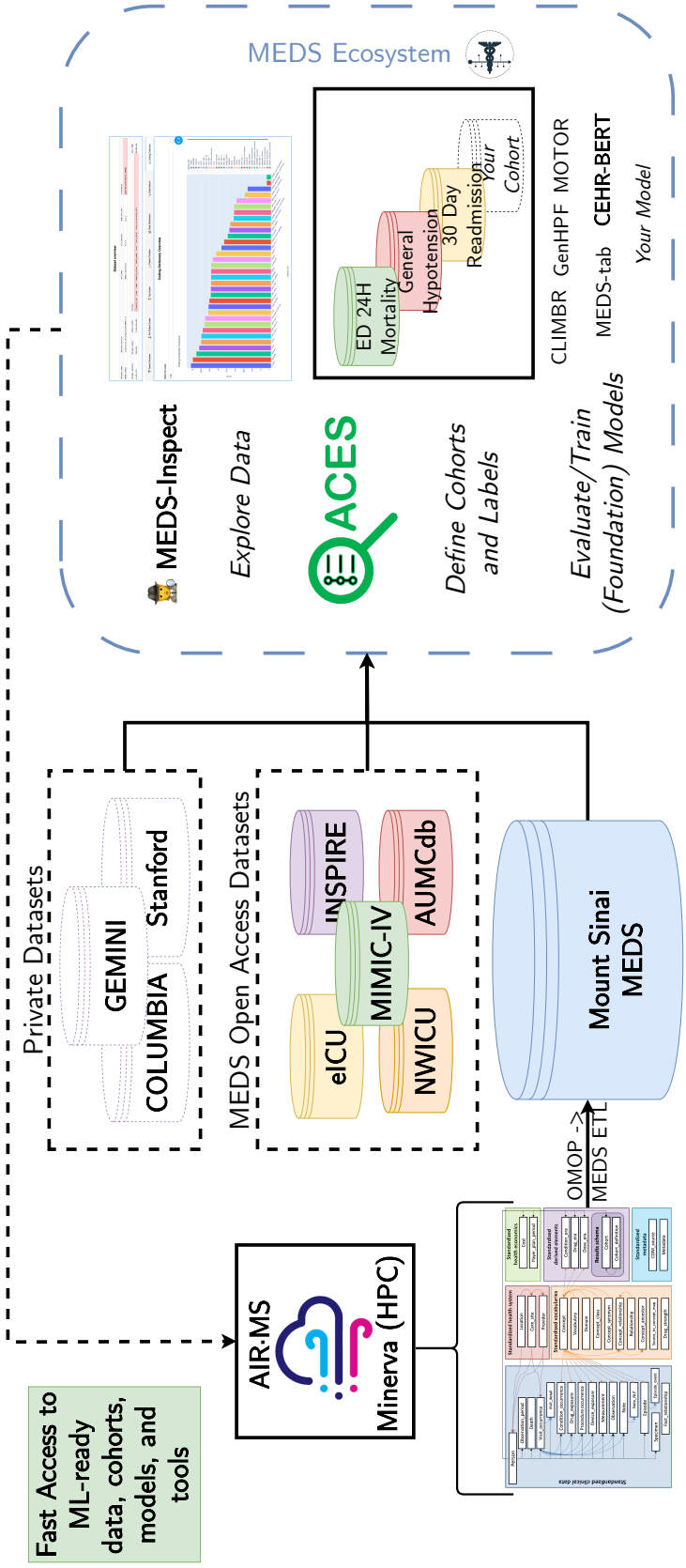


Figure 5.27: Architecture for creating a low-barrier Mount Sinai AI ecosystem. Structure resembles a simplified version of Figure 5.15 for AI experiments. OMOP serves as the backbone of Mount Sinai's observational data science efforts. MEDS complements this by providing an event-based data standard and tooling ecosystem that enables efficient model training and evaluation. The MEDS ecosystem facilitates adaptation from tooling that works on large health systems, making it straightforward to enable rapid model development and deployment.

The Mount Sinai Health System is already well-equipped to start this process with the partial compliance of items 1-4. Figure 5.27 demonstrates a simplified version of how the MEDS ecosystem could be integrated into the existing MSHS data infrastructure to facilitate AI model development and deployment. The MEDS codebases, further developed and maintained by key ML4H researchers across the world, could be adapted as an application layer to the AIR·MS environment [Gue+25]. This would be a modular system that can be extended internally or externally as needed. One can take advantage of existing tooling developed by the community, and community-driven improvements can be incorporated. Moreover, maintaining internal configuration files and a limited amount of internal software infrastructure is desirable to safeguard the private health data in MSHS or any MEDS-converted MSHS data. With our work, we aim to lower the barrier to entry for conceiving, developing, and deploying AI tools with clinical impact.

6

Discussion and Outlook

This dissertation presents various methods and applications of machine learning in healthcare, focusing on clinical prediction models using Electronic Health Record (EHR) data. First, Chapter 1 introduced this thesis by outlining its contributions. Chapter 2 introduced the context and preliminaries. In Chapter 3, we introduced *Yet Another ICU Benchmark* (*YAIB*), a modular framework for developing and evaluating machine learning models on EHR data. Chapter 4 demonstrated the application of continuous multimodal AI using wearable vital signs to predict postoperative complications in the general ward. Finally, in Chapter 5, we presented the Medical Event Data Standard (MEDS) and a variety of tools — ETLs, MEDS-Inspect, HarmonizEHR, and *MEDS-DEV* — a decentralized and extensible ecosystem for developing healthcare prediction models across multiple institutions. Moreover, we showcased the use of MEDS in training and evaluating foundation models on EHR data on the Mount Sinai Health System (MSHS).

6.1 Ethical Considerations

The use and research of automated tools, AI, and, by extension, ML in healthcare raise ethical considerations. It is a shared responsibility of the community involved in this field to ensure that these technologies are used responsibly and equitably. This goes beyond the data protection laws outlined in the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR). Here, we highlight some key ethical considerations to consider when developing and deploying machine learning models in healthcare.

Misaligned Interests Incentives to deploy AI could be driven by an increase in profits, rather than improved patient outcomes or staff well-being [Gaf+25]. For example, Jang [Jan25] describes the case of insurer UnitedHealth's "nH Predict" denied coverage for a large number of elderly patients and was accused of intentionally using a faulty AI model to increase profits.

Model Failure Modes Even without alleged harmful intent, there are several cases where pre- or post-deployment evaluation has revealed erroneous model behavior. For

example, missing model calibration, such as in the Epic Sepsis model case uncovered in 2021 [Won+21]. Moreover, training data can contain biases that lead to unequal treatment of different patient groups [Bor+23]; Obermeyer et al. [Obe+19] showed that an algorithm, arguably not classified as AI, widely used in US hospitals, exhibited significant racial bias towards black patients as they historically receive less care, leading to disparities in care. LLMs introduce new types of failure modes, such as hallucinations, where the model generates plausible but incorrect information [Asg+25]. Transparency and explainability can detect or mitigate some of these issues, but one needs to consider monitoring fairness and performance over the entire lifecycle of the model [Has+25].

Governance Regulatory bodies, such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA), have begun to develop frameworks for evaluating and approving AI-based medical devices; however, the exact requirements are often vague and vary between individual implementations [Fen+25]. Moreover, there is doubt that all incentives are aligned to ensure safe and effective AI deployment in healthcare [van+25a]. Regulatory bodies worldwide will continue to evolve their frameworks to address the unique challenges posed by AI applied to healthcare.

Environmental impact Finally, the environmental impact of training and use of (large) ML models has become a growing concern in recent years. The computational resources required to train and deploy these models can lead to significant energy consumption [HGV25] and water usage [Li+25]. Within healthcare, the use of AI has the potential to improve efficiency by optimizing the use of existing resources, such as diagnostic tools, in a targeted manner. However, for this and other well-intentioned applications, this can turn into a "catch all" argument to justify using and deploying models with ever more parameters, which use more resources [Zha+25a].

6.2 Future Work

Based on this dissertation and the current state of the field, we identify several areas for future research and development in machine learning for healthcare.

Open Science and Reproducibility Multiple works have claimed an issue or even crisis of reproducibility in ML/AI (or even because of AI) [Pin+21; McD+21; KN23; Bal23]. A central issue is a lack of incentive structures to promote open science [KN23; Bal23]. It seems that authors are often reluctant to share code and data, which hinders reproducibility and slows down progress in the field [BMG20]. Several reasons exist for this, including legal constraints, commercial interests, and the additional effort required to document and share code and data. Moreover, if researchers do share code and data, it is often not straightforward to reproduce results due to (partially) missing code, documentation, dependencies, or data access issues [BMG20]. Future work should focus on developing incentive structures that encourage researchers to share their code and data openly. These could include recognition and rewards for open science practices (such as releasing Free and Open Source Software (FOSS)), as well as the development and enforcement of standardized protocols for sharing code and data [Mab+25]. Major ML

conferences and journals have become aware of this issue and developed (mandatory) checklists [Pin+21; Col+24; Kap+24]. For *YAIB*, the *machine learning Reproducibility Checklist* is listed in Appendix C.8, in addition to the hyperparameters in Appendix C.7. The latter, a TRIPOD+AI checklist for the Clinical ASSist AND aleRt Algorithms (CASSANDRA) study, is found in Appendix D.6 and contains highly detailed information about the study and the developed solution.

Standardized benchmarks and datasets: There is a need for standardized benchmarks and datasets for evaluating machine learning models in healthcare. Many studies use different datasets and evaluation metrics, making it difficult to compare results across studies. While progress has been made in this area, for example, through various frameworks and benchmarks [Jar+21; Yan+23a; Wor+23a; Van+24a; McD+24] and datasets, there is no one-size-fits-all solution. Many benchmarks, benchmarking frameworks, and datasets for ML and AI require careful consideration of trade-offs between generalizability and flexibility. Moreover, due to the highly competitive nature of ML research and the limited resources available, researchers often quickly move on to the next endeavor. Deployability is often considered a "pie in the sky" (unattainable) due to the considerable effort and resources required to develop and get approval for AI systems. Future work should focus on developing standardized benchmarks and datasets that are widely accepted by the research community and regulators. The process to determine these should involve collaboration between researchers, clinicians, funding organizations, companies, and policy makers to ensure support and adoption [Mab+25]. In this regard, ML for health could learn from other fields, such as computer vision [MR10] and natural language processing [Wol+20], which are not without limitations [Reu+24], but where standardized benchmarks and datasets have been instrumental in driving progress [Don24].

Data provenance: The quality of healthcare datasets is a factor that can significantly impact the performance and generalizability of machine learning models [Sch+24b]. Data provenance is often hard to track. However, in healthcare datasets, this is a critical consideration [Joh+23a]. For example, if the accuracy of timestamps is unknown due to retrospective data entry, the training and evaluation of temporal models can suffer from data leakage [KN23]. Moreover, healthcare data is often collected for administrative and clinical care purposes (Real-world Data), rather than for developing machine learning models. As a result, the data may be incomplete, inconsistent, or biased. Assumptions on the provenance of the data used need to be clearly documented and (automatically) verified where possible; this takes the initiative of researchers, engineers, and institutions to work together to recognize the importance of data quality and provenance.

Feature standardization: In this work, we recognize that transformations to data are often needed to, for example, simplify modelling or improve the amount of (training) data available, and we employ several of them. To an extent, there is a trade-off similar to that of precision and recall: more standardization can lead to more data and potentially better model performance, but it can also lead to loss of important information and nuances in the data [DHV15]. We have seen this in Chapter 5, where the Mount

Sinai Health System data has a long tail of specific codes that are 1) not standardized (i.e., mapped to a standardized vocabulary) due to a lack of resources 2) have not been grouped due to a lack of clinical consensus on how to do so or 3), where grouping leads to a loss of information deemed unacceptable to the mapping team. This list is non-exhaustive as there could be other reasons that a mapping is incomplete. While a growing body of work exists on the standardization of healthcare data, there is a need for more transparency and dialogue on the trade-offs involved in standardization. While perhaps not the most exciting research topic, it is an essential consideration for the successful application of machine learning in healthcare [Hen+25]. For example, consider the use of off-the-shelf models from companies like Epic [Wax+25]. Here, the mapping of local data could be the difference between a successful or failed deployment.

The road to deployment: There is a significant gap between health care sites when it comes to AI-tool deployment. While some countries, states, or hospitals have the resources, infrastructure, expertise, and legal framework to deploy models, many others do not (yet). Different parts of the world face varying regulations [Sch+24a; 24b] and different levels of maturity [Cas+24] in terms of digitalization and necessary infrastructure. Future work should make model deployment strategies agnostic to these differences as much as possible. Finally, to ensure we gain and retain support from clinicians, patients, and policymakers, there needs to be more emphasis on how models are deployed and monitored in clinical settings. For example, simultaneously improving healthcare policy communication to patients along with system deployment leads to a more trustworthy deployment of AI in healthcare [Van+23].

6.3 Recommendations

Based on the findings and experiences gained throughout this work, we provide several recommendations for future research and development in this field.

Data modalities: What is commonly construed as EHR data in current academic settings is a derived, limited subset of "real" healthcare data as routinely used in hospitals. We might often be unaware of what type of data (i.e., modalities, patients, data silos) are missing and the importance of this data; data that is needed to create a model that provides a considerable benefit to the clinical workflow might not even be recorded in the current state. Data provenance (see above) and quality are also essential considerations, as they can significantly impact model performance and generalizability. As shown in Chapter 5, the exact provenance and mapping of data have significant effects on model development and evaluation. Our work in Chapter 4 shows that adding high-resolution vital sign data can improve model performance for postoperative complication prediction and, specifically, for the deep organ space surgical site infections (SSI-III) endpoint as defined in our study. However, many other data types could also be relevant, such as imaging data and genomics data. Here, we can rely on clinical and systems-biological hypotheses to drive the search for additional modalities and test them. A driving question should be: what hypothetically useful data can realistically be collected in a clinical setting without overburdening staff, patients, and hospital resources? Second, how do we integrate this data into models and, ultimately, clinical workflows?

Benchmarking and reproducibility: These are critical components of scientific research, particularly in the field of machine learning for healthcare. As discussed in Chapter 3 and the previous section, the lack of standardized benchmarks and reproducible experiments can hinder progress and make it difficult to compare different models and approaches. Implementing models into modular frameworks, such as *YAIB* and *MEDS-DEV*, is a first step towards addressing these challenges. However, there is an exponential number of choices made during model development that can bias results, whether intentionally or not [Sem+25]. There is an inherent interest in publishing positive results as quickly as possible, which, combined with the complexity of ML experiments and a largely voluntary reviewer pool, can lead to a lack of transparency and trust in this research. To alleviate this, we need to reform incentives, for example, through a refutations track [Sch+25a], and encourage open science, which could hopefully lead to the "frictionless" reproduction of existing work [McD+24; Don24].

Implementing ML in the clinical workflow: Each clinical setting is unique, and implementing ML models into the clinical workflow requires a deep understanding of the specific context, including the needs of both healthcare providers and patients. End-to-end implementation involves not only technical integration but also addressing ethical, legal, and social implications. Successful implementation often requires collaboration between data scientists, clinicians, and other stakeholders to ensure that the models are helpful, reliable, and accepted by end-users. Moreover, we acknowledge that AI-based solutions are often not the best *first step* to improve clinical care; once the workflow has been optimized, safe, dependable, and validated AI-based solutions can be considered.

Foundation models compared to small-scale specialized models: Foundation models, which are large-scale models trained on vast amounts of data, have shown promise in this and other works [Wax+25; Ste+23; Pan+25a]. However, as argued above, context is a crucial consideration of many applications: what works in visceral surgery might not work in oncology or, perhaps, not even in other surgical domains. The success of these models often depends on the quality and relevance of the training data to the downstream task. Earlier work trained on curated datasets has shown to be strong performers in various tasks; however, they might not generalize well to out-of-distribution data. Foundation models have drawbacks, including the need for substantial computational resources and their potential to perpetuate biases inherent in the training data. Current work is not yet conclusive on whether we should continue using the type of pipeline and models introduced in Chapter 4 or, instead, use institution-wide models, if viable, like in Chapter 5. Therefore, we recommend using foundation models as a starting point, when considered beneficial for the specific task, and fine-tuning or adapting them to specific clinical contexts as needed. Embeddings from foundation models can also be used as extra input features for smaller models, thereby combining the strengths of both approaches.

6.4 Final Remarks

In recent years, there have been many exciting developments in machine learning and its applications to the healthcare domain. More datasets are becoming available: Phy-

sioNet [Gol+00] now contains several large-scale EHR datasets with different cohorts, characteristics and modalities contributed by the community. Moreover, the rise and maturity of Large Language Model (LLM)s and foundation model (FM)s have opened new avenues of research and potential deployments using massive amounts of available retrospective data. Furthermore, the community is becoming more aware of the importance of reproducibility and transparency [Sem+25], leading to the development of standards and tools that facilitate these goals [Fen+25].

However, it is essential to approach these developments critically, recognizing that these developments, on their own, will not lead to meaningful improvements in clinical practice. The successful application of machine learning, AI, and, more broadly, computer science applied to healthcare requires careful consideration of the unique challenges and complexities of each medical subdomain, including data quality, ethical considerations, and the need for robust validation and evaluation of models in real-world clinical settings.

Bibliography

[Aba+16] Martin Abadi et al. "TensorFlow: A System for Large-Scale Machine Learning". In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. 2016, pp. 265–283 (cit. on p. 41).

[Adi+25] Amin Adibi et al. *Recent Advances, Applications and Open Challenges in Machine Learning for Health: Reflections from Research Roundtables at ML4H 2024 Symposium*. 2025-02. DOI: 10.48550/arXiv.2502.06693. arXiv: 2502.06693 [cs]. (Visited on 2025-07-21) (cit. on p. 13).

[Adi+16] Mohammad Adibuzzaman, Ken Musselman, Alistair Johnson, Paul Brown, Zachary Pitluk Pitluk, and Ananth Grama. "Closing the Data Loop: An Integrated Open Access Analysis Platform for the MIMIC Database". In: *2016 Computing in Cardiology Conference*. 2016-09. DOI: 10.22489/CinC.2016.043-205. (Visited on 2023-02-13) (cit. on pp. 29, 102).

[Agy+23] Kwaku Agyeman-Manu, Tedros Adhanom Ghebreyesus, Mohamed Maait, Alexandru Rafila, Lino Tom, Nísia Trindade Lima, and Dechen Wangmo. "Prioritising the Health and Care Workforce Shortage: Protect, Invest, Together". In: *The Lancet Global Health* 11.8 (2023-08), e1162–e1164. ISSN: 2214-109X. DOI: 10.1016/S2214-109X(23)00224-3. (Visited on 2026-01-15) (cit. on p. 1).

[Aki+19] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. New York, NY, USA: Association for Computing Machinery, 2019-07, pp. 2623–2631. ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330701. (Visited on 2025-08-18) (cit. on p. 41).

[Ale+25] Katharina Alefs, Susanne Ibing, Pia Francesca Rissom, Jan Carlo Schmid, Arkadiusz Kwasigroch, Robin van de Water, Bernhard Y. Renard, and Eugenia Alleva. "Towards Foundation Model-Based Propensity Score Matching from Electronic Health Records". In: *Machine Learning for Health Symposium 2025*. 2025-09. (Visited on 2025-11-22) (cit. on pp. xiv, 130).

[Alm+22] Malak Abdullah Almarshad, Md Saiful Islam, Saad Al-Ahmadi, and Ahmed S. BaHammam. "Diagnostic Features and Potential Applications of PPG Signal in Healthcare: A Systematic Review". In: *Healthcare* 10.3 (2022-03), p. 547. ISSN: 2227-9032. DOI: 10.3390/healthcare10030547. (Visited on 2024-02-02) (cit. on p. 96).

[AD52] T. W. Anderson and D. A. Darling. "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes". In: *The Annals of Mathematical Statistics* 23.2 (1952), pp. 193–212. ISSN: 0003-4851. JSTOR: 2236446. (Visited on 2026-01-05) (cit. on p. 61).

[Aro+23] Anmol Arora et al. "The Value of Standards for Health Datasets in Artificial Intelligence-Based Applications". In: *Nature Medicine* 29.11 (2023-11), pp. 2929–2938. ISSN: 1546-170X. DOI: 10.1038/s41591-023-02608-w. (Visited on 2025-12-03) (cit. on p. 101).

[Asg+25] Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. "A Framework to Assess Clinical Safety and Hallucination Rates of LLMs for Medical Text Summarisation". In: *npj Digital Medicine* 8.1 (2025-05), p. 274. ISSN: 2398-6352. DOI: 10.1038/s41746-025-01670-7. (Visited on 2025-11-30) (cit. on p. 148).

[Att+25] Rafi Al Attrach, Rajna Fani, Yugang Jia, David Restrepo, Leo Anthony Celi, and Peter Schueffler. "Rethinking Tokenization for Clinical Time Series: When Less Is More". In: *Machine Learning for Health 2025*. 2025-11. (Visited on 2025-11-30) (cit. on p. 129).

[Aut+15] Adam Auton et al. "A Global Reference for Human Genetic Variation". In: *Nature* 526.7571 (2015-10), pp. 68–74. ISSN: 1476-4687. DOI: 10.1038/nature15393. (Visited on 2026-01-17) (cit. on p. 15).

[Bac+21] Joshua D. Backman et al. "Exome Sequencing and Analysis of 454,787 UK Biobank Participants". In: *Nature* 599.7886 (2021-11), pp. 628–634. ISSN: 1476-4687. DOI: 10.1038/s41586-021-04103-z (cit. on p. 15).

[Bag+20] Sean M. Bagshaw, Dat T. Tran, Dawn Opgenorth, Xiaoming Wang, Danny J. Zuege, Armann Ingolfsson, Henry T. Stelfox, and Nguyen X. Thanh. "Assessment of Costs of Avoidable Delays in Intensive Care Unit Discharge". In: *JAMA Network Open* 3.8 (2020-08), e2013913. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2020.13913. (Visited on 2024-02-16) (cit. on p. 57).

[BKK18] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. *An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling*. 2018-04. DOI: 10.48550/arXiv.1803.01271. arXiv: 1803.01271 [cs]. (Visited on 2023-04-18) (cit. on pp. 47, 82, 85, 94, 211).

[BXA20] Stephanie Baker, Wei Xiang, and Ian Atkinson. "Continuous and Automatic Mortality Risk Prediction Using Vital Signs in the Intensive Care Unit: A Hybrid Neural Network Approach". In: *Scientific Reports* 10.1 (2020-12), p. 21282. ISSN: 2045-2322. DOI: 10.1038/s41598-020-78184-7. (Visited on 2022-01-27) (cit. on p. 46).

[Bal23] Philip Ball. "Is AI Leading to a Reproducibility Crisis in Science?" In: *Nature* 624.7990 (2023-12), pp. 22–25. DOI: 10.1038/d41586-023-03817-6. (Visited on 2025-12-03) (cit. on p. 148).

[Bar+20] Sebastiano Barbieri, James Kemp, Oscar Perez-Concha, Sradha Kotwal, Martin Gallagher, Angus Ritchie, and Louisa Jorm. "Benchmarking Deep Learning Architectures for Predicting Readmission to the ICU and Describing Patients-at-Risk". In: *Scientific Reports* 10.1 (2020-01), p. 1111. ISSN: 2045-2322. DOI: 10.1038/s41598-020-58053-z. (Visited on 2023-01-12) (cit. on pp. 29, 31).

[Bas+23] Mujeeb A. Basit, Mereeja Varghese, Aamirah Vadsariya, Bhavini Nayee, Margaret Langley, Ashley Huynh, Jennifer Cai, Donglu Xie, Cindy Kao, Eric Nguyen, Todd Boutte, Shiby Antony, Tammye Garrett, Christoph U. Lehmann, and Duwayne L. Willett. ""OMOP Anywhere": Daily Updates from EHR Data Leveraging Epic's Native Tools". In: *Proceedings of the 2023 OHDSI Global Symposium*. Washington, DC, USA: Observational Health Data Sciences and Informatics (OHDSI), 2023-10 (cit. on p. 103).

[Bas+17] Claudio Bassi et al. "The 2016 Update of the International Study Group (ISGPS) Definition and Grading of Postoperative Pancreatic Fistula: 11 Years After". In: *Surgery* 161.3 (2017-03), pp. 584–591. ISSN: 1532-7361. DOI: 10.1016/j.surg.2016.11.014 (cit. on pp. 56, 71).

[BMV16] Nathaniel D. Bastian, David Munoz, and Marta Ventura. "A Mixed-Methods Research Framework for Healthcare Process Improvement". In: *Journal of Pediatric Nursing* 31.1 (2016), e39–51. ISSN: 1532-8449. DOI: 10.1016/j.pedn.2015.09.003 (cit. on p. 14).

[Bau+19] Philip Baum, Johannes Diers, Sven Lichthardt, Carolin Kastner, Nicolas Schlegel, Christoph-Thomas Germer, and Armin Wiegering. "Mortality and Complications Following Visceral Surgery". In: *Deutsches Ärzteblatt international* (2019-11). ISSN: 1866-0452. DOI: 10.3238/arzteb1.2019.0739. (Visited on 2024-11-20) (cit. on p. 55).

[BMG20] Andrew L. Beam, Arjun K. Manrai, and Marzyeh Ghassemi. "Challenges to the Reproducibility of Machine Learning Models in Health Care". In: *JAMA* 323.4 (2020-01), pp. 305–306. ISSN: 0098-7484. DOI: 10.1001/jama.2019.20866. (Visited on 2025-12-03) (cit. on p. 148).

[Bec+23] Femke L. Becking-Verhaar, Robin P. H. Verweij, Marjan de Vries, Hester Vermeulen, Harry van Goor, and Getty J. Huisman-de Waal. "Continuous Vital Signs Monitoring with a Wireless Device on a General Ward: A Survey to Explore Nurses' Experiences in a Post-Implementation Period". In: *International Journal of Environmental Research and Public Health* 20.10 (2023-05), p. 5794. ISSN: 1660-4601. DOI: 10.3390/ijerph20105794 (cit. on p. 93).

[BS13] Duane Bender and Kamran Sartipi. "HL7 FHIR: An Agile and RESTful Approach to Healthcare Information Exchange". In: *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. 2013-06, pp. 326–331. DOI: 10.1109/CBMS.2013.6627810. (Visited on 2025-10-11) (cit. on pp. 9, 101, 104).

[Ben+25] Alex Bennett, Hannes Ulrich, Joshua Wiedekopf, Piotr Szul, John Grimes, and Alistair Johnson. *MIMIC-IV Clinical Database Demo on FHIR*. 2025. DOI: 10.13026/VPHG-Y548. (Visited on 2025-10-19) (cit. on p. 109).

[Ben+23] Nicolas Bennett, Drago Plecko, Ida-Fong Ukor, Nicolai Meinshausen, and Peter Bühlmann. "Ricu: R's Interface to Intensive Care Data". In: *GigaScience* 12 (2023-06). DOI: 10.1093/gigascience/giad041 (cit. on pp. 28, 29, 40, 204).

[Ber06] Ulf Bergman. "The History of the Drug Utilization Research Group in Europe". In: *Pharmacoepidemiology and Drug Safety* 15.2 (2006), pp. 95–98. ISSN: 1099-1557. DOI: 10.1002/pds.1171. (Visited on 2025-12-19) (cit. on p. 102).

[Ber+18] Dimitris Bertsimas, Jack Dunn, George C. Velmahos, and Haytham M. A. Kaafarani. "Surgical Risk Is Not Linear: Derivation and Validation of a Novel, User-friendly, and Machine-learning-based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) Calculator". In: *Annals of Surgery* 268.4 (2018-10), p. 574. ISSN: 0003-4932. DOI: 10.1097/SLA.0000000000002956. (Visited on 2026-01-05) (cit. on p. 67).

[Bha+23] Hemanga Kumar Bhattacharjee, M.p Kaviyarasan, Kh. Jitenkumar Singh, K. Don Jose, Suhani Suhani, Mohit Joshi, and Rajinder Parshad. "Age Adjusted Charlson Comorbidity Index (a-CCI) AS a Tool to Predict 30-Day Post-Operative Outcome in General Surgery Patients". In: *ANZ Journal of Surgery* 93.1-2 (2023), pp. 132–138. ISSN: 1445-2197. DOI: 10.1111/ans.18178. (Visited on 2026-01-05) (cit. on p. 68).

[Bie20] Lukas Biewald. *Experiment Tracking with Weights and Biases*. 2020 (cit. on p. 41).

[Blo+20] Joshua A. Bloomstone, Benjamin T. Houseman, Evora Vicents Sande, Ann Brantley, Jessica Curran, Gerald A. Maccioli, Tania Haddad, James Steinshouer, David Walker, and Ramani Moonesinghe. "Documentation of Individualized Preoperative Risk Assessment: A Multi-Center Study". In: *Perioperative Medicine* 9.1 (2020-09), p. 28. ISSN: 2047-0525. DOI: 10.1186/s13741-020-00156-2. (Visited on 2026-01-05) (cit. on p. 68).

[Boa+24] William Boag, Alifia Hasan, Jee Young Kim, Mike Revoir, Marshall Nichols, William Ratliff, Michael Gao, Shira Zilberstein, Zainab Samad, Zahra Hoodbhoy, Mushyada Ali, Nida Saddaf Khan, Manesh Patel, Suresh Balu, and Mark Sendak. "The Algorithm Journey Map: A Tangible Approach to Implementing AI Solutions in Healthcare". In: *npj Digital Medicine* 7.1 (2024-04), p. 87. ISSN: 2398-6352. DOI: 10.1038/s41746-024-01061-4. (Visited on 2025-07-28) (cit. on p. 25).

[Bod04] Olivier Bodenreider. "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology". In: *Nucleic Acids Research* 32.suppl_1 (2004-01), pp. D267–D270. ISSN: 0305-1048. DOI: 10.1093/nar/gkh061. (Visited on 2025-10-28) (cit. on pp. 103, 118).

[Bom+21] Rishi Bommasani et al. "On the Opportunities and Risks of Foundation Models". In: *arXiv:2108.07258 [cs]* (2021-08). arXiv: 2108.07258 [cs]. (Visited on 2022-03-10) (cit. on pp. 1, 13, 21, 105).

[Bor+24] Florian Borchert, Ignacio Llorca, Roland Roller, Bert Arnrich, and Matthieu-P Schapranow. "xMEN: A Modular Toolkit for Cross-Lingual Medical Entity Normalization". In: *JAMIA Open* 8.1 (2024-12), ooae147. ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooae147. (Visited on 2025-10-28) (cit. on pp. 103, 118).

[Bor+23] Israel Júnior Borges do Nascimento, Hebatullah Abdulazeem, Lenny Thinagaran Vasanthan, Edson Zangiacomi Martinez, Miriane Lucindo Zucoloto, Lasse Østengaard, Natasha Azzopardi-Muscat, Tomas Zapata, and David Novillo-Ortiz. "Barriers and Facilitators to Utilizing Digital Health Technologies by Healthcare Professionals". In: *npj Digital Medicine* 6.1 (2023-09), p. 161. ISSN: 2398-6352. DOI: 10.1038/s41746-023-00899-4. (Visited on 2026-01-19) (cit. on pp. 101, 148).

[Box+15] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2015-05. ISBN: 978-1-118-67492-5 (cit. on p. 19).

[Bra+20] Noé Brasier, Lukas Geissmann, Miro Käch, Markus Mutke, Bianca Hoelz, Fiorangelo De Ieso, and Jens Eckstein. "Device- and Analytics-Agnostic Infrastructure for Continuous Inpatient Monitoring: A Technical Note". In: *Digital Biomarkers* 4.2 (2020-08), pp. 62–68. ISSN: 2504-110X. DOI: 10.1159/000509279. (Visited on 2023-11-14) (cit. on p. 73).

[Bro+14a] Harvey Brown, Jamie Terrence, Patricia Vasquez, David W. Bates, and Eyal Zimlichman. "Continuous Monitoring in an Inpatient Medical-Surgical Unit: A Controlled Clinical Trial". In: *The American Journal of Medicine* 127.3 (2014-03), pp. 226–232. ISSN: 0002-9343, 1555-7162. DOI: 10.1016/j.amjmed.2013.12.004. (Visited on 2024-12-17) (cit. on p. 58).

[Bro+25] Katherine E Brown, Chao Yan, Zhuohang Li, Xinmeng Zhang, Benjamin X Collins, You Chen, Ellen Wright Clayton, Murat Kantarcioglu, Yevgeniy Vorobeychik, and Bradley A Malin. "Large Language Models Are Less Effective at Clinical Prediction Tasks than Locally Trained Machine Learning Models". In: *Journal of the American Medical Informatics Association* 32.5 (2025-05), pp. 811–822. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaf038. (Visited on 2025-08-14) (cit. on pp. 21, 82, 85).

[Bro+14b] Sarah R. Brown, Ronnie Mathew, Ada Keding, Helen C. Marshall, Julia M. Brown, and David G. Jayne. "The Impact of Postoperative Complications on Long-Term Quality of Life after Curative Colorectal Cancer Surgery". In: *Annals of Surgery* 259.5 (2014-05), pp. 916–923. ISSN: 1528-1140. DOI: 10.1097/SLA.0000000000000407 (cit. on p. 56).

[Bro+20] Tom B. Brown et al. *Language Models Are Few-Shot Learners*. 2020-07. DOI: 10.48550/arXiv.2005.14165. arXiv: 2005.14165 [cs]. (Visited on 2025-07-21) (cit. on pp. 1, 13, 21, 94, 105, 107).

[Bun25] Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). *OPS: Operationen- Und Prozedurenschlüssel. Amtliche Klassifikation Zur Verschlüsselung von Operationen, Prozeduren Und Allgemeinen Medizinischen Maßnahmen*. Manual. Bonn, 2025 (cit. on pp. 60, 233).

[CKP09] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. "Building Classifiers with Independency Constraints". In: *2009 IEEE International Conference on Data Mining Workshops*. 2009-12, pp. 13–18. DOI: 10.1109/ICDMW.2009.83 (cit. on p. 210).

[Cal+23] Alison Callahan, Euan Ashley, Somalee Datta, Priyamvada Desai, Todd A Ferris, Jason A Fries, Michael Halaas, Curtis P Langlotz, Sean Mackey, José D Posada, Michael A Pfeffer, and Nigam H Shah. "The Stanford Medicine Data Science Ecosystem for Clinical and Translational Research". In: *JAMIA Open* 6.3 (2023-10), ooad054. ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooad054. (Visited on 2025-10-24) (cit. on pp. 22, 105, 106).

[Cal+16] Jacob S. Calvert, Daniel A. Price, Uli K. Chettipally, Christopher W. Barton, Mitchell D. Feldman, Jana L. Hoffman, Melissa Jay, and Ritankar Das. "A Computational Approach to Early Sepsis Detection". In: *Computers in Biology and Medicine* 74 (2016-07), pp. 69–73. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2016.05.003. (Visited on 2023-02-10) (cit. on pp. 50, 51).

[Car+19] Heather Carmichael, Douglas M. Overbey, Patrick Hosokawa, Christina M. Goode, Teresa S. Jones, Carlton C. Barnett, Edward L. Jones, and Thomas N. Robinson. "Wearable Technology-A Pilot Study to Define "Normal" Postoperative Recovery Trajectories". In: *The Journal of Surgical Research* 244 (2019-12), pp. 368–373. ISSN: 1095-8673. DOI: 10.1016/j.jss.2019.06.057 (cit. on p. 58).

[Cas25] Fidelia Cascini. "Electronic Health Data Categories". In: *Secondary Use of Electronic Health Data: Public Health Perspectives, Use Cases and Challenges*. Ed. by Fidelia Cascini. Cham: Springer Nature Switzerland, 2025, pp. 1–49. ISBN: 978-3-031-88497-9. DOI: 10.1007/978-3-031-88497-9_1. (Visited on 2025-07-22) (cit. on pp. 13, 14).

[Cas+24] Alexandre Castonguay, Gerit Wagner, Aude Motulsky, and Guy Paré. "AI Maturity in Health Care: An Overview of 10 OECD Countries". In: *Health Policy* 140 (2024-02), p. 104938. ISSN: 0168-8510. DOI: 10.1016/j.healthpol.2023.104938. (Visited on 2025-12-03) (cit. on p. 150).

[CT10] Gavin C. Cawley and Nicola LC Talbot. "On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation". In: *The Journal of Machine Learning Research* 11 (2010), pp. 2079–2107. (Visited on 2026-01-05) (cit. on p. 62).

[Cer+24] Mariagiulia Dal Cero et al. "International External Validation of Risk Prediction Model of 90-Day Mortality after Gastrectomy for Cancer Using Machine Learning". In: *Cancers* 16.13 (2024-07). ISSN: 2072-6694. DOI: 10.3390/cancers16132463. (Visited on 2026-01-05) (cit. on p. 67).

[Cha+16] Chun-Ming Chang, Wen-Yao Yin, Chang-Kao Wei, Chin-Chia Wu, Yu-Chieh Su, Chia-Hui Yu, and Ching-Chih Lee. "Adjusted Age-Adjusted Charlson Comorbidity Index Score as a Risk Measure of Perioperative Mortality before Cancer Surgery". In: *PLOS ONE* 11.2 (2016-02), e0148076. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0148076. (Visited on 2026-01-05) (cit. on p. 68).

[Cha+87] Mary E. Charlson, Peter Pompei, Kathy L. Ales, and C. Ronald MacKenzie. "A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation". In: *Journal of Chronic Diseases* 40.5 (1987-01), pp. 373–383. ISSN: 0021-9681. DOI: 10.1016/0021-9681(87)90171-8. (Visited on 2026-01-04) (cit. on p. 61).

[Cha+25] Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. "RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs". In: *ACM Comput. Surv.* 58.2 (2025-09), 53:1–53:37. ISSN: 0360-0300. DOI: 10.1145/3743127. (Visited on 2026-01-21) (cit. on p. 20).

[Cha+02] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: Synthetic Minority over-Sampling Technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357. (Visited on 2026-01-05) (cit. on p. 62).

[Che+24] Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Danielle Bitterman, Fei Wang, and Kai Shu. *ClinicalBench: Can LLMs Beat Traditional ML Models in Clinical Prediction?* 2024-11. DOI: 10.48550/arXiv.2411.06469. arXiv: 2411.06469. (Visited on 2024-11-13) (cit. on pp. 13, 21, 94).

[Che+18a] David Chen, Naveed Afzal, Sunghwan Sohn, Elizabeth B. Habermann, James M. Naessens, David W. Larson, and Hongfang Liu. "Postoperative Bleeding Risk Prediction for Patients Undergoing Colorectal Surgery". In: *Surgery* 164.6 (2018-12), pp. 1209–1216. ISSN: 0039-6060. DOI: 10.1016/j.surg.2018.05.043. (Visited on 2026-01-05) (cit. on p. 67).

[Che+23] Kevin A. Chen, Chinmaya U. Joisa, Jonathan M. Stem, Jose G. Guillem, Shawn M. Gomez, and Muneera R. Kapadia. "Prediction of Ureteral Injury During Colorectal Surgery Using Machine Learning". In: *The American Surgeon™* 89.12 (2023-12), pp. 5702–5710. ISSN: 0003-1348. DOI: 10.1177/00031348231173981. (Visited on 2026-01-05) (cit. on p. 67).

[Che+22] Kevin A. Chen, Chinmaya U. Joisa, Karyn B. Stitzenberg, Jonathan Stem, Jose G. Guillem, Shawn M. Gomez, and Muneera R. Kapadia. "Development and Validation of Machine Learning Models to Predict Readmission After Colorectal Surgery". In: *Journal of Gastrointestinal Surgery* 26.11 (2022-11), pp. 2342–2350. ISSN: 1091-255X. DOI: 10.1007/s11605-022-05443-5. (Visited on 2026-01-05) (cit. on p. 67).

[CG16] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016-08, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. (Visited on 2022-12-24) (cit. on pp. 43, 62, 82, 85, 94, 232).

[Che+18b] Peng Cheng, Lemuel R. Waitman, Yong Hu, and Mei Liu. "Predicting Inpatient Acute Kidney Injury over Different Time Horizons: How Early and Accurate?" In: *AMIA Annual Symposium Proceedings* 2017 (2018-04), pp. 565–574. ISSN: 1942-597X. (Visited on 2023-06-05) (cit. on p. 46).

[Chi+22] Maria Michela Chiarello, Pietro Fransvea, Maria Cariati, Neill James Adams, Valentina Bianchi, and Giuseppe Brisinda. "Anastomotic Leakage in Colorectal Cancer Surgery". In: *Surgical Oncology* 40 (2022-03), p. 101708. ISSN: 1879-3320. DOI: 10.1016/j.suronc.2022.101708 (cit. on p. 57).

[Chi+20] Calvin J. Chiew, Nan Liu, Ting Hway Wong, Yilin E. Sim, and Hairil R. Abdullah. "Utilizing Machine Learning Methods for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit Admission". In: *Annals of Surgery* 272.6 (2020-12), p. 1133. ISSN: 0003-4932. DOI: 10.1097/SLA.0000000000003297. (Visited on 2026-01-05) (cit. on p. 67).

[Cho+14a] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. *Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation*. 2014-09. DOI: 10.48550/arXiv.1406.1078. arXiv: 1406.1078 [cs]. (Visited on 2024-12-17) (cit. on pp. 82, 85, 94).

[Cho+14b] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*. 2014-10. arXiv: 1409.1259 [cs, stat]. (Visited on 2023-01-18) (cit. on p. 47).

[Chr+17] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. "Deep Reinforcement Learning from Human Preferences". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. (Visited on 2025-08-18) (cit. on p. 20).

[Cli+14] Lei Clifton, David A. Clifton, Marco A. F. Pimentel, Peter J. Watkinson, and Lionel Tarassenko. "Predictive Monitoring of Mobile Patients by Combining Clinical Observations With Data From Wearable Sensors". In: *IEEE Journal of Biomedical and Health Informatics* 18.3 (2014-05), pp. 722–730. ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2013.2293059. (Visited on 2023-10-11) (cit. on pp. 57, 58).

[Col+14] Francis S Collins, Kathy L Hudson, Josephine P Briggs, and Michael S Lauer. "PCORnet: Turning a Dream into Reality". In: *Journal of the American Medical Informatics Association* 21.4 (2014-07), pp. 576–577. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2014-002864. (Visited on 2025-10-11) (cit. on p. 101).

[Col+15] G S Collins, J B Reitsma, D G Altman, and K G M Moons. "Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement". In: *British Journal of Surgery* 102.3 (2015-01), pp. 148–158. ISSN: 0007-1323, 1365-2168. DOI: 10.1002/bjs.9736. (Visited on 2022-05-24) (cit. on p. 59).

[Col+24] Gary S. Collins et al. "TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods". In: *BMJ* 385 (2024-04), e078378. ISSN: 1756-1833. DOI: 10.1136/bmj-2023-078378. (Visited on 2024-12-18) (cit. on pp. 59, 69, 149, 242).

[24a] *COPRA PDMS (Patientendatenmanagementsystem) | COPRA System GmbH*. https://copra-system.de. 2024-02. (Visited on 2024-02-16) (cit. on p. 73).

[DGS18] Dan Holtmann-Rice, Sergio Guadarrama, and Nathan Silberman. *Gin-Config*. Google. 2018 (cit. on p. 41).

[DG24] Tri Dao and Albert Gu. *Transformers Are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality*. 2024-05. arXiv: 2405.21060 [cs]. (Visited on 2024-06-19) (cit. on p. 108).

[De 21] Jens De Bruijn. *Python Implementation of the Nadeau and Bengio Correction of Dependent Student's t-Test*. https://gist.github.com/jensdebruijn/13e8eeda85eb8644ac2a4ac4c3b8e732. 2021. (Visited on 2026-01-21) (cit. on p. 240).

[DSW21] Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. "MINIROCKET: A Very Fast (Almost) Deterministic Transform for Time Series Classification". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021-08, pp. 248–257. DOI: 10.1145/3447548.3467231. arXiv: 2012.08791 [cs, stat]. (Visited on 2024-09-24) (cit. on p. 86).

[Dem06] Janez Demšar. "Statistical Comparisons of Classifiers over Multiple Data Sets". In: *J. Mach. Learn. Res.* 7 (2006-12), pp. 1–30. ISSN: 1532-4435. (Visited on 2026-01-21) (cit. on p. 241).

[Dev+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019-05. DOI: 10.48550/arXiv.1810.04805. arXiv: 1810.04805 [cs]. (Visited on 2025-07-25) (cit. on pp. 20, 21, 105, 107).

[DHV15] Brian E. Dixon, John Hook, and Daniel J. Vreeman. "Learning From the Crowd in Terminology Mapping: The LOINC Experience". In: *Laboratory Medicine* 46.2 (2015-05), pp. 168–174. ISSN: 0007-5027. DOI: 10.1309/LMWJ730SVKTUBAOJ. (Visited on 2025-12-03) (cit. on p. 149).

[Do+23] Hyungrok Do, Yuxin Chang, Yoon Sang Cho, Padhraic Smyth, and Judy Zhong. "Fair Survival Time Prediction via Mutual Information Minimization". In: *Proceedings of machine learning research* 219 (2023-08), pp. 128–149. ISSN: 2640-3498. (Visited on 2025-07-30) (cit. on p. 24).

[Dob20] Geoffrey P. Dobson. "Trauma of Major Surgery: A Global Problem That Is Not Going Away". In: *International Journal of Surgery (London, England)* 81 (2020-09), pp. 47–54. ISSN: 1743-9191. DOI: 10.1016/j.ijsu.2020.07.017. (Visited on 2025-08-26) (cit. on p. 55).

[Doh+17] Aiden Doherty et al. "Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study". In: *PLOS ONE* 12.2 (2017-02), e0169649. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0169649. (Visited on 2026-01-17) (cit. on p. 15).

[Don06] Kevin Donnelly. "SNOMED-CT: The Advanced Terminology and Coding System for eHealth". In: *Studies in health technology and informatics* 121 (2006), p. 279. (Visited on 2025-10-28) (cit. on pp. 102, 118).

[Don24] David Donoho. "Data Science at the Singularity". In: *Harvard Data Science Review* 6.1 (2024-01). ISSN: 2644-2353, 2688-8513. DOI: 10.1162/99608f92.b91339ef. (Visited on 2026-01-19) (cit. on pp. 101, 149, 151).

[Dow+19] Candice Downey, Shu Ng, David Jayne, and David Wong. "Reliability of a Wearable Wireless Patch for Continuous Remote Monitoring of Vital Signs in Patients Recovering from Major Surgery: A Clinical Validation Study from the TRaCINg Trial". In: *BMJ Open* 9.8 (2019-08), e031150. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2019-031150. (Visited on 2024-12-17) (cit. on pp. 58, 95).

[Dow+23] Candice L. Downey, Jamie Bainbridge, David G. Jayne, and David M. Meads. "Impact of In-Hospital Postoperative Complications on Quality of Life up to 12 Months after Major Abdominal Surgery". In: *The British Journal of Surgery* 110.9 (2023-08), pp. 1206–1212. ISSN: 1365-2168. DOI: 10.1093/bjs/znad167 (cit. on p. 56).

[Du23] Wenjie Du. *PyPOTS: A Python Toolbox for Data Mining on Partially-Observed Time Series*. 2023-05. DOI: 10.48550/arXiv.2305.18811. arXiv: 2305.18811 [cs, stat]. (Visited on 2023-07-06) (cit. on p. 207).

[Dun+20] Heather P. Duncan, Balazs Fule, Iain Rice, Alice J. Sitch, and David Lowe. "Wireless Monitoring and Real-Time Adaptive Predictive Indicator of Deterioration". In: *Scientific Reports* 10.1 (2020-07), p. 11366. ISSN: 2045-2322. DOI: 10.1038/s41598-020-67835-4. (Visited on 2024-11-11) (cit. on p. 59).

[Edd+22] Yassin Eddahchouri, Roel V. Peelen, Mats Koeneman, Hugo R. W. Touw, Harry van Goor, and Sebastian J. H. Bredie. "Effect of Continuous Wireless Vital Sign Monitoring on Unplanned ICU Admissions and Rapid Response Team Calls: A before-and-after Study". In: *British Journal of Anaesthesia* 128.5 (2022-05), pp. 857–863. ISSN: 0007-0912, 1471-6771. DOI: 10.1016/j.bja.2022.01.036. (Visited on 2024-12-17) (cit. on p. 58).

[Ede+24] Dana P. Edelson, Matthew M. Churpek, Kyle A. Carey, Zhenqiu Lin, Chenxi Huang, Jonathan M. Siner, Jennifer Johnson, Harlan M. Krumholz, and Deborah J. Rhodes. "Early Warning Scores With and Without Artificial Intelligence". In: *JAMA Network Open* 7.10 (2024-10), e2438986. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2024.38986. (Visited on 2024-11-22) (cit. on pp. 58, 59).

[Ehl+17] Anne P. Ehlers, Senjuti Basu Roy, Sara Khor, Prathyusha Mandagani, Moushumi Maria, Rafael Alfonso-Cristancho, and David R. Flum. "Improved Risk Prediction Following Surgery Using Machine Learning Algorithms". In: *eGEMs* 5.2 (2017-04), p. 3. ISSN: 2327-9214. DOI: 10.13063/2327-9214.1278. (Visited on 2026-01-05) (cit. on p. 67).

[Ein+22] Bar Eini-Porat, Ofra Amir, Danny Eytan, and Uri Shalit. "Tell Me Something Interesting: Clinical Utility of Machine Learning Prediction Models in the ICU". In: *Journal of Biomedical Informatics* 132 (2022-08), p. 104107. ISSN: 15320464. DOI: 10.1016/j.jbi.2022.104107. (Visited on 2022-07-21) (cit. on p. 27).

[Ele96] Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology. "Heart Rate Variability". In: *Circulation* 93.5 (1996-03), pp. 1043–1065. DOI: 10.1161/01.CIR.93.5.1043. (Visited on 2025-08-11) (cit. on p. 78).

[Elf+21] Omar Elfanagely, Yoshiko Toyoda, Sammy Othman, Joseph A. Mellia, Marten Basta, Tony Liu, Konrad Kording, Lyle Ungar, and John P. Fischer. "Machine Learning and Surgical Outcomes Prediction: A Systematic Review". In: *Journal of Surgical Research* 264 (2021-08), pp. 346–361. ISSN: 0022-4804. DOI: 10.1016/j.jss.2021.02.045. (Visited on 2022-01-28) (cit. on p. 24).

[Eng+20] Milo Engoren, Troy Seelhammer, Robert E. Freundlich, Michael D. Maile, Matthew J. G. Sigakis, and Thomas A. Schwann. "A Comparison of Sepsis-2 (Systemic Inflammatory Response Syndrome Based) to Sepsis-3 (Sequential Organ Failure Assessment Based) Definitions—A Multicenter Retrospective Study*". In: *Critical Care Medicine* 48.9 (2020-09), pp. 1258–1264. ISSN: 0090-3493. DOI: 10.1097/CCM.0000000000004449. (Visited on 2023-05-02) (cit. on p. 50).

[Ft23] William Falcon and The PyTorch Lightning team. *PyTorch Lightning*. Zenodo. 2023-04. DOI: 10.5281/zenodo.7859091 (cit. on pp. 41, 205, 211).

[Fal+24] Adibvafa Fallahpour, Mahshid Alinoori, Wenqian Ye, Xu Cao, Arash Afkanpour, and Amrit Krishnan. *EHRMamba: Towards Generalizable and Scalable Foundation Models for Electronic Health Records*. 2024-11. DOI: 10.48550/arXiv.2405.14567. arXiv: 2405.14567 [cs]. (Visited on 2025-10-12) (cit. on pp. 22, 106–108, 142).

[Fär+12] Franz Färber, Norman May, Wolfgang Lehner, Philipp Große, Ingo Müller, Hannes Rauhe, and Jonathan Dees. "The SAP HANA Database–an Architecture Overview." In: *IEEE Data Eng. Bull.* 35.1 (2012), pp. 28–33 (cit. on p. 110).

[Fel+16] Shara I. Feld, Sarah E. Tevis, Alexander G. Cobian, Mark W. Craven, and Gregory D. Kennedy. "Multiple Postoperative Complications: Making Sense of the Trajectories". In: *Surgery* 160.6 (2016-12), pp. 1666–1674. ISSN: 1532-7361. DOI: 10.1016/j.surg.2016.08.047 (cit. on p. 57).

[Fen+25] Jean Feng, Fan Xia, Karandeep Singh, and Romain Pirracchio. "Not All Clinical AI Monitoring Systems Are Created Equal: Review and Recommendations". In: *NEJM AI* 2.2 (2025-01). ISSN: 2836-9386. DOI: 10.1056/AIra2400657. (Visited on 2025-01-30) (cit. on pp. 96, 148, 152).

[Fle+20a] Lucas M. Fleuren, Thomas L. T. Klausch, Charlotte L. Zwager, Linda J. Schoonmade, Tingjie Guo, Luca F. Roggeveen, Eleonora L. Swart, Armand R. J. Girbes, Patrick Thoral, Ari Ercole, Mark Hoogendoorn, and Paul W. G. Elbers. "Machine Learning for the Prediction of Sepsis: A Systematic Review and Meta-Analysis of Diagnostic Test Accuracy". In: *Intensive Care Medicine* 46.3 (2020), pp. 383–400. ISSN: 0342-4642. DOI: 10.1007/s00134-019-05872-y. (Visited on 2023-05-16) (cit. on p. 28).

[Fle+20b] Lucas M. Fleuren, Patrick Thoral, Duncan Shillan, Ari Ercole, Paul W. G. Elbers, Mark Hoogendoorn, Ben Gibbison, Thomas L. T. Klausch, Tingjie Guo, Luca F. Roggeveen, Eleonora L. Swart, Armand R. J. Girbes, and Right Data Right Now Collaborators. "Machine Learning in Intensive Care Medicine: Ready for Take-Off?" In: *Intensive Care Medicine* 46.7 (2020-07), pp. 1486–1488. ISSN: 1432-1238. DOI: 10.1007/s00134-020-06045-y. (Visited on 2023-05-16) (cit. on pp. 27, 46, 50).

[Fom+20] V. Fomin, J. Anmol, S. Desroziers, J. Kriss, and A. Tejani. *High-Level Library to Help with Training Neural Networks in PyTorch*. 2020 (cit. on pp. 42, 209).

[Fri40] Milton Friedman. "A Comparison of Alternative Tests of Significance for the Problem of m Rankings". In: *The annals of mathematical statistics* 11.1 (1940), pp. 86–92. JSTOR: 2235971. (Visited on 2026-01-21) (cit. on pp. 88, 241).

[FXC25] Keren Fuentes, Mimee Xu, and Irene Chen. *Privacy-Preserving Dataset Combination*. 2025-02. DOI: 10.48550/arXiv.2502.05765. arXiv: 2502.05765 [cs]. (Visited on 2025-04-24) (cit. on p. 54).

[Fut+16] Joseph Futoma, Mark Sendak, Blake Cameron, and Katherine Heller. "Predicting Disease Progression with a Model for Multivariate Longitudinal Clinical Data". In: *Proceedings of the 1st Machine Learning for Healthcare Conference*. PMLR, 2016-12, pp. 42–54. (Visited on 2023-06-05) (cit. on p. 46).

[Fut+20] Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. "The Myth of Generalisability in Clinical Research and Machine Learning in Health Care". In: *The Lancet Digital Health* 2.9 (2020-09), e489–e492. ISSN: 2589-7500. DOI: 10.1016/S2589-7500(20)30186-2. (Visited on 2025-07-30) (cit. on p. 24).

[Gaf+25] Adam Gaffney, Steffie Woolhandler, David U. Himmelstein, and Danny McCormick. "Health Care in the USA: Money Has Become the Mission". In: *The Lancet* 406.10519 (2025-11), pp. 2588–2600. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(25)01669-1. (Visited on 2025-12-01) (cit. on p. 147).

[Gal21] Soledad Galli. "Feature-Engine: A Python Package for Feature Engineering for Machine Learning". In: *Journal of Open Source Software* 6.65 (2021-09), p. 3642. ISSN: 2475-9066. DOI: 10.21105/joss.03642. (Visited on 2025-07-22) (cit. on p. 44).

[GM21] Jeff Gao and Aziz M. Merchant. "A Machine Learning Approach in Predicting Mortality Following Emergency General Surgery". In: *The American Surgeon™* 87.9 (2021-09), pp. 1379–1385. ISSN: 0003-1348. DOI: 10.1177/00031348211038568. (Visited on 2026-01-05) (cit. on p. 67).

[Gao+20] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*. 2020-12. DOI: 10.48550/arXiv.2101.00027. arXiv: 2101.00027 [cs]. (Visited on 2025-12-11) (cit. on pp. 139, 141).

[GYH22] Michael Gao, Jiayu Yao, and Ricardo Henao. "A Framework for the Evaluation of Clinical Time Series Models". In: *NeurIPS 2022 Workshop on Learning from Time Series for Health*. 2022-12. (Visited on 2025-07-28) (cit. on pp. 83, 87, 88, 91).

[Gel+03] G. Gell, P. Schmücker, M. Pedevilla, H. Leitner, J. Naumann, H. Fuchs, H. Pitz, and W. Köle. "SAP and Partners: IS-H™ and IS-H*MED™". In: *Methods of Information in Medicine* 42.1 (2003-02), pp. 16–24. ISSN: 0026-1270, 2511-705X. DOI: 10.1055/s-0038-1634205. (Visited on 2023-11-08) (cit. on p. 73).

[Gha+20] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L. Beam, Irene Y. Chen, and Rajesh Ranganath. "A Review of Challenges and Opportunities in Machine Learning for Health". In: *AMIA Summits on Translational Science Proceedings* 2020 (2020-05), pp. 191–200. ISSN: 2153-4063. (Visited on 2025-07-21) (cit. on p. 13).

[Gol+00] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals". In: *Circulation* 101.23 (2000-06). ISSN: 0009-7322, 1524-4539. DOI: 10.1161/01.CIR.101.23.e215 (cit. on pp. 13, 15, 152).

[Gra+23] Martin Graeßner, Bettina Jungwirth, Elke Frank, Stefan Josef Schaller, Eberhard Kochs, Kurt Ulm, Manfred Blobner, Bernhard Ulm, Armin Horst Podtschaske, and Simone Maria Kagerbauer. "Enabling Personalized Perioperative Risk Prediction by Using a Machine-Learning Model Based on Preoperative Data". In: *Scientific Reports* 13.1 (2023-05), p. 7128. ISSN: 2045-2322. DOI: 10.1038/s41598-023-33981-8. (Visited on 2026-01-05) (cit. on p. 67).

[GOV22a] Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. "Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data?" In: *Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2022-09. (Visited on 2023-06-06) (cit. on p. 85).

[GOV22b] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. *Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?* 2022-07. arXiv: 2207.08815 [cs, stat]. (Visited on 2022-12-16) (cit. on pp. 49, 82, 94).

[Gro24] MEDS Working Group. "Medical Event Data Standard (MEDS): Facilitating Machine Learning for Health". In: *ICLR 2024 Workshop on Learning from Time Series for Health*. 2024. (Visited on 2024-10-23) (cit. on pp. xv, 6, 99, 100, 142).

[Gue+25] Pablo Guerrero et al. "The AIR·MS Data Platform for Artificial Intelligence in Healthcare". In: *JAMIA Open* 8.6 (2025-12), ooaf145. ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooaf145. (Visited on 2025-12-19) (cit. on p. 146).

[GK18] Odd Erik Gundersen and Sigbjørn Kjensmo. "State of the Art: Reproducibility in Artificial Intelligence". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press, 2018-02, pp. 1644–1651. ISBN: 978-1-57735-800-8. DOI: 10.1609/aaai.v32i1.11503. (Visited on 2026-01-02) (cit. on p. 44).

[GLC20] Chonghui Guo, Menglin Lu, and Jingfeng Chen. "An Evaluation of Time Series Summary Statistics as Features for Clinical Prediction Tasks". In: *BMC Medical Informatics and Decision Making* 20.1 (2020-12), p. 48. ISSN: 1472-6947. DOI: 10.1186/s12911-020-1063-x. (Visited on 2022-12-06) (cit. on p. 46).

[Guo+24] Lin Lawrence Guo, Jason Fries, Ethan Steinberg, Scott Lanyon Fleming, Keith Morse, Catherine Aftandilian, Jose Posada, Nigam Shah, and Lillian Sung. "A Multi-Center Study on the Adaptability of a Shared Foundation Model for Electronic Health Records". In: *npj Digital Medicine* 7.1 (2024-06), pp. 1–9. ISSN: 2398-6352. DOI: 10.1038/s41746-024-01166-w. (Visited on 2024-11-21) (cit. on pp. 108, 117, 130).

[Gup+22] Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, and Rahmatollah Beheshti. *An Extensive Data Processing Pipeline for MIMIC-IV*. 2022-09. DOI: 10.48550/arXiv.2204.13841. arXiv: 2204.13841 [cs]. (Visited on 2022-10-24) (cit. on pp. 29, 31, 102).

[Haa+22] Camilla Haahr-Raunkjaer, Jesper Mølgaard, Mikkel Elvekjaer, Søren M. Rasmussen, Michael P. Achiam, Lars N. Jorgensen, Mette I. V. Søgaard, Katja K. Grønbaek, Anne-Britt Oxbøll, Helge B. D. Sørensen, Christian S. Meyhoff, and Eske K. Aasvang. "Continuous Monitoring of Vital Sign Abnormalities; Association to Clinical Complications in 500 Postoperative Patients". In: *Acta Anaesthesiologica Scandinavica* 66.5 (2022-05), pp. 552–562. ISSN: 1399-6576. DOI: 10.1111/aas.14048 (cit. on p. 93).

[Har+16] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. "Equality of Opportunity in Supervised Learning". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2016. (Visited on 2023-08-14) (cit. on p. 210).

[Har+19] Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. "Multitask Learning and Benchmarking with Clinical Time Series Data". In: *Scientific Data* 6.1 (2019-12), p. 96. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0103-9. (Visited on 2022-03-09) (cit. on pp. 28, 29, 31).

[Has+25] Fereshteh Hasanzadeh, Colin B. Josephson, Gabriella Waters, Demilade Adedinsewo, Zahra Azizi, and James A. White. "Bias Recognition and Mitigation Strategies in Artificial Intelligence Healthcare Applications". In: *npj Digital Medicine* 8.1 (2025-03), p. 154. ISSN: 2398-6352. DOI: 10.1038/s41746-025-01503-7. (Visited on 2025-11-30) (cit. on p. 148).

[Hea+21] Tim Head, Manoj Kumar, Holger Nahrstaedt, Gilles Louppe, and Iaroslav Shcherbatyi. *Scikit-Optimize/Scikit-Optimize*. Zenodo. 2021-10. DOI: 10.5281/zenodo.5565057. (Visited on 2023-05-30) (cit. on p. 41).

[Hea+24] Garrett L. Healy, Christina M. Stuart, Adam R. Dyas, Michael R. Bronsert, Robert A. Meguid, Tochi Anioke, Ahmad M. Hider, Richard D. Schulick, and William G. Henderson. "Association between Postoperative Complications and Hospital Length of Stay: A Large-Scale Observational Study of 4,495,582 Patients in the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) Registry". In: *Patient Safety in Surgery* 18.1 (2024-10), p. 29. ISSN: 1754-9493. DOI: 10.1186/s13037-024-00409-9. (Visited on 2025-10-30) (cit. on p. 93).

[Heg+25] Stefan Hegselmann, Georg von Arnim, Tillmann Rheude, Noel Kronenberg, David Sontag, Gerhard Hindricks, Roland Eils, and Benjamin Wild. *Large Language Models Are Powerful Electronic Health Record Encoders*. 2025-10. DOI: 10.48550/arXiv.2502.17403. arXiv: 2502.17403 [cs]. (Visited on 2025-11-28) (cit. on pp. 140, 143).

[Hen+25] Elisa Henke, Stephan Lorenz, Michele Zoch, Martin Sedlmayr, and Yuan Peng. "Mapping National Vocabularies to International Standards Using OHDSI Standardized Vocabularies". In: *Envisioning the Future of Health Informatics and Digital Health*. IOS Press, 2025, pp. 349–353. DOI: 10.3233/SHTI250110. (Visited on 2025-12-03) (cit. on p. 150).

[Her+15] Emily Herrett, Arlene M Gallagher, Krishnan Bhaskaran, Harriet Forbes, Rohini Mathur, Tjeerd van Staa, and Liam Smeeth. "Data Resource Profile: Clinical Practice Research Datalink (CPRD)". In: *International Journal of Epidemiology* 44.3 (2015-06), pp. 827–836. ISSN: 0300-5771. DOI: 10.1093/ije/dyv098. (Visited on 2025-10-24) (cit. on pp. 22, 106).

[Hil+19] Brian L. Hill, Robert Brown, Eilon Gabel, Nadav Rakocz, Christine Lee, Maxime Cannesson, Pierre Baldi, Loes Olde Loohuis, Ruth Johnson, Brandon Jew, Uri Maoz, Aman Mahajan, Sriram Sankararaman, Ira Hofer, and Eran Halperin. "An Automated Machine Learning-Based Model Predicts Postoperative Mortality Using Readily-Extractable Preoperative Electronic Health Record Data". In: *British Journal of Anaesthesia* 123.6 (2019-12), pp. 877–886. ISSN: 0007-0912. DOI: 10.1016/j.bja.2019.07.030. (Visited on 2026-01-05) (cit. on p. 67).

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. (Visited on 2023-08-09) (cit. on p. 207).

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997-11), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735 (cit. on pp. 21, 47, 85, 199).

[Hom15] Homer Homer. *The Odyssey*. Xist Publishing, 2015-04. ISBN: 978-1-62395-832-9 (cit. on p. 139).

[HGV25] Mohammad Hosseini, Peng Gao, and Carolina Vivas-Valencia. "A Social-Environmental Impact Perspective of Generative Artificial Intelligence". In: *Environmental Science and Ecotechnology* 23 (2025-01), p. 100520. ISSN: 2666-4984. DOI: 10.1016/j.ese.2024.100520. (Visited on 2025-12-01) (cit. on p. 148).

[Hu19] Xiao Hu. "An Algorithm Strategy for Precise Patient Monitoring in a Connected Healthcare Enterprise". In: *npj Digital Medicine* 2.1 (2019-04), p. 30. ISSN: 2398-6352. DOI: 10.1038/s41746-019-0107-z. (Visited on 2025-11-06) (cit. on p. 95).

[Hua+25] Hongzhi Huang, Defa Zhu, Banggu Wu, Yutao Zeng, Ya Wang, Qiyang Min, and Zhou Xun. "Over-Tokenized Transformer: Vocabulary Is Generally Worth Scaling". In: *Forty-Second International Conference on Machine Learning*. 2025-06. (Visited on 2026-01-20) (cit. on p. 140).

[Hua+21] Wei Huang, Yuwen Chen, Peng Wang, Xiang Liu, and Shuguang Liu. "An Interpretable Temporal Convolutional Network Model for Acute Kidney Injury Prediction in the Intensive Care Unit". In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2021-12, pp. 3021–3028. DOI: 10.1109/BIBM52615.2021.9669653 (cit. on p. 46).

[Huf+98] Stanley M. Huff et al. "Development of the Logical Observation Identifier Names and Codes (LOINC) Vocabulary". In: *Journal of the American Medical Informatics Association* 5.3 (1998-05), pp. 276–292. ISSN: 1067-5027. DOI: 10.1136/jamia.1998.0050276. (Visited on 2025-10-28) (cit. on pp. 103, 118).

[Hur+24] Kyunghoon Hur, Jungwoo Oh, Junu Kim, Jiyoun Kim, Min Jae Lee, Eunbyeol Cho, Seong-Eun Moon, Young-Hak Kim, Louis Atallah, and Edward Choi. "GenHPF: General Healthcare Predictive Framework for Multi-Task Multi-Source Learning". In: *IEEE Journal of Biomedical and Health Informatics* 28.1 (2024-01), pp. 502–513. ISSN: 2168-2208. DOI: 10.1109/JBHI.2023.3327951. (Visited on 2025-10-12) (cit. on pp. 22, 106, 121, 125, 130).

[Hyl20] Stephanie L Hyland. "Early Prediction of Circulatory Failure in the Intensive Care Unit Using Machine Learning". In: *Nature MedIcIne* 26 (2020), p. 28 (cit. on pp. 9, 16, 28, 30, 33, 34, 47, 56, 57, 92, 93, 104, 109, 121).

[Hym+07] Neil Hyman, Thomas L. Manchester, Turner Osler, Betsy Burns, and Peter A. Cataldo. "Anastomotic Leaks After Intestinal Anastomosis". In: *Annals of Surgery* 245.2 (2007-02), pp. 254–258. ISSN: 0003-4932. DOI: 10.1097/01.sla.0000225083.27182.85. (Visited on 2025-09-14) (cit. on p. 56).

[Imb04] Guido W. Imbens. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review". In: *The Review of Economics and Statistics* 86.1 (2004-02), pp. 4–29. ISSN: 0034-6535. DOI: 10.1162/003465304323023651. (Visited on 2025-08-11) (cit. on p. 20).

[J+19] Eric J., Zachary Barry, Sam Zuckerman, and Zachary Sailer. "Pyjanitor: A Cleaner API for Cleaning Data". In: *Proceedings of the Python in Science Conference*. Austin, Texas: SciPy, 2019, pp. 50–53. DOI: 10.25080/majora-7ddc1dd1-007. (Visited on 2025-07-22) (cit. on p. 44).

[Jan25] Eugene Jang. "Oops? Interdisciplinary Stories of Sociotechnical Error| When Faulty AI Falls Into the Wrong Hands: The Risks of Erroneous AI-Driven Healthcare Decisions". In: *International Journal of Communication* 19 (2025-04), pp. 7–7. ISSN: 1932-8036. (Visited on 2025-12-01) (cit. on p. 147).

[Jar+21] Daniel Jarrett, Ioana Bica, Ari Ercole, Jinsung Yoon, and Zhaozhi Qian. "CLAIRVOYANCE: A PIPELINE TOOLKIT FOR MEDICAL TIME SERIES". In: (2021), p. 32 (cit. on pp. 29, 31, 41, 102, 149).

[Jar+22] Daniel Jarrett, Bogdan Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. *HyperImpute: Generalized Iterative Imputation with Automatic Model Selection*. 2022-06. DOI: 10.48550/arXiv.2206.07769. arXiv: 2206.07769 [cs, stat]. (Visited on 2022-10-19) (cit. on p. 207).

[Jea24] Vincent Jeanselme. "Clinical Presence: Impact on Predictive Modelling and Algorithmic Fairness". In: (2024-09). DOI: 10.17863/CAM.112135. (Visited on 2025-07-29) (cit. on pp. 15, 24).

[Jen+25] Marie Said Vang Jensen, Vibeke Ramsgaard Eriksen, Søren Straarup Rasmussen, Christian Sylvest Meyhoff, and Eske Kvanner Aasvang. "Time to Detection of Serious Adverse Events by Continuous Vital Sign Monitoring versus Clinical Practice". In: *Acta Anaesthesiologica Scandinavica* 69.1 (2025), e14541. ISSN: 1399-6576. DOI: 10.1111/aas.14541. (Visited on 2025-08-27) (cit. on p. 59).

[Jha+08] S. Jhanji, B. Thomas, A. Ely, D. Watson, C. J. Hinds, and R. M. Pearse. "Mortality and Utilisation of Critical Care Resources amongst High-Risk Surgical Patients in a Large NHS Trust". In: *Anaesthesia* 63.7 (2008), pp. 695–700. ISSN: 1365-2044. DOI: 10.1111/j.1365-2044.2008.05560.x. (Visited on 2026-01-05) (cit. on p. 68).

[Jin+23] Senjun Jin, Lin Chen, Kun Chen, Chaozhou Hu, Sheng'an Hu, and Zhongheng Zhang. "Establishment of a Chinese Critical Care Database from Electronic Healthcare Records in a Tertiary Care Medical Center". In: *Scientific Data* 10.1 (2023-01), p. 49. ISSN: 2052-4463. DOI: 10.1038/s41597-023-01952-3. (Visited on 2023-05-30) (cit. on p. 33).

[Joh+23a] Marco Johns, Thierry Meurers, Felix N. Wirth, Anna C. Haber, Armin Müller, Mehmed Halilovic, Felix Balzer, and Fabian Prasser. "Data Provenance in Biomedical Research: Scoping Review". In: *Journal of Medical Internet Research* 25.1 (2023-03), e42289. DOI: 10.2196/42289. (Visited on 2025-07-25) (cit. on pp. 16, 149).

[Joh+22] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, S Horng, LA Celi, and R Mark. *MIMIC-IV*. PhysioNet. 2022-06 (cit. on pp. 9, 22, 29, 57, 104–106).

[Joh+23b] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. *MIMIC-IV Clinical Database Demo*. 2023. DOI: 10.13026/DP1F-EX47. (Visited on 2025-10-27) (cit. on p. 113).

[JPM19] Alistair Johnson, Tom Pollard, and Roger Mark. *MIMIC-III Clinical Database Demo*. 2019. DOI: 10.13026/C2HM2Q. (Visited on 2022-12-05) (cit. on p. 16).

[Joh+23c] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. "MIMIC-IV, a Freely Accessible Electronic Health Record Dataset". In: *Scientific Data* 10.1 (2023-01), p. 1. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01899-x. (Visited on 2023-04-13) (cit. on pp. 15, 16, 33, 34).

[JPM17] Alistair E. W. Johnson, Tom J. Pollard, and Roger G. Mark. "Reproducibility in Critical Care: A Mortality Prediction Case Study". In: *Proceedings of the 2nd Machine Learning for Healthcare Conference*. PMLR, 2017-11, pp. 361–376. (Visited on 2023-04-18) (cit. on pp. 27, 29, 31, 44).

[Joh+16] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. "MIMIC-III, a Freely Accessible Critical Care Database". In: *Scientific Data* 3.1 (2016-12), p. 160035. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.35. (Visited on 2022-03-09) (cit. on pp. 30, 33, 34, 101, 105).

[Jor13] Davide Anguita Jorge Reyes-Ortiz. *Human Activity Recognition Using Smartphones*. 2013. DOI: 10.24432/C54S4K. (Visited on 2026-01-17) (cit. on p. 15).

[Jos+25] Shalmali Joshi, Iñigo Urteaga, Wouter AC Van Amsterdam, George Hripcsak, Pierre Elias, Benjamin Recht, Noémie Elhadad, James Fackler, Mark P. Sendak, and Jenna Wiens. "AI as an Intervention: Improving Clinical Outcomes Relies on a Causal Approach to AI Development and Validation". In: *Journal of the American Medical Informatics Association* 32.3 (2025), pp. 589–594. (Visited on 2025-07-26) (cit. on p. 23).

[JM21] Rachel Hae-Soo Joung and Ryan P. Merkow. "Is It Time to Abandon 30-Day Mortality as a Quality Measure?" In: *Annals of Surgical Oncology* 28.3 (2021-03), pp. 1263–1264. ISSN: 1534-4681. DOI: 10.1245/s10434-020-09262-3. (Visited on 2026-01-05) (cit. on p. 67).

[JJL18] Gordana Jovanović, Dea Karaba Jakovljević, and Mirka Lukić-Šarkanović. "Enhanced Recovery in Surgical Intensive Care: A Review". In: *Frontiers in Medicine* 5 (2018). ISSN: 2296-858X. (Visited on 2024-02-16) (cit. on pp. 57, 95).

[Joy22] James Joyce. *Ulysses by James Joyce Paperback Book*. Sahara Publisher Books, 1922-02. ISBN: 978-2-38226-147-7 (cit. on p. 139).

[Jun+23] Jin-On Jung, Juan I Pisula, Kasia Bozek, Felix Popp, Hans F Fuchs, Wolfgang Schröder, Christiane J Bruns, and Thomas Schmidt. "Prediction of Postoperative Complications after Oesophagectomy Using Machine-Learning Methods". In: *BJS* 110.10 (2023-10), pp. 1361–1366. ISSN: 1365-2168. DOI: 10.1093/bjs/znad181. (Visited on 2026-01-05) (cit. on p. 67).

[Jus+06] Amy C. Justice, Joseph Erdos, Cynthia Brandt, Joseph Conigliaro, William Tierney, and Kendall Bryant. "The Veterans Affairs Healthcare System: A Unique Laboratory for Observational and Interventional Research". In: *Medical Care* 44.8 (2006-08), S7. ISSN: 0025-7079. DOI: 10.1097/01.mlr.0000228027.80012.c5. (Visited on 2025-10-24) (cit. on pp. 15, 106).

[K25] Boriharn K. *Finding the Best Color-blind Friendly Palette on Python Seaborn*. 2025-08. (Visited on 2026-01-16) (cit. on pp. 136, 291).

[KVM17] Tadashi Kamio, Tomoaki Van, and Ken Masamune. "Use of Machine-Learning Approaches to Predict Clinical Deterioration in Critically Ill Patients: A Systematic Review". In: *INTERNATIONAL JOURNAL OF MEDICAL RESEARCH & HEALTH SCIENCES* 6.6 (2017), pp. 1–7. ISSN: 2319-5886 (cit. on p. 25).

[KN23] Sayash Kapoor and Arvind Narayanan. "Leakage and the Reproducibility Crisis in Machine-Learning-Based Science". In: *Patterns* 4.9 (2023-09), p. 100804. ISSN: 2666-3899. DOI: 10.1016/j.patter.2023.100804. (Visited on 2025-12-03) (cit. on pp. 148, 149).

[Kap+24] Sayash Kapoor et al. "REFORMS: Consensus-based Recommendations for Machine-learning-based Science". In: *Science Advances* 10.18 (2024-05), eadk3452. DOI: 10.1126/sciadv.adk3452. (Visited on 2025-12-03) (cit. on p. 149).

[Kat+16] Hiroshi Katayama, Yukinori Kurokawa, Kenichi Nakamura, Hiroyuki Ito, Yukihide Kanemitsu, Norikazu Masuda, Yasuhiro Tsubosa, Toyomi Satoh, Akira Yokomizo, Haruhiko Fukuda, and Mitsuru Sasako. "Extended Clavien-Dindo Classification of Surgical Complications: Japan Clinical Oncology Group Postoperative Complications Criteria". In: *Surgery Today* 46 (2016), pp. 668–685. ISSN: 0941-1291. DOI: 10.1007/s00595-015-1236-x. (Visited on 2025-09-22) (cit. on p. 71).

[Kat+24] Jakob Nikolas Kather, Dyke Ferber, Isabella C. Wiest, Stephen Gilbert, and Daniel Truhn. "Large Language Models Could Make Natural Language Again the Universal Interface of Healthcare". In: *Nature Medicine* 30.10 (2024-10), pp. 2708–2710. ISSN: 1546-170X. DOI: 10.1038/s41591-024-03199-w. (Visited on 2025-12-14) (cit. on p. 143).

[Kaz+19] Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. *Time2Vec: Learning a Vector Representation of Time*. 2019-07. DOI: 10.48550/arXiv.1907.05321. arXiv: 1907.05321 [cs]. (Visited on 2025-10-24) (cit. on p. 107).

[KDI12] KDIGO. "Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group: KDIGO Clinical Practice Guideline for Acute Kidney Injury." In: *Kidney International Supplements* 2.1 (2012-03). ISSN: 21571716. DOI: 10.1038/kisup.2012.2 (cit. on pp. 28, 30, 35, 195).

[Ke+17] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. (Visited on 2023-09-25) (cit. on pp. 47, 82, 85, 94, 132, 205, 211).

[Kel+19a] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. "Key Challenges for Delivering Clinical Impact with Artificial Intelligence". In: *BMC Medicine* 17.1 (2019-12), p. 195. ISSN: 1741-7015. DOI: 10.1186/s12916-019-1426-2 (cit. on p. 27).

[Kel+19b] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. "Key Challenges for Delivering Clinical Impact with Artificial Intelligence". In: *BMC Medicine* 17.1 (2019-12), p. 195. ISSN: 1741-7015. DOI: 10.1186/s12916-019-1426-2. (Visited on 2023-04-12) (cit. on p. 44).

[Kem+00] B. Kemp, A.H. Zwinderman, B. Tuk, H.A.C. Kamphuisen, and J.J.L. Oberye. "Analysis of a Sleep-Dependent Neuronal Feedback Loop: The Slow-Wave Microcontinuity of the EEG". In: *IEEE Transactions on Biomedical Engineering* 47.9 (2000-09), pp. 1185–1194. ISSN: 1558-2531. DOI: 10.1109/10.867928. (Visited on 2026-01-17) (cit. on p. 15).

[Kha+25] Wasif Khan, Seowung Leem, Kyle B. See, Joshua K. Wong, Shaoting Zhang, and Ruogu Fang. "A Comprehensive Survey of Foundation Models in Medicine". In: *IEEE reviews in biomedical engineering* PP (2025-05). ISSN: 1941-1189. DOI: 10.1109/RBME.2025.3531360 (cit. on p. 100).

[Kir+25a] Matthias Kirchler, Matteo Ferro, Veronica Lorenzini, Christoph Lippert, and Andrea Ganna. "Large Language Models Improve Transferability of Electronic Health Record-Based Predictions across Countries and Coding Systems". In: *medRxiv* (2025-02), p. 2025.02.03.25321597. DOI: 10.1101/2025.02.03.25321597. (Visited on 2025-08-15) (cit. on pp. 119, 130, 143).

[Kir+25b] Matthias Kirchler, Matteo Ferro, Veronica Lorenzini, Robin van de Water, Christoph Lippert, and Andrea Ganna. "Large Language Models Improve Transferability of Electronic Health Record-Based Predictions across Countries and Coding Systems". In: *Accepted to npj Digital Medicine* (2025) (cit. on pp. xiv, 10, 140).

[Kni+21] Stephen R. Knight, Nathan Ng, Athanasios Tsanas, Kenneth Mclean, Claudia Pagliari, and Ewen M. Harrison. "Mobile Devices and Wearable Technology for Measuring Patient Outcomes after Surgery: A Systematic Review". In: *npj Digital Medicine* 4.1 (2021-11), p. 157. ISSN: 2398-6352. DOI: 10.1038/s41746-021-00525-1. (Visited on 2025-08-26) (cit. on p. 58).

[Koc+11] Moritz Koch et al. "Bile Leakage after Hepatobiliary and Pancreatic Surgery: A Definition and Grading of Severity by the International Study Group of Liver Surgery". In: *Surgery* 149.5 (2011-05), pp. 680–688. ISSN: 1532-7361. DOI: 10.1016/j.surg.2010.12.002 (cit. on pp. 56, 71).

[Kok+20] Christopher Kok, V. Jahmunah, Shu Lih Oh, Xujuan Zhou, Raj Gururajan, Xiaohui Tao, Kang Hao Cheong, Rashmi Gururajan, Filippo Molinari, and U.Rajendra Acharya. "Automated Prediction of Sepsis Using Temporal Convolutional Network". In: *Computers in Biology and Medicine* 127 (2020-12), p. 103957. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2020.103957. (Visited on 2022-04-28) (cit. on p. 46).

[Kon+23a] Orhan Konak, Robin van de Water, Valentin Döring, Tobias Fiedler, Lucas Liebe, Leander Masopust, Kirill Postnov, Franz Sauerwald, Felix Treykorn, and Alexander Wischmann. "HARE: Unifying the Human Activity Recognition Engineering Workflow". In: *Sensors* 23.23 (2023), p. 9571. (Visited on 2024-10-23) (cit. on p. 10).

[Kon+23b] Orhan Konak, Alexander Wischmann, Robin van de Water, and Bert Arnrich. "A Real-time Human Pose Estimation Approach for Optimal Sensor Placement in Sensor-based Human Activity Recognition". In: *Proceedings of the 8th International Workshop on Sensor-Based Activity Recognition and Artificial Intelligence*. Lübeck Germany: ACM, 2023-09, pp. 1–6. ISBN: 979-8-4007-0816-9. DOI: 10.1145/3615834.3615848. (Visited on 2024-10-23) (cit. on p. 10).

[Kon+21a] Junjie Kong, Guangbing Li, Jiawei Chai, Guangsheng Yu, Yong Liu, and Jun Liu. "ASO Visual Abstract: Impact of Postoperative Complications on Long-Term Survival After Resection of Hepatocellular Carcinoma—A Systematic Review and Meta-Analysis". In: *Annals of Surgical Oncology* 28.S3 (2021-12), pp. 499–500. ISSN: 1068-9265, 1534-4681. DOI: 10.1245/s10434-021-10341-2. (Visited on 2025-09-18) (cit. on p. 56).

[Kon+21b] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. *DiffWave: A Versatile Diffusion Model for Audio Synthesis*. 2021-03. arXiv: 2009.09761 [cs, eess, stat]. (Visited on 2023-03-07) (cit. on p. 207).

[Koy+18a] Jay L. Koyner, Kyle A. Carey, Dana P. Edelson, and Matthew M. Churpek. "The Development of a Machine Learning Inpatient Acute Kidney Injury Prediction Model*:" in: *Critical Care Medicine* 46.7 (2018-07), pp. 1070–1077. ISSN: 0090-3493. DOI: 10.1097/CCM.0000000000003123 (cit. on p. 30).

[Koy+18b] Jay L. Koyner, Kyle A. Carey, Dana P. Edelson, and Matthew M. Churpek. "The Development of a Machine Learning Inpatient Acute Kidney Injury Prediction Model*:" in: *Critical Care Medicine* 46.7 (2018-07), pp. 1070–1077. ISSN: 0090-3493. DOI: 10.1097/CCM.0000000000003123. (Visited on 2022-11-29) (cit. on p. 46).

[Kra+24] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Balston, Jack Ross, Esther Idowu, James T. Teo, and Richard J. B. Dobson. "Foresight—a Generative Pretrained Transformer for Modelling of Patient Timelines Using Electronic Health Records: A Retrospective Modelling Study". In: *The Lancet Digital Health* 6.4 (2024-04), e281–e290. ISSN: 2589-7500. DOI: 10.1016/S2589-7500(24)00025-6. (Visited on 2025-08-13) (cit. on pp. 22, 106, 107).

[Kra+21] Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. *MedGPT: Medical Concept Prediction from Clinical Narratives*. 2021-07. DOI: 10.48550/arXiv.2107.03134. arXiv: 2107.03134 [cs]. (Visited on 2025-08-13) (cit. on pp. 22, 106, 107).

[Kri+22] Amrit Krishnan, Vallijah Subasri, Kaden McKeen, Ali Kore, Franklin Ogidi, Mahshid Alinoori, Nadim Lalani, Azra Dhalla, Amol Verma, Fahad Razak, Deval Pandya, and Elham Dolatabadi. *CyclOps: Cyclical Development towards Operationalizing ML Models for Health*. Preprint. Health Informatics, 2022-12. DOI: 10.1101/2022.12.02.22283021. (Visited on 2023-05-02) (cit. on p. 96).

[Kuh+24] Max Kuhn, Hadley Wickham, Emil Hvitfeldt, Posit Software, and PBC. *Recipes: Preprocessing and Feature Engineering Steps for Modeling*. 2024-07. (Visited on 2024-07-09) (cit. on p. 44).

[Kul19] Casimir A. Kulikowski. "Beginnings of Artificial Intelligence in Medicine (AIM): Computational Artifice Assisting Scientific Inquiry and Clinical Art – with Reflections on Present AIM Challenges". In: *Yearbook of Medical Informatics* 28.1 (2019-08), pp. 249–256. ISSN: 0943-4747. DOI: 10.1055/s-0039-1677895. (Visited on 2025-07-21) (cit. on p. 13).

[Kun+21] Kyle N Kunze, Laura M Krivicich, Ian M Clapp, Blake M Bodendorfer, Benedict U Nwachukwu, Jorge Chahla, and Shane J Nho. "Machine Learning Algorithms Predict Achievement of Clinically Significant Outcomes after Orthopaedic Surgery: A Systematic Review." In: *Arthroscopy : the journal of arthroscopic & related surgery : official publication of the Arthroscopy Association of North America and the International Arthroscopy Association* (2021). ISSN: 1526-3231. DOI: 10.1016/j.arthro.2021.12.030 (cit. on p. 25).

[Lam+25] Siu Kwan Lam et al. *Numba/Numba: 0.63.1*. Zenodo. 2025-12. DOI: 10.5281/zenodo.17874185. (Visited on 2026-01-04) (cit. on p. 77).

[Lam+23] Sophie Isabelle Lambert, Murielle Madi, Saša Sopka, Andrea Lenes, Hendrik Stange, Claus-Peter Buszello, and Astrid Stephan. "An Integrative Review on the Acceptance of Artificial Intelligence among Healthcare Professionals in Hospitals". In: *npj Digital Medicine* 6.1 (2023-06), pp. 1–14. ISSN: 2398-6352. DOI: 10.1038/s41746-023-00852-5. (Visited on 2024-12-20) (cit. on p. 56).

[Lau+20] Simon Meyer Lauritsen, Mads Ellersgaard Kalør, Emil Lund Kongsgaard, Katrine Meyer Lauritsen, Marianne Johansson Jørgensen, Jeppe Lange, and Bo Thiesson. "Early Detection of Sepsis Utilizing Deep Learning on Electronic Health Record Event Sequences". In: *Artificial Intelligence in Medicine* 104 (2020-04), p. 101820. ISSN: 09333657. DOI: 10.1016/j.artmed.2020.101820. (Visited on 2022-02-04) (cit. on p. 46).

[LRH25] John Lawrence, Mike Rayo, and Timothy Huerta. *ALarms, Outcomes Telemetry with Timing (ALOTT): A Bedside-EMR Database*. 2025-03. DOI: 10.13026/SBQ5-DY17. (Visited on 2025-04-23) (cit. on p. 16).

[Lee+22a] Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. "VitalDB, a High-Fidelity Multi-Parameter Vital Signs Database in Surgical Patients". In: *Scientific Data* 9.1 (2022-06), p. 279. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01411-5. (Visited on 2024-11-06) (cit. on p. 16).

[Lee+22b] Seung Wook Lee, Hyung-Chul Lee, Jungyo Suh, Kyung Hyun Lee, Heonyi Lee, Suryang Seo, Tae Kyong Kim, Sang-Wook Lee, and Yi-Jun Kim. "Multi-Center Validation of Machine Learning Model for Preoperative Prediction of Postoperative Mortality". In: *npj Digital Medicine* 5.1 (2022-07), p. 91. ISSN: 2398-6352. DOI: 10.1038/s41746-022-00625-6. (Visited on 2026-01-05) (cit. on p. 67).

[Lee+24] Jobbe P L Leenen, Vera Ardesch, Cor J Kalkman, Lisette Schoonhoven, and Gijs A Patijn. "Impact of Wearable Wireless Continuous Vital Sign Monitoring in Abdominal Surgical Patients: Before–after Study". In: *BJS Open* 8.1 (2024-01), zrad128. ISSN: 2474-9842. DOI: 10.1093/bjsopen/zrad128. (Visited on 2024-12-17) (cit. on pp. 57, 94).

[Lee+22c] Jobbe P. L. Leenen, Henriëtte J. M. Rasing, Joris D. van Dijk, Cor J. Kalkman, Lisette Schoonhoven, and Gijs A. Patijn. "Feasibility of Wireless Continuous Monitoring of Vital Signs without Using Alarms on a General Surgical Ward: A Mixed Methods Study". In: *PLOS ONE* 17.3 (2022-03), e0265435. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0265435. (Visited on 2025-10-30) (cit. on p. 93).

[AJ25] Obada Al-leimon and Malik Eid Juweid. ""Publish or Perish" Paradigm and Medical Research: Replication Crisis in the Context of Artificial Intelligence Trend". In: *Annals of Biomedical Engineering* 53.1 (2025-01), pp. 3–4. ISSN: 1573-9686. DOI: 10.1007/s10439-024-03625-7. (Visited on 2025-07-28) (cit. on p. 24).

[LNA17] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: *Journal of Machine Learning Research* 18.17 (2017), pp. 1–5 (cit. on pp. 43, 62).

[Les+16] T. M. Leslie-Mazwi, J. A. Bello, R. Tu, G. N. Nicola, W. D. Donovan, R. M. Barr, and J. A. Hirsch. "Current Procedural Terminology: History, Structure, and Relationship to Valuation for the Neuroradiologist". In: *American Journal of Neuroradiology* 37.11 (2016-11), pp. 1972–1976. ISSN: 0195-6108, 1936-959X. DOI: 10.3174/ajnr.A4863. (Visited on 2025-12-19) (cit. on p. 103).

[Lev+03] Mitchell M. Levy, Mitchell P. Fink, John C. Marshall, Edward Abraham, Derek Angus, Deborah Cook, Jonathan Cohen, Steven M. Opal, Jean-Louis Vincent, Graham Ramsay, and International Sepsis Definitions Conference. "2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference". In: *Intensive Care Medicine* 29.4 (2003-04), pp. 530–538. ISSN: 0342-4642. DOI: 10.1007/s00134-003-1662-x (cit. on p. 50).

[LZY20] Haomin Li, Xian Zeng, and Gang Yu. *Paediatric Intensive Care Database*. 2020-11. DOI: 10.13026/32X9-WV38. (Visited on 2025-04-23) (cit. on p. 16).

[Li+25] Pengfei Li, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. "Making AI Less 'Thirsty'". In: *Commun. ACM* 68.7 (2025-06), pp. 54–61. ISSN: 0001-0782. DOI: 10.1145/3724499. (Visited on 2025-12-01) (cit. on p. 148).

[Li+20] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. "BEHRT: Transformer for Electronic Health Records". In: *Scientific Reports* 10.1 (2020-04), p. 7155. ISSN: 2045-2322. DOI: 10.1038/s41598-020-62922-y. (Visited on 2025-07-21) (cit. on pp. 13, 22, 106, 107).

[Lia+25] Qi-Ming Liao, Wahab Hussain, Zhong-Xia Liao, Sarfraz Hussain, Zhi-Liang Jiang, Yong-Hao Zhu, Huang-Yin Luo, Xin-Ying Ji, Hong-Wei Wen, and Dong-Dong Wu. "Computer-Aided Application in Medicine and Biomedicine". In: *International Journal of Computational Intelligence Systems* 18.1 (2025-08), p. 221. ISSN: 1875-6883. DOI: 10.1007/s44196-025-00936-y. (Visited on 2026-01-21) (cit. on p. 13).

[Lim+21] Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. "Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting". In: *International Journal of Forecasting* 37.4 (2021-10), pp. 1748–1764. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2021.03.012. (Visited on 2023-08-09) (cit. on pp. 206, 207).

[Lim+24] Leerang Lim, Hyeonhoon Lee, Chul-Woo Jung, Dayeon Sim, Xavier Borrat, Tom J. Pollard, Leo A. Celi, Roger G. Mark, Simon T. Vistisen, and Hyung-Chul Lee. "INSPIRE, a Publicly Available Research Dataset for Perioperative Medicine". In: *Scientific Data* 11.1 (2024-06), p. 655. ISSN: 2052-4463. DOI: 10.1038/s41597-024-03517-4. (Visited on 2025-04-23) (cit. on pp. 9, 16, 104, 109, 121).

[LHM93] D. A. Lindberg, B. L. Humphreys, and A. T. McCray. "The Unified Medical Language System". In: *Methods of Information in Medicine* 32.4 (1993-08), pp. 281–291. ISSN: 0026-1270. DOI: 10.1055/s-0038-1634945 (cit. on p. 13).

[Liu+21] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. "Self-Alignment Pretraining for Biomedical Entity Representations". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou. Online: Association for Computational Linguistics, 2021-06, pp. 4228–4238. DOI: 10.18653/v1/2021.naacl-main.334. (Visited on 2024-06-11) (cit. on p. 118).

[Lom+22] Sara Lombardi, Petri Partanen, Piergiorgio Francia, Italo Calamai, Rossella Deodati, Marco Luchini, Rosario Spina, and Leonardo Bocchi. "Classifying Sepsis from Photoplethysmography". In: *Health Information Science and Systems* 10.1 (2022-12), p. 30. ISSN: 2047-2501. DOI: 10.1007/s13755-022-00199-3 (cit. on p. 94).

[Lon+13] John Lonsdale et al. "The Genotype-Tissue Expression (GTEx) Project". In: *Nature Genetics* 45.6 (2013-06), pp. 580–585. ISSN: 1546-1718. DOI: 10.1038/ng.2653. (Visited on 2026-01-17) (cit. on p. 15).

[Lu+22] Sheng-Chieh Lu, Cai Xu, Chandler H. Nguyen, Yimin Geng, André Pfob, and Chris Sidey-Gibbons. "Machine Learning–Based Short-Term Mortality Prediction Models for Patients With Cancer Using Electronic Health Record Data: Systematic Review and Critical Appraisal". In: *JMIR Medical Informatics* 10.3 (2022-03), e33182. DOI: 10.2196/33182. (Visited on 2023-04-18) (cit. on p. 46).

[Lud22] Guy L. Ludbrook. "The Hidden Pandemic: The Cost of Postoperative Complications". In: *Current Anesthesiology Reports* 12.1 (2022), pp. 1–9. ISSN: 1523-3855. DOI: 10.1007/s40140-021-00493-y (cit. on p. 93).

[LL17a] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017 (cit. on pp. 43, 91).

[LL17b] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017 (cit. on pp. 62, 85, 87).

[Lun+20] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. "From Local Explanations to Global Understanding with Explainable AI for Trees". In: *Nature machine intelligence* 2.1 (2020-01), pp. 56–67. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0138-9. (Visited on 2025-11-06) (cit. on p. 95).

[LL25] Xiaolong Luo and Michael Lingzhi Li. *The CRITICAL Records Integrated Standardization Pipeline (CRISP): End-to-End Processing of Large-scale Multi-institutional OMOP CDM Data*. 2025-09. DOI: 10.48550/arXiv.2509.08247. arXiv: 2509.08247 [cs]. (Visited on 2025-10-11) (cit. on p. 102).

[Luo+20] Xueying Luo, Feng Li, Haofei Hu, Baoer Liu, Sujing Zheng, Liping Yang, Rui Gao, Ya Li, Rao Xi, and Jinsong He. "Anemia and Perioperative Mortality in Non-Cardiac Surgery Patients: A Secondary Analysis Based on a Single-Center Retrospective Study". In: *BMC Anesthesiology* 20.1 (2020-05), p. 112. ISSN: 1471-2253. DOI: 10.1186/s12871-020-01024-8. (Visited on 2026-01-05) (cit. on p. 68).

[Maa+20] Lientje Maas, Mathan Geurtsen, Florian Nouwt, Stefan F. Schouten, Robin van de Water, Sandra Van Dulmen, Fabiano Dalpiaz, Kees Van Deemter, and Sjaak Brinkkemper. "The Care2Report System: Automated Medical Reporting as an Integrated Solution to Reduce Administrative Burden in Healthcare." In: *HICSS*. 2020, pp. 1–10. (Visited on 2024-10-23) (cit. on p. 10).

[Mab+25] Laurence Mabile, Hanna Shmagun, Christopher Erdmann, Anne Cambon-Thomsen, Mogens Thomsen, and Florencia Grattarola. "Recommendations on Open Science Rewards and Incentives: Guidance for Multiple Stakeholders in Research | Data Science Journal". In: (2025-01). DOI: 10.5334/dsj-2025-015. (Visited on 2025-12-03) (cit. on pp. 148, 149).

[Mak+21] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. "NeuroKit2: A Python Toolbox for Neurophysiological Signal Processing". In: *Behavior Research Methods* 53.4 (2021-08), pp. 1689–1696. ISSN: 1554-3528. DOI: 10.3758/s13428-020-01516-y. (Visited on 2025-11-23) (cit. on p. 78).

[Man+21] Aishwarya Mandyam, Elizabeth C. Yoo, Jeff Soules, Krzysztof Laudanski, and Barbara E. Engelhardt. "COP-E-CAT: Cleaning and Organization Pipeline for EHR Computational and Analytic Tasks". In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. Gainesville Florida: ACM, 2021-08, pp. 1–9. ISBN: 978-1-4503-8450-6. DOI: 10.1145/3459930.3469536. (Visited on 2023-03-07) (cit. on pp. 29, 31, 102).

[MW47] H. B. Mann and D. R. Whitney. "On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other". In: *The Annals of Mathematical Statistics* 18.1 (1947), pp. 50–60. ISSN: 0003-4851. JSTOR: 2236101. (Visited on 2026-01-05) (cit. on p. 61).

[MR10] Sébastien Marcel and Yann Rodriguez. "Torchvision the Machine-Vision Package of Torch". In: *Proceedings of the 18th ACM International Conference on Multimedia*. MM '10. New York, NY, USA: Association for Computing Machinery, 2010-10, pp. 1485–1488. ISBN: 978-1-60558-933-6. DOI: 10.1145/1873951.1874254. (Visited on 2025-12-03) (cit. on pp. 142, 149).

[Mau+25] Max M. Maurer, Bjarne Pfitzner, Robin P. van de Water, Lara Faraj, Christoph Riepe, Daniela Zuluaga, Felix Krenzien, Nathanael Raschzok, Robert Siegel, Christian Schineis, Bert Arnrich, Katharina Beyer, Johann Pratschke, Igor M. Sauer, and Axel Winter. "Privacy Preserving Federated Learning for 90-Day Mortality Prediction in Colorectal Surgery: A Multicenter Retrospective Development and Comparison Study". In: *International Journal of Surgery (London, England)* (2025-08). ISSN: 1743-9159. DOI: 10.1097/JS9.0000000000003084 (cit. on pp. xv, 55, 58).

[McD25] Matthew McDermott. "The (Lack of?) Science of Machine Learning for Healthcare". In: *Proceedings of the 4th Machine Learning for Health Symposium*. PMLR, 2025-02, pp. 19–29. (Visited on 2025-08-08) (cit. on p. 101).

[McD+25a] Matthew McDermott, Oufattole, Robin van de Water, Ryan King, Patrick Rockenschaub, and Simon Lee. *Mmcdermott/MEDS_transforms: 0.6.1*. Zenodo. 2025-11. DOI: 10.5281/ZENODO.17535783. (Visited on 2025-11-15) (cit. on p. 142).

[McD+25b] Matthew McDermott, Ethan Steinberg, Robin Water, Patrick Rockenschaub, Tom Pollard, EthanSteinbergPrealize, Michael Wornow, and juancaCeb. *Medical-Event-Data-Standard/Meds: 0.4.1*. Zenodo. 2025-11. DOI: 10.5281/zenodo.17535827. (Visited on 2025-12-15) (cit. on p. 100).

[MX25] Matthew McDermott and Justin Xu. *MIMIC-IV MEDS: An ETL Pipeline to Extract MIMIC-IV Data into the MEDS Format*. Zenodo. 2025-11. DOI: 10.5281/ZENODO.17535580. (Visited on 2025-11-15) (cit. on pp. 100, 142).

[McD+25c] Matthew McDermott et al. "MEDS: A Simple, Interoperable Data Standard and Ecosystem for Health AI Research". In: *Under review at NEJM AI* (2025) (cit. on pp. xvii, 99, 142).

[McD+23] Matthew B. A. McDermott, Bret Nestor, Peniel Argaw, and Isaac Kohane. *Event Stream GPT: A Data Preprocessing and Modeling Library for Generative, Pre-trained Transformers over Continuous-time Sequences of Complex Events*. 2023-06. DOI: 10.48550/arXiv.2306.11547. arXiv: 2306.11547 [cs]. (Visited on 2023-06-26) (cit. on p. 138).

[McD+21] Matthew B. A. McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. "Reproducibility in Machine Learning for Health Research: Still a Ways to Go". In: *Science Translational Medicine* 13.586 (2021-03), eabb1655. DOI: 10.1126/scitranslmed.abb1655. (Visited on 2025-07-26) (cit. on pp. 44, 148).

[McD+25d] Matthew B. A. McDermott, Justin Xu, Teya S. Bergamaschi, Hyewon Jeong, Simon A. Lee, Nassim Oufattole, Patrick Rockenschaub, Kamilė Stankevičiūtė, Ethan Steinberg, Jimeng Sun, Robin P. van de Water, Michael Wornow, John Wu, and Zhenbang Wu. "MEDS: Building Models and Tools in a Reproducible Health AI Ecosystem". In: *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*. Toronto ON Canada: ACM, 2025-08, pp. 6243–6244. ISBN: 979-8-4007-1454-2. DOI: 10.1145/3711896.3737608. (Visited on 2025-08-10) (cit. on pp. xv, 6, 99).

[McD+24] Matthew B. A. McDermott et al. "MEDS Decentralized, Extensible Validation (MEDS-DEV) Benchmark: Establishing Reproducibility and Comparability in ML for Health". In: *Machine Learning for Health (ML4H), 15-16 December 2024, Vancouver, Canada* (2024-11). (Visited on 2024-12-24) (cit. on pp. xv, 6, 9, 22, 99, 100, 106, 107, 121, 123, 125, 127, 142, 149, 151).

[McE+23] Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. "When Do Neural Nets Outperform Boosted Trees on Tabular Data?" In: *Advances in Neural Information Processing Systems* 36 (2023-12), pp. 76336–76369. (Visited on 2025-09-01) (cit. on pp. 67, 94, 136).

[MHM20] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020-09. DOI: 10.48550/arXiv.1802.03426. arXiv: 1802.03426 [stat]. (Visited on 2025-12-05) (cit. on pp. 79, 137, 291).

[McK10] Wes McKinney. "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a (cit. on p. 44).

[McL+23] Kenneth A McLean, Tanvi Goel, Samuel Lawday, Aya Riad, Joana Simoes, Stephen R Knight, Dhruva Ghosh, James C Glasbey, Aneel Bhangu, Ewen M Harrison, and NIHR Global Health Research Unit on Global Surgery. "Prognostic Models for Surgical-Site Infection in Gastrointestinal Surgery: Systematic Review". In: *British Journal of Surgery* 110.11 (2023-11), pp. 1441–1450. ISSN: 1365-2168. DOI: 10.1093/bjs/znad187. (Visited on 2024-01-31) (cit. on p. 56).

[McL+21] Kenneth A. McLean et al. "Remote Diagnosis of Surgical-Site Infection Using a Mobile Digital Intervention: A Randomised Controlled Trial in Emergency Surgery Patients". In: *npj Digital Medicine* 4.1 (2021-11), p. 160. ISSN: 2398-6352. DOI: 10.1038/s41746-021-00526-0. (Visited on 2025-08-26) (cit. on p. 58).

[Med+19] Goran Medic, Melodi Kosaner Kließ, Louis Atallah, Jochen Weichert, Saswat Panda, Maarten Postma, and Amer El-Kerdi. "Evidence-Based Clinical Decision Support Systems for the Prediction and Detection of Three Disease States in Critical Care: A Systematic Literature Review." In: *F1000Research* 8 (2019), p. 1728. ISSN: 2046-1402. DOI: 10.12688/f1000research.20498.2 (cit. on pp. 24, 25, 27, 46).

[Meh+17] Hemalkumar B. Mehta, Eric Sieloff, Sreenivas P. Veeranki, Sneha D. Sura, Taylor S. Riall, Anthony J. Senagore, and James S. Goodwin. "Risk Prediction Models for Hospital Readmission in Surgery: A Systematic Review". In: *JOURNAL OF THE AMERICAN COLLEGE OF SURGEONS* 225.4, 2 (2017-10), E113. ISSN: 1072-7515. DOI: 10.1016/j.jamcollsurg.2017.07.832 (cit. on pp. 24, 25).

[Mej+22] Elisa Mejía-Mejía, John Allen, Karthik Budidha, Chadi El-Hajj, Panicos A. Kyriacou, and Peter H. Charlton. "Photoplethysmography Signal Processing and Synthesis". In: *Photoplethysmography*. Elsevier, 2022, pp. 69–146. ISBN: 978-0-12-823374-0. DOI: 10.1016/B978-0-12-823374-0.00015-3. (Visited on 2023-12-19) (cit. on p. 73).

[Men+15] Bjoern H. Menze et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)". In: *IEEE Transactions on Medical Imaging* 34.10 (2015-10), pp. 1993–2024. ISSN: 1558-254X. DOI: 10.1109/TMI.2014.2377694. (Visited on 2025-08-01) (cit. on p. 15).

[Mer+20a] Katiuscha Merath, J. Madison Hyer, Rittal Mehta, Ayesha Farooq, Fabio Bagante, Kota Sahara, Diamantis I. Tsilimigras, Eliza Beal, Anghela Z. Paredes, Lu Wu, Aslam Ejaz, and Timothy M. Pawlik. "Use of Machine Learning for Prediction of Patient Risk of Postoperative Complications After Liver, Pancreatic, and Colorectal Surgery". In: *Journal of Gastrointestinal Surgery* 24.8 (2020-08), pp. 1843–1851. ISSN: 1091-255X, 1873-4626. DOI: 10.1007/s11605-019-04338-2. (Visited on 2022-03-09) (cit. on p. 46).

[Mer+20b] Katiuscha Merath, J. Madison Hyer, Rittal Mehta, Ayesha Farooq, Fabio Bagante, Kota Sahara, Diamantis I. Tsilimigras, Eliza Beal, Anghela Z. Paredes, Lu Wu, Aslam Ejaz, and Timothy M. Pawlik. "Use of Machine Learning for Prediction of Patient Risk of Postoperative Complications After Liver, Pancreatic, and Colorectal Surgery". In: *Journal of Gastrointestinal Surgery* 24.8 (2020-01), pp. 1843–1851. ISSN: 1091-255X. DOI: 10.1007/s11605-019-04338-2. (Visited on 2026-01-05) (cit. on p. 67).

[18] *Merative MarketScan Research Databases*. https://www.merative.com/documents/merative-marketscan-research-databases. 2018. (Visited on 2025-10-24) (cit. on pp. 15, 22, 106).

[Mer+16] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. *Pointer Sentinel Mixture Models*. https://arxiv.org/abs/1609.07843v1. 2016-09. (Visited on 2025-12-10) (cit. on pp. 139–141).

[Mey+18] Alexander Meyer, Dina Zverinski, Boris Pfahringer, Jorg Kempfert, Titus Kuehne, Simon H Sundermann, Christof Stamm, Thomas Hofmann, Volkmar Falk, and Carsten Eickhoff. "Machine Learning for Real-Time Prediction of Complications in Critical Care: A Retrospective Study." In: *The Lancet. Respiratory medicine* 6.12 (2018), pp. 905–914. ISSN: 2213-2619. DOI: 10.1016/S2213-2600(18)30300-X (cit. on pp. 30, 93).

[Mih+11] Borislava Mihaylova, Andrew Briggs, Anthony O'Hagan, and Simon G. Thompson. "Review of Statistical Methods for Analysing Healthcare Resources and Costs". In: *Health Economics* 20.8 (2011), pp. 897–916. ISSN: 1099-1050. DOI: 10.1002/hec.1653. (Visited on 2025-08-11) (cit. on p. 20).

[Møl+25] Jesper Mølgaard, Katja K. Grønbæk, Søren S. Rasmussen, Jonas P. Eiberg, Lars N. Jørgensen, Michael P. Achiam, Malene Rohrsted, Upender M. Singh, Tuyet-Hoa Hoang, Marlene Søgaard, Christian S. Meyhoff, and Eske K. Aasvang. "Continuous Vital Sign Monitoring at the Surgical Ward for Improved Outcomes After Major Noncardiac Surgery: A Randomized Clinical Trial". In: *Anesthesia and Analgesia* 141.4 (2025-10), pp. 807–817. ISSN: 0003-2999. DOI: 10.1213/ANE.0000000000007606. (Visited on 2025-11-05) (cit. on p. 93).

[Moo+22] Benjamin Moody, Sicheng Hao, Brian Gow, Tom Pollard, Wei Zong, and Roger Mark. *MIMIC-IV Waveform Database*. 2022-08. DOI: 10.13026/A2MW-F949. (Visited on 2026-01-17) (cit. on p. 15).

[MM96] G.B. Moody and R.G. Mark. "A Database to Support Development and Evaluation of Intelligent Intensive Care Monitoring". In: *Computers in Cardiology 1996*. 1996-09, pp. 657–660. DOI: 10.1109/CIC.1996.542622. (Visited on 2025-07-21) (cit. on p. 13).

[MM01] G.B. Moody and R.G. Mark. "The Impact of the MIT-BIH Arrhythmia Database". In: *IEEE Engineering in Medicine and Biology Magazine* 20.3 (2001-05), pp. 45–50. ISSN: 1937-4186. DOI: 10.1109/51.932724. (Visited on 2026-01-17) (cit. on p. 15).

[Moo+21a] Michael Moor, Nicolas Bennet, Drago Plecko, Max Horn, Bastian Rieck, Nicolai Meinshausen, Peter Bühlmann, and Karsten Borgwardt. *Predicting Sepsis in Multi-Site, Multi-National Intensive Care Cohorts Using Deep Learning*. 2021-07. arXiv: 2107.05230 [cs]. (Visited on 2022-08-11) (cit. on pp. 46, 50, 51, 196, 197).

[Moo+19] Michael Moor, Max Horn, Bastian Rieck, Damian Roqueiro, and Karsten Borgwardt. "Early Recognition of Sepsis with Gaussian Process Temporal Convolutional Networks and Dynamic Time Warping". In: *Proceedings of the 4th Machine Learning for Healthcare Conference*. PMLR, 2019-10, pp. 2–26. (Visited on 2023-04-18) (cit. on p. 46).

[Moo+21b] Michael Moor, Bastian Rieck, Max Horn, Catherine R. Jutzeler, and Karsten Borgwardt. "Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review". In: *Frontiers in Medicine* 8 (2021). ISSN: 2296-858X. (Visited on 2022-01-27) (cit. on pp. 24, 25, 27, 46, 48).

[Mor21] Amy Morgan. "Long-Term Outcomes from Critical Care". In: *Surgery (Oxford)* 39.1 (2021-01), pp. 53–57. ISSN: 0263-9319. DOI: 10.1016/j.mpsur.2020.11.005. (Visited on 2024-02-16) (cit. on p. 57).

[Mou+24] Dana Moukheiber et al. *Northwestern ICU (NWICU) Database*. 2024. DOI: 10.13026/S84W-1829. (Visited on 2025-01-31) (cit. on pp. 9, 16, 101, 104, 105, 109, 116).

[Mua+21] Noor M. Mualla, Maryam R. Hussain, Muhammad Akrmah, Preeti Malik, Sadia Bashir, and Jenny J. Lin. "The Impact of Postoperative Complications on Long-Term Oncological Outcomes Following Curative Resection of Colorectal Cancer (Stage I-III): A Systematic Review and Meta-Analysis". In: *Cureus* 13.1 (2021-01), e12837. ISSN: 2168-8184. DOI: 10.7759/cureus.12837 (cit. on p. 56).

[Mur+21] Sankavi Muralitharan, Walter Nelson, Shuang Di, Michael McGillion, Pj Devereaux, Neil Grant Barr, and Jeremy Petch. "Machine Learning–Based Early Warning Systems for Clinical Deterioration: Systematic Scoping Review". In: *Journal of Medical Internet Research* 23.2 (2021-02), e25187. ISSN: 1438-8871. DOI: 10.2196/25187. (Visited on 2022-03-22) (cit. on pp. 25, 26, 46, 57).

[Mur+09] Shawn Murphy, Susanne Churchill, Lynn Bry, Henry Chueh, Scott Weiss, Ross Lazarus, Qing Zeng, Anil Dubey, Vivian Gainer, Michael Mendis, John Glaser, and Isaac Kohane. "Instrumenting the Health Care Enterprise for Discovery Research in the Genomic Era". In: *Genome Research* 19.9 (2009-01), pp. 1675–1681. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.094615.109. (Visited on 2025-07-21) (cit. on p. 13).

[Mur+10] Shawn N Murphy, Griffin Weber, Michael Mendis, Vivian Gainer, Henry C Chueh, Susanne Churchill, and Isaac Kohane. "Serving the Enterprise and beyond with Informatics for Integrating Biology and the Bedside (I2b2)". In: *Journal of the American Medical Informatics Association* 17.2 (2010-03), pp. 124–130. ISSN: 1067-5027. DOI: 10.1136/jamia.2009.000893. (Visited on 2025-10-11) (cit. on p. 101).

[NB03] Claude Nadeau and Yoshua Bengio. "Inference for the Generalization Error". In: *Machine Learning* 52.3 (2003-09), pp. 239–281. ISSN: 1573-0565. DOI: 10.1023/A:1024068626366. (Visited on 2026-01-21) (cit. on pp. 87, 240).

[Nah+18] Christopher B Nahm, Saxon J Connor, Jaswinder S Samra, and Anubhav Mittal. "Postoperative Pancreatic Fistula: A Review of Traditional and Emerging Concepts". In: *Clinical and Experimental Gastroenterology* 11 (2018-03), pp. 105–118. ISSN: 1178-7023. DOI: 10.2147/CEG.S120217. (Visited on 2024-12-17) (cit. on pp. 56, 71).

[Nel+11] Stuart J Nelson, Kelly Zeng, John Kilbourne, Tammy Powell, and Robin Moore. "Normalized Names for Clinical Drugs: RxNorm at 6 Years". In: *Journal of the American Medical Informatics Association* 18.4 (2011), pp. 441–448 (cit. on p. 103).

[Nem+18] Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D. Stanley, Gari D. Clifford, and Timothy G. Buchman. "An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU". In: *Critical Care Medicine* 46.4 (2018-04), pp. 547–553. ISSN: 1530-0293. DOI: 10.1097/CCM.0000000000002936 (cit. on pp. 56, 93).

[Nep+25] Dmitri Nepogodiev et al. "Surgical Health Policy 2025–35: Strengthening Essential Services for Tomorrow's Needs". In: *The Lancet* 406.10505 (2025-08), pp. 860–880. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(25)00985-7. (Visited on 2025-08-26) (cit. on p. 1).

[Ngu+25] Van Nguyen, Brendan MacDonald, Anthony Cignarella, and Charne Miller. "A Descriptive Investigation of Alarm Activation in a Critical Care Setting". In: *Nursing in Critical Care* 30.2 (2025-03), e13302. ISSN: 1362-1017. DOI: 10.1111/nicc.13302. (Visited on 2025-11-19) (cit. on p. 91).

[Nic+22] Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. *TorchMetrics - Measuring Reproducibility in PyTorch*. 2022-02. DOI: 10.21105/joss.04101 (cit. on pp. 42, 209, 210).

[Nik+22] Okke Nikkinen, Timo Kolehmainen, Toni Aaltonen, Elias Jämsä, Seppo Alahuhta, and Merja Vakkala. "Developing a Supervised Machine Learning Model for Predicting Perioperative Acute Kidney Injury in Arthroplasty Patients". In: *Computers in Biology and Medicine* 144 (2022-05), p. 105351. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2022.105351. (Visited on 2022-04-04) (cit. on p. 46).

[Nun+09] Joseph William Nunoo-Mensah, Michael Rosen, Linday S. Chan, Nir Wasserberg, and Robert W. Beart. "Prevalence of Intra-Abdominal Surgery: What Is an Individual's Lifetime Risk?" In: *Southern Medical Journal* 102.1 (2009-01), pp. 25–29. ISSN: 1541-8243. DOI: 10.1097/SMJ.0b013e318182575b (cit. on pp. 55, 68).

[Obe+19] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations". In: *Science* 366.6464 (2019-10), pp. 447–453. DOI: 10.1126/science.aax2342. (Visited on 2025-11-30) (cit. on p. 148).

[Odg+24] Mikkel Odgaard, Kiril Vadimovic Klein, Sanne Møller Thysen, Espen Jimenez-Solem, Martin Sillesen, and Mads Nielsen. *CORE-BEHRT: A Carefully Optimized and Rigorously Evaluated BEHRT*. 2024-10. DOI: 10.48550/arXiv.2404.15201. arXiv: 2404.15201 [cs]. (Visited on 2025-08-13) (cit. on pp. 22, 106, 107).

[oMed00] National Library of Medicine (U.S.) *Medical Subject Headings*. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Library of Medicine, 2000 (cit. on p. 103).

BIBLIOGRAPHY

[Oli+23] Matthieu Oliver, Jérôme Allyn, Rémi Carencotte, Nicolas Allou, and Cyril Ferdynus. "Introducing the BlendedICU Dataset, the First Harmonized, International Intensive Care Dataset". In: *Journal of Biomedical Informatics* 146 (2023-10), p. 104502. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2023.104502. (Visited on 2023-11-13) (cit. on pp. 29, 31).

[Org92] World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. Vol. 1. World Health Organization, 1992. (Visited on 2025-10-28) (cit. on pp. 102, 118).

[Ouf+24a] Nassim Oufattole, Teya Bergamaschi, Aleksia Kolo, Hyewon Jeong, Hanna Gaggin, Collin M. Stultz, and Matthew B. A. McDermott. *MEDS-Tab: Automated Tabularization and Baseline Methods for MEDS Datasets*. 2024-10. DOI: 10.48550/arXiv.2411.00200. arXiv: 2411.00200 [cs]. (Visited on 2025-09-16) (cit. on pp. 121, 125, 142).

[Ouf+24b] Nassim Oufattole, Hyewon Jeong, Matthew B. A. McDermott, Aparna Balagopalan, Bryan Jangeesingh, Marzyeh Ghassemi, and Collin Stultz. "Event-Based Contrastive Learning for Medical Time Series". In: *Proceedings of the 9th Machine Learning for Healthcare Conference*. PMLR, 2024-11. (Visited on 2025-08-13) (cit. on pp. 22, 106, 107).

[Pag+24] Bethan Page, Dulcie Irving, Rene Amalberti, and Charles Vincent. "Health Services under Pressure: A Scoping Review and Development of a Taxonomy of Adaptive Strategies". In: *BMJ Quality & Safety* 33.11 (2024-11), pp. 738–747. ISSN: 2044-5415, 2044-5423. DOI: 10.1136/bmjqs-2023-016686. (Visited on 2025-11-18) (cit. on p. 97).

[Pan+19] Ziyuan Pan, Hao Du, Kee Yuan Ngiam, Fei Wang, Ping Shum, and Mengling Feng. "A Self-Correcting Deep Learning Approach to Predict Acute Conditions in Critical Care". In: *arXiv:1901.04364 [cs, stat]* (2019-01). arXiv: 1901.04364 [cs, stat]. (Visited on 2022-03-09) (cit. on p. 46).

[Pan+25a] Chao Pang, Vincent Jeanselme, Young Sang Choi, Xinzhuo Jiang, Zilin Jing, Aparajita Kashyap, Yuta Kobayashi, Yanwei Li, Florent Pollet, Karthik Natarajan, and Shalmali Joshi. *FoMoH: A Clinically Meaningful Foundation Model Evaluation for Structured Electronic Health Records*. 2025-06. DOI: 10.48550/arXiv.2505.16941. arXiv: 2505.16941 [cs]. (Visited on 2025-06-25) (cit. on pp. 21, 22, 100, 105–107, 110, 111, 117, 123, 127–129, 134, 151).

[Pan+21] Chao Pang, Xinzhuo Jiang, Krishna S. Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. *CEHR-BERT: Incorporating Temporal Information from Structured EHR Data to Improve Prediction Tasks*. 2021-11. DOI: 10.48550/arXiv.2111.08585. arXiv: 2111.08585 [cs]. (Visited on 2025-08-13) (cit. on pp. 22, 105–107, 121, 125, 129, 132).

[Pan+24] Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S. Kalluri, Elise L. Minto, Jason Patterson, Linying Zhang, George Hripcsak, Gamze Gürsoy, Noémie Elhadad, and Karthik Natarajan. *CEHR-GPT: Generating Electronic Health Records with Chronological Patient Timelines*. 2024-05. DOI: 10.48550/arXiv.2402.04400. arXiv: 2402.04400 [cs]. (Visited on 2025-08-13) (cit. on pp. 22, 105–107, 127).

[Pan+25b] Chao Pang, Jiheum Park, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S. Kalluri, Shalmali Joshi, Noémie Elhadad, and Karthik Natarajan. *CEHR-XGPT: A Scalable Multi-Task Foundation Model for Electronic Health Records*. 2025-09. DOI: 10.48550/arXiv.2509.03643. arXiv: 2509.03643 [cs]. (Visited on 2025-09-08) (cit. on pp. 22, 106, 127, 142).

[Pas+19] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. DOI: 10.48550/arXiv.1912.01703. (Visited on 2023-04-18) (cit. on p. 44).

[Pat+20] Colleen M. Pater, Tina K. Sosa, Jacquelyn Boyer, Rhonda Cable, Melinda Egan, Timothy K. Knilans, Amanda C. Schondelmeyer, Christine L. Schuler, and Nicolas L. Madsen. "Time Series Evaluation of Improvement Interventions to Reduce Alarm Notifications in a Paediatric Hospital". In: *BMJ Quality & Safety* 29.9 (2020-09), pp. 717–726. ISSN: 2044-5415, 2044-5423. DOI: 10.1136/bmjqs-2019-010368. (Visited on 2025-11-20) (cit. on p. 91).

[PB18] Emily F. Patridge and Tania P. Bardyn. "Research Electronic Data Capture (REDCap)". In: *Journal of the Medical Library Association : JMLA* 106.1 (2018-01), pp. 142–144. ISSN: 1536-5050. DOI: 10.5195/jmla.2018.319. (Visited on 2023-11-08) (cit. on p. 73).

[Pea+06] Rupert M. Pearse, David A. Harrison, Philip James, David Watson, Charles Hinds, Andrew Rhodes, R. Michael Grounds, and E. David Bennett. "Identification and Characterisation of the High-Risk Surgical Population in the United Kingdom". In: *Critical Care* 10.3 (2006-06), R81. ISSN: 1364-8535. DOI: 10.1186/cc4928. (Visited on 2026-01-05) (cit. on p. 68).

[Pea00] Karl Pearson. "X. *On the Criterion That a given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling*". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (1900-07), pp. 157–175. ISSN: 1941-5982, 1941-5990. DOI: 10.1080/14786440009463897. (Visited on 2026-01-05) (cit. on p. 61).

[Ped+11a] F. Pedregosa et al. “Scikit-Learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on pp. 42, 47, 209, 211).

[Ped+11b] Fabian Pedregosa et al. “Scikit-Learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. ISSN: 1533-7928. (Visited on 2022-12-21) (cit. on pp. 43, 44, 85).

[Per+22] Manuel Pera et al. “Machine Learning Risk Prediction Model of 90-Day Mortality After Gastrectomy for Cancer”. In: *Annals of Surgery* 276.5 (2022-11), p. 776. ISSN: 0003-4932. DOI: 10.1097/SLA.0000000000005616. (Visited on 2026-01-05) (cit. on p. 67).

[Per+15] Adler Perotte, Rajesh Ranganath, Jamie S Hirsch, David Blei, and Noémie Elhadad. “Risk Prediction for Chronic Kidney Disease Progression Using Heterogeneous Electronic Health Record Data and Time Series Analysis”. In: *Journal of the American Medical Informatics Association : JAMIA* 22.4 (2015-07), pp. 872–880. ISSN: 1067-5027. DOI: 10.1093/jamia/ocv024. (Visited on 2023-06-05) (cit. on p. 46).

[Per+23] Bharath Perugu, Varun Wadhwa, Jin Kim, Jenny Cai, Auey Shin, and Amar Gupta. “Surmounting Barriers to Healthcare Data and Information: Cases in Point, the U.S. Experience”. In: *Telehealth and Medicine Today* 8.4 (2023). ISSN: 24716960. DOI: 10.30953/thmt.v8.425. (Visited on 2026-01-19) (cit. on p. 101).

[Pfi+24] Bjarne Pfitzner, Max M. Maurer, Axel Winter, Christoph Riepe, Igor M. Sauer, Robin van de Water, and Bert Arnrich. “Differentially-Private Federated Learning with Non-IID Data for Surgical Risk Prediction”. In: *2024 IEEE First International Conference on Artificial Intelligence for Medicine, Health and Care (AIMHC)*. IEEE, 2024, pp. 120–129. (Visited on 2024-10-23) (cit. on pp. xvi, 10, 24, 55, 58).

[Pfi+25] Bjarne Pfitzner, Max M. Maurer, Axel Winter, Christoph Riepe, Igor M. Sauer, Robin van de Water, Christian Denecke, Johann Pratschke, and Bert Arnrich. “Differentially-Private Federated Learning with Non-IID Data For Surgical Risk Prediction.” In: *International Journal of Semantic Computing* 19.3 (2025). (Visited on 2025-11-23) (cit. on pp. xiv, 10, 55, 58, 68, 117).

[Pia14] Steven T. Piantadosi. “Zipf’s Word Frequency Law in Natural Language: A Critical Review and Future Directions”. In: *Psychonomic bulletin & review* 21.5 (2014-10), pp. 1112–1130. ISSN: 1069-9384. DOI: 10.3758/s13423-014-0585-6. (Visited on 2025-12-11) (cit. on p. 138).

[Pie+95] M. Piepoli, C. S. Garrard, D. A. Kontoyannis, and L. Bernardi. “Autonomic Control of the Heart and Peripheral Vessels in Human Septic Shock”. In: *Intensive Care Medicine* 21.2 (1995-02), pp. 112–119. ISSN: 0342-4642. DOI: 10.1007/BF01726532 (cit. on p. 94).

[Pin+21] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d’Alche-Buc, Emily Fox, and Hugo Larochelle. “Improving Reproducibility in Machine Learning Research(A Report from the NeurIPS 2019 Reproducibility Program)”. In: *Journal of Machine Learning Research* 22.164 (2021), pp. 1–20. ISSN: 1533-7928. (Visited on 2025-12-03) (cit. on pp. 148, 149).

[Pir+15] Romain Pirracchio, Maya L Petersen, Marco Carone, Matthieu Resche Rigon, Sylvie Chevret, and Mark J van der Laan. “Mortality Prediction in Intensive Care Units with the Super ICU Learner Algorithm (SICULA): A Population-Based Study”. In: *The Lancet Respiratory Medicine* 3.1 (2015-01), pp. 42–52. ISSN: 22132600. DOI: 10.1016/S2213-2600(14)70239-5 (cit. on p. 47).

[Pla+22] Deborah Plana, Dennis L. Shung, Alyssa A. Grimshaw, Anurag Saraf, Joseph J. Y. Sung, and Benjamin H. Kann. “Randomized Clinical Trials of Machine Learning Interventions in Health Care: A Systematic Review”. In: *JAMA Network Open* 5.9 (2022-09), e2233946. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2022.33946. (Visited on 2025-07-26) (cit. on p. 15).

[Plo24a] Plotly. *Dash*. Zenodo. 2024-11. DOI: 10.5281/ZENODO.14182630. (Visited on 2025-10-27) (cit. on p. 113).

[Plo24b] Plotly. *Plotly PY*. Zenodo. 2024-12. DOI: 10.5281/ZENODO.14503524. (Visited on 2025-10-27) (cit. on p. 113).

[Pol+18] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. “The eICU Collaborative Research Database, a Freely Available Multi-Center Database for Critical Care Research”. In: *Scientific Data* 5.1 (2018-12), p. 180178. ISSN: 2052-4463. DOI: 10.1038/sdata.2018.178. (Visited on 2022-03-09) (cit. on pp. 9, 15, 16, 33, 34, 104, 105).

[Pop73] Harry E. Pople. “On the Mechanization of Abductive Logic”. In: *Proceedings of the 3rd International Joint Conference on Artificial Intelligence*. IJCAI’73. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1973-08, pp. 147–152. (Visited on 2025-08-15) (cit. on p. 13).

[Por+24] Sebastian Porsdam Mann, Anuraag A. Vazirani, Mateo Aboy, Brian D. Earp, Timo Minssen, I. Glenn Cohen, and Julian Savulescu. “Guidelines for Ethical Use and Acknowledgement of Large Language Models in Academic Writing”. In: *Nature Machine Intelligence* 6.11 (2024-11), pp. 1272–1274. ISSN: 2522-5839. DOI: 10.1038/s42256-024-00922-7. (Visited on 2025-12-17) (cit. on p. 184).

[Pos+24] Linda M. Posthuma, Martine J. M. Breteler, Philipp B. Lirk, Els J. Nieveen Van Dijkum, Maarten J. Visscher, Jennifer S. Breel, Carin A. G. L. Wensing, Jimmy Schenk, Lyan B. Vlaskamp, Mathilde C. Van Rossum, Jelle P. Ruurda, Marcel G. W. Dijkgraaf, Markus W. Hollmann, Cor J. Kalkman, and Benedikt Preckel. "Surveillance of High-Risk Early Postsurgical Patients for Real-Time Detection of Complications Using Wireless Monitoring (SHEPHERD Study): Results of a Randomized Multicenter Stepped Wedge Cluster Trial". In: *Frontiers in Medicine* 10 (2024-01), p. 1295499. ISSN: 2296-858X. DOI: 10.3389/fmed.2023.1295499. (Visited on 2024-12-17) (cit. on pp. 58, 59, 93).

[Pra+22] Nisha Pradhan, Adam R. Dyas, Michael R. Bronsert, Anne Lambert-Kerzner, William G. Henderson, Howe Qiu, Kathryn L. Colborn, Nicholas J. Mason, and Robert A. Meguid. "Attitudes about Use of Preoperative Risk Assessment Tools: A Survey of Surgeons and Surgical Residents in an Academic Health System". In: *Patient Safety in Surgery* 16.1 (2022-03), p. 13. ISSN: 1754-9493. DOI: 10.1186/s13037-022-00320-1. (Visited on 2026-01-05) (cit. on p. 68).

[Pro+19] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. *CatBoost: Unbiased Boosting with Categorical Features*. 2019-01. DOI: 10.48550/arXiv.1706.09516. arXiv: 1706.09516 [cs]. (Visited on 2025-08-14) (cit. on pp. 43, 82, 85, 94).

[Pur+18] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. "Benchmarking Deep Learning Models on Large Healthcare Datasets". In: *Journal of Biomedical Informatics* 83 (2018-07), pp. 112–134. ISSN: 15320464. DOI: 10.1016/j.jbi.2018.04.007. (Visited on 2022-03-09) (cit. on pp. 29, 31).

[Qua+11] Hude Quan, Bing Li, Chantal M. Couris, Kiyohide Fushimi, Patrick Graham, Phil Hider, Jean-Marie Januel, and Vijaya Sundararajan. "Updating and Validating the Charlson Comorbidity Index and Score for Risk Adjustment in Hospital Discharge Abstracts Using Data From 6 Countries". In: *American Journal of Epidemiology* 173.6 (2011-03), pp. 676–682. ISSN: 0002-9262. DOI: 10.1093/aje/kwq433. (Visited on 2026-01-04) (cit. on p. 61).

[Qua+05] Hude Quan, Vijaya Sundararajan, Patricia Halfon, Andrew Fong, Bernard Burnand, Jean-Christophe Luthi, L. Duncan Saunders, Cynthia A. Beck, Thomas E. Feasby, and William A. Ghali. "Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data". In: *Medical Care* 43.11 (2005-11), p. 1130. ISSN: 0025-7079. DOI: 10.1097/01.mlr.0000182534.19832.83. (Visited on 2026-01-05) (cit. on p. 61).

[Rad+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*. 2021-02. DOI: 10.48550/arXiv.2103.00020. arXiv: 2103.00020 [cs]. (Visited on 2025-10-22) (cit. on p. 22).

[Raf19] Edward Raff. "A Step toward Quantifying Independently Reproducible Machine Learning Research". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019-12, pp. 5485–5495. (Visited on 2026-01-02) (cit. on p. 44).

[RTK24] Oona Rainio, Jarmo Teuho, and Riku Klén. "Evaluation Metrics and Statistical Tests for Machine Learning". In: *Scientific Reports* 14.1 (2024-03), p. 6086. ISSN: 2045-2322. DOI: 10.1038/s41598-024-56706-x. (Visited on 2026-01-21) (cit. on p. 241).

[Raj+18] Alvin Rajkomar et al. "Scalable and Accurate Deep Learning with Electronic Health Records". In: *npj Digital Medicine* 1.1 (2018-05), p. 18. ISSN: 2398-6352. DOI: 10.1038/s41746-018-0029-1. (Visited on 2025-10-11) (cit. on p. 105).

[Raj+20] Hema Sekhar Reddy Rajula, Giuseppe Verlato, Mirko Manchia, Nadia Antonucci, and Vassilios Fanos. "Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment". In: *Medicina* 56.9 (2020-09), p. 455. ISSN: 1648-9144. DOI: 10.3390/medicina56090455. (Visited on 2025-08-11) (cit. on p. 20).

[Ran+20] Nina Rank, Boris Pfahringer, Jörg Kempfert, Christof Stamm, Titus Kühne, Felix Schoenrath, Volkmar Falk, Carsten Eickhoff, and Alexander Meyer. "Deep-Learning-Based Real-Time Prediction of Acute Kidney Injury Outperforms Human Predictive Performance". In: *npj Digital Medicine* 3.1 (2020-12), p. 139. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00346-8. (Visited on 2022-03-17) (cit. on p. 46).

[Ras+21a] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. "Med-BERT: Pretrained Contextualized Embeddings on Large-Scale Structured Electronic Health Records for Disease Prediction". In: *npj Digital Medicine* 4.1 (2021-05), p. 86. ISSN: 2398-6352. DOI: 10.1038/s41746-021-00455-y. (Visited on 2025-08-13) (cit. on pp. 22, 106, 107).

[Ras+21b] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. "Med-BERT: Pretrained Contextualized Embeddings on Large-Scale Structured Electronic Health Records for Disease Prediction". In: *npj Digital Medicine* 4.1 (2021-05), p. 86. ISSN: 2398-6352. DOI: 10.1038/s41746-021-00455-y. (Visited on 2025-10-12) (cit. on p. 143).

[RNC94] A. L. Rector, W. A. Nowlan, and Galen Consortium. "The Galen Project". In: *Computer Methods and Programs in Biomedicine*. AIM - Advanced Informatics in Medicine 45.1 (1994-10), pp. 75–78. ISSN: 0169-2607. DOI: 10.1016/0169-2607(94)90020-5. (Visited on 2025-07-21) (cit. on p. 13).

[24b] *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA Relevance)*. 2024-06. (Visited on 2025-12-03) (cit. on p. 150).

[Rei+24] Christian Reich, Anna Ostropolets, Patrick Ryan, Peter Rijnbeek, Martijn Schuemie, Alexander Davydov, Dmitry Dymshyts, and George Hripcsak. "OHDSI Standardized Vocabularies—a Large-Scale Centralized Reference Ontology for International Data Harmonization". In: *Journal of the American Medical Informatics Association* 31.3 (2024-03), pp. 583–590. ISSN: 1527-974X. DOI: 10.1093/jamia/ocad247. (Visited on 2025-04-11) (cit. on pp. 112, 113, 127).

[Ren+24] Pawel Renc, Yugang Jia, Anthony E. Samir, Jaroslaw Was, Quanzheng Li, David W. Bates, and Arkadiusz Sitek. "Zero Shot Health Trajectory Prediction Using Transformer". In: *npj Digital Medicine* 7.1 (2024-09), p. 256. ISSN: 2398-6352. DOI: 10.1038/s41746-024-01235-0. (Visited on 2025-08-13) (cit. on pp. 22, 106, 107, 142).

[Res+21] Benjamin J. Resio, Lou Gonsalves, Maureen Canavan, Lloyd Mueller, Cathryn Phillips, Sathe Tejas, Katrina Swett, and Daniel J. Boffa. "Where the Other Half Dies: Analysis of Mortalities Occurring More than 30 Days after Complex Cancer Surgery". In: *Annals of Surgical Oncology* 28.3 (2021), pp. 1278–1286. (Visited on 2026-01-05) (cit. on p. 67).

[Reu+24] Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. "BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices". In: *Advances in Neural Information Processing Systems* 37 (2024-12), pp. 21763–21813. DOI: 10.52202/079017-0685. (Visited on 2025-12-03) (cit. on p. 149).

[Rey+19] Matthew A Reyna, Chris Josef, Salman Seyedi, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Ashish Sharma, Shamim Nemati, and Gari D Clifford. "Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019". In: *2019 Computing in Cardiology (CinC)*. 2019-09, Page 1–Page 4. DOI: 10.23919/CinC49843.2019.9005736 (cit. on pp. 46, 50, 196).

[Rie+25a] Christoph Riepe, Robin van de Water, Axel Winter, Bjarne Pfitzner, Lara Faraj, Robert Ahlborn, Maximilian Schulze, Daniela Zuluaga, Christian Schineis, Katharina Beyer, Johann Pratschke, Bert Arnrich, Igor M. Sauer, and Max M. Maurer. "90-Day Mortality Prediction in Elective Visceral Surgery Using Machine Learning: A Retrospective Multicenter Development, Validation and Comparison Study". In: *International Journal of Surgery (London, England)* (2025-04). ISSN: 1743-9159. DOI: 10.1097/JS9.0000000000002372 (cit. on p. 8).

[Rie+25b] Christoph Riepe, Robin van de Water, Axel Winter, Bjarne Pfitzner, Lara Faraj, Robert Ahlborn, Maximilian Schulze, Daniela Zuluaga, Christian Schineis, Katharina Beyer, Johann Pratschke, Bert Arnrich, Igor M. Sauer, and Max M. Maurer. "90-Day Mortality Prediction in Elective Visceral Surgery Using Machine Learning: A Retrospective Multicenter Development, Validation, and Comparison Study". In: *International Journal of Surgery* 111.6 (2025-06), p. 3742. DOI: 10.1097/JS9.0000000000002372. (Visited on 2025-08-15) (cit. on pp. xiv, 5, 55, 58–61, 64–66, 228–230, 232).

[Rit+24] Álvaro Ritoré, Andreea M Oprescu, Alberto Estirado Bronchalo, and Miguel Ángel Armengol de la Hoz. *COVID Data for Shared Learning (CDSL): A Comprehensive, Multimodal COVID-19 Dataset from HM Hospitales*. 2024-10. DOI: 10.13026/1176-6C44. (Visited on 2025-04-23) (cit. on p. 16).

[Roc+25] Patrick Rockenschaub, Ela Marie Akay, Benjamin Gregory Carlisle, Adam Hilbert, Joshua Wendland, Falk Meyer-Eschenbach, Anatol-Fiete Näher, Dietmar Frey, and Vince Istvan Madai. "External Validation of AI-based Scoring Systems in the ICU: A Systematic Review and Meta-Analysis". In: *BMC Medical Informatics and Decision Making* 25.1 (2025-01), p. 5. ISSN: 1472-6947. DOI: 10.1186/s12911-024-02830-7. (Visited on 2025-07-30) (cit. on p. 24).

[Roc+24] Patrick Rockenschaub, Adam Hilbert, Tabea Kossen, Paul Elbers, Falk von Dincklage, Vince Istvan Madai, and Dietmar Frey. "The Impact of Multi-Institution Datasets on the Generalizability of Machine Learning Prediction Models in the ICU". In: *Critical Care Medicine* 52.11 (2024-11), p. 1710. ISSN: 1530-0293. DOI: 10.1097/CCM.0000000000006359. (Visited on 2025-04-24) (cit. on p. 54).

[Rod+23a] Niklas Rodemund, Bernhard Wernly, Christian Jung, Crispiana Cozowicz, and Andreas Koköfer. "The Salzburg Intensive Care Database (SICdb): An Openly Available Critical Care Dataset". In: *Intensive Care Medicine* (2023-04). ISSN: 1432-1238. DOI: 10.1007/s00134-023-07046-3. (Visited on 2023-05-30) (cit. on pp. 9, 16, 33, 57, 104, 105, 202).

[Rod+23b] Niklas Rodemund, Bernhard Wernly, Christian Jung, Crispiana Cozowicz, and Andreas Koköfer. "The Salzburg Intensive Care Database (SICdb): An Openly Available Critical Care Dataset". In: *Intensive Care Medicine* 49.6 (2023-06), pp. 700–702. ISSN: 1432-1238. DOI: 10.1007/s00134-023-07046-3. (Visited on 2024-02-02) (cit. on pp. 109, 119, 144).

[Roj+24] Juan C. Rojas et al. "A Common Longitudinal Intensive Care Unit Data Format (CLIF) to Enable Multi-Institutional Federated Critical Illness Research". In: *medRxiv: The Preprint Server for Health Sciences* (2024-09), p. 2024.09.04.24313058. DOI: 10.1101/2024.09.04.24313058 (cit. on p. 102).

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4. DOI: 10.1007/978-3-319-24574-4_28 (cit. on p. 207).

[RKK24] Julia Stefanie Roppelt, Dominik K. Kanbach, and Sascha Kraus. "Artificial Intelligence in Healthcare Institutions: A Systematic Literature Review on Influencing Factors". In: *Technology in Society* 76 (2024-03), p. 102443. ISSN: 0160-791X. DOI: 10.1016/j.techsoc.2023.102443. (Visited on 2024-12-20) (cit. on p. 56).

[Rot+23] Ylenia Rotalinti, Allan Tucker, Michael Lonergan, Puja Myles, and Richard Branson. "Detecting Drift in Healthcare AI Models Based on Data Availability". In: *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Ed. by Irena Koprinska et al. Cham: Springer Nature Switzerland, 2023, pp. 243–258. ISBN: 978-3-031-23633-4. DOI: 10.1007/978-3-031-23633-4_17 (cit. on pp. 24, 96).

[Rud19] Cynthia Rudin. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". In: *Nature Machine Intelligence* 1.5 (2019-05), pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x. (Visited on 2022-06-07) (cit. on p. 21).

[SR15] Takaya Saito and Marc Rehmsmeier. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets". In: *PLOS ONE* 10.3 (2015-03), e0118432. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0118432. (Visited on 2026-01-05) (cit. on p. 67).

[San+25a] Ricardo Santos, André V. Carreiro, Xi Peng, Hugo Gamboa, and Holger Fröhlich. *Improving Representation Learning of Complex Critical Care Data with ICU-BERT*. 2025-02. DOI: 10.48550/arXiv.2502.19593. arXiv: 2502.19593 [cs]. (Visited on 2025-04-24) (cit. on p. 54).

[San+25b] Raul Santos-Rodriguez, Kacper Sokol, Julia E. Vogt, and Sven Wellmann. "From Research to Certification with Data-Driven Medical Decision Support Systems (Dagstuhl Seminar 25052)". In: *Dagstuhl Reports* 15.1 (2025). Ed. by Raul Santos-Rodriguez, Kacper Sokol, Julia E. Vogt, and Sven Wellmann, pp. 201–220. ISSN: 2192-5283. DOI: 10.4230/DagRep.15.1.201. (Visited on 2025-11-26) (cit. on pp. xviii, 11).

[Sar+14] Massimo Sartelli et al. "Complicated Intra-Abdominal Infections Worldwide: The Definitive Data of the CIAOW Study". In: *World journal of emergency surgery: WJES* 9 (2014), p. 37. ISSN: 1749-7922. DOI: 10.1186/1749-7922-9-37 (cit. on p. 56).

[Sar+23] Tabinda Sarwar, Sattar Seifollahi, Jeffrey Chan, Xiuzhen Zhang, Vural Aksakalli, Irene Hudson, Karin Verspoor, and Lawrence Cavedon. "The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges". In: *ACM Computing Surveys* 55.2 (2023-03), pp. 1–40. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3490234. (Visited on 2022-05-02) (cit. on p. 27).

[Sau+22a] Christopher M Sauer, Li-Ching Chen, Stephanie L Hyland, Armand Girbes, Paul Elbers, and Leo A Celi. "Leveraging Electronic Health Records for Data Science: Common Pitfalls and How to Avoid Them". In: *The Lancet Digital Health* 4.12 (2022-12), e893–e898. ISSN: 25897500. DOI: 10.1016/S2589-7500(22)00154-6. (Visited on 2022-12-02) (cit. on p. 28).

[Sau+22b] Christopher M. Sauer, Tariq A. Dam, Leo A. Celi, Martin Faltys, Miguel A. A. de la Hoz, Lasith Adhikari, Kirsten A. Ziesemer, Armand Girbes, Patrick J. Thoral, and Paul Elbers. "Systematic Review and Comparison of Publicly Available ICU Data Sets—A Decision Guide for Clinicians and Data Scientists". In: *Critical Care Medicine* 50.6 (2022-06), e581–e588. ISSN: 0090-3493. DOI: 10.1097/CCM.0000000000005517. (Visited on 2022-08-08) (cit. on pp. 15, 33).

[Sv23] Evgeny S. Saveliev and Mihaela van der Schaar. *TemporAI: Facilitating Machine Learning Innovation in Time Domain Tasks for Medicine*. 2023-01. arXiv: 2301.12260 [cs]. (Visited on 2023-11-13) (cit. on pp. 29, 31, 32).

[Sch+99] Gunther Schadow, Clement J McDonald, Jeffrey G Suico, Ulrich Föhring, and Thomas Tolxdorff. "Units of Measure in Clinical Information Systems". In: *Journal of the American Medical Informatics Association* 6.2 (1999), pp. 151–162 (cit. on p. 103).

[Sch+25a] Rylan Schaeffer, Joshua Kazdan, Yegor Denisov-Blanch, Brando Miranda, Matthias Gerstgrasser, Susan Zhang, Andreas Haupt, Isha Gupta, Elyas Obbad, Jesse Dodge, Jessica Zosa Forde, Francesco Orabona, Sanmi Koyejo, and David Donoho. *Position: Machine Learning Conferences Should Establish a "Refutations and Critiques" Track*. 2025-07. DOI: 10.48550/arXiv.2506.19882. arXiv: 2506.19882 [cs]. (Visited on 2025-12-01) (cit. on p. 151).

[Sch+25b] Jan Carlo Schmid, Susanne Ibing, Stefan Kalabakov, Katharina Alefs, Robin P. van de Water, Arkadiusz Kwasigroch, Eugenia Alleva, Bert Arnrich, Bernhard Y. Renard, and Maia Kayal. "Expert-Defined, Chart-Reviewed Features Outperform EHR Foundation Models for Predicting Crohn's-like Disease of the Pouch". In: *Under review at The American Journal of Gastroenterology* (2025) (cit. on pp. xvii, 11).

[Sch+24a] Jelena Schmidt, Nienke M. Schutte, Stefan Buttigieg, David Novillo-Ortiz, Eric Sutherland, Michael Anderson, Bart de Witte, Michael Peolsson, Brigid Unim, Milena Pavlova, Ariel Dora Stern, Elias Mossialos, and Robin van Kessel. "Mapping the Regulatory Landscape for Artificial Intelligence in Health within the European Union". In: *npj Digital Medicine* 7.1 (2024-08), p. 229. ISSN: 2398-6352. DOI: 10.1038/s41746-024-01221-6. (Visited on 2025-12-03) (cit. on p. 150).

[Sch+24b] Daniel Schwabe, Katinka Becker, Martin Seyferth, Andreas Klaß, and Tobias Schaeffter. "The METRIC-framework for Assessing Data Quality for Trustworthy AI in Medicine: A Systematic Review". In: *npj Digital Medicine* 7.1 (2024-08), p. 203. ISSN: 2398-6352. DOI: 10.1038/s41746-024-01196-4. (Visited on 2025-12-03) (cit. on p. 149).

[Sem+25] Harald Semmelrock, Tony Ross-Hellauer, Simone Kopeinik, Dieter Theiler, Armin Haberl, Stefan Thalmann, and Dominik Kowald. "Reproducibility in Machine-Learning-Based Research: Overview, Barriers, and Drivers". In: *AI Magazine* 46.2 (2025), e70002. ISSN: 2371-9621. DOI: 10.1002/aaai.70002. (Visited on 2025-07-23) (cit. on pp. 24, 44, 101, 151, 152).

[Sey+16] Christopher W. Seymour, Vincent X. Liu, Theodore J. Iwashyna, Frank M. Brunkhorst, Thomas D. Rea, André Scherag, Gordon Rubenfeld, Jeremy M. Kahn, Manu Shankar-Hari, Mervyn Singer, Clifford S. Deutschman, Gabriel J. Escobar, and Derek C. Angus. "Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)". In: *JAMA* 315.8 (2016-02), p. 762. ISSN: 0098-7484. DOI: 10.1001/jama.2016.0288. (Visited on 2023-05-26) (cit. on pp. 50, 51, 196).

[SG17] Fred Shaffer and J. P. Ginsberg. "An Overview of Heart Rate Variability Metrics and Norms". In: *Frontiers in Public Health* 5 (2017-09), p. 258. ISSN: 2296-2565. DOI: 10.3389/fpubh.2017.00258. (Visited on 2025-08-11) (cit. on p. 78).

[SZC21] Farah Shamout, Tingting Zhu, and David A. Clifton. "Machine Learning for Clinical Outcome Prediction". In: *IEEE Reviews in Biomedical Engineering* 14 (2021), pp. 116–126. ISSN: 1937-3333, 1941-1189. DOI: 10.1109/RBME.2020.3007816. (Visited on 2022-05-02) (cit. on p. 46).

[Sha53] Lloyd S. Shapley. "A Value for N-Person Games". In: (1953). (Visited on 2026-01-05) (cit. on p. 62).

[Sha+17] Alok Sharma, Anupam Shukla, Ritu Tiwari, and Apoorva Mishra. "Mortality Prediction of ICU Patients Using Machine Leaning: A Survey". In: *Proceedings of the International Conference on Compute and Data Analysis - ICCDA '17*. Lakeland, FL, USA: ACM Press, 2017, pp. 49–53. ISBN: 978-1-4503-5241-3. DOI: 10.1145/3093241.3093267. (Visited on 2022-03-22) (cit. on p. 46).

[SBO20] Seyedmostafa Sheikhalishahi, Vevake Balaraman, and Venet Osmani. "Benchmarking Machine Learning Models on Multi-Centre eICU Critical Care Dataset". In: *PLOS ONE* 15.7 (2020-07). Ed. by Kyoung-Sae Na, e0235424. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0235424. (Visited on 2022-08-04) (cit. on pp. 28, 31).

[SRC24] Judy Hanwen Shen, Inioluwa Deborah Raji, and Irene Y. Chen. *The Data Addition Dilemma*. 2024-08. DOI: 10.48550/arXiv.2408.04154. arXiv: 2408.04154 [cs]. (Visited on 2025-04-24) (cit. on p. 54).

[She20] Alex Sherstinsky. "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network". In: *Physica D: Nonlinear Phenomena* 404 (2020-03), p. 132306. ISSN: 0167-2789. DOI: 10.1016/j.physd.2019.132306. (Visited on 2022-11-29) (cit. on p. 21).

[Shi+12] Hon-Yi Shi, King-Teh Lee, Hao-Hsien Lee, Wen-Hsien Ho, Ding-Ping Sun, Jhi-Joung Wang, and Chong-Chi Chiu. "Comparison of Artificial Neural Network and Logistic Regression Models for Predicting In-Hospital Mortality after Primary Liver Cancer Surgery". In: *PLOS ONE* 7.4 (2012-04), e35781. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0035781. (Visited on 2026-01-05) (cit. on p. 68).

[Shi+23] Benjamin Shickel, Tyler J. Loftus, Matthew Ruppert, Gilbert R. Upchurch, Tezcan Ozrazgat-Baslanti, Parisa Rashidi, and Azra Bihorac. "Dynamic Predictions of Postoperative Complications from Explainable, Uncertainty-Aware, and Multi-Task Deep Neural Networks". In: *Scientific Reports* 13.1 (2023-01), p. 1224. ISSN: 2045-2322. DOI: 10.1038/s41598-023-27418-5 (cit. on pp. 58, 93).

[Shi+19a] Duncan Shillan, Jonathan A. C. Sterne, Alan Champneys, and Ben Gibbison. "Use of Machine Learning to Analyse Routinely Collected Intensive Care Unit Data: A Systematic Review". In: *Critical Care* 23.1 (2019-12), p. 284. ISSN: 1364-8535. DOI: 10.1186/s13054-019-2564-9. (Visited on 2023-02-10) (cit. on pp. 27, 29, 46, 57).

[Shi+19b] Duncan Shillan, Jonathan A. C. Sterne, Alan Champneys, and Ben Gibbison. "Use of Machine Learning to Analyse Routinely Collected Intensive Care Unit Data: A Systematic Review". In: *Critical Care* 23.1 (2019-12), p. 284. ISSN: 1364-8535. DOI: 10.1186/s13054-019-2564-9 (cit. on p. 28).

[Shr+15] Mark G Shrime, Stephen W Bickler, Blake C Alkire, and Charlie Mock. "Global Burden of Surgical Disease: An Estimation from the Provider Perspective". In: *The Lancet Global Health*. Global Surgery 3 (2015-04), S8–S9. ISSN: 2214-109X. DOI: 10.1016/S2214-109X(14)70384-5. (Visited on 2025-08-06) (cit. on p. 55).

[Si+21] Yuqi Si, Jingcheng Du, Zhao Li, Xiaoqian Jiang, Timothy Miller, Fei Wang, W. Jim Zheng, and Kirk Roberts. "Deep Representation Learning of Patient Data from Electronic Health Records (EHR): A Systematic Review". In: *Journal of Biomedical Informatics* 115 (2021-03), p. 103671. ISSN: 15320464. DOI: 10.1016/j.jbi.2020.103671. (Visited on 2022-03-11) (cit. on pp. 24, 26).

[Sil+12] Ikaro Silva, George Moody, Daniel J. Scott, Leo A. Celi, and Roger G. Mark. "Predicting In-Hospital Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012". In: *Computing in Cardiology* 39 (2012), pp. 245–248. ISSN: 2325-8861 (cit. on p. 47).

[Sin+16] Mervyn Singer et al. "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)". In: *JAMA* 315.8 (2016-02), pp. 801–810. ISSN: 0098-7484. DOI: 10.1001/jama.2016.0287. (Visited on 2022-11-07) (cit. on pp. 28, 30, 50, 196).

[Smi+13] Gary B. Smith, David R. Prytherch, Paul Meredith, Paul E. Schmidt, and Peter I. Featherstone. "The Ability of the National Early Warning Score (NEWS) to Discriminate Patients at Risk of Early Cardiac Arrest, Unanticipated Intensive Care Unit Admission, and Death". In: *Resuscitation* 84.4 (2013-04), pp. 465–470. ISSN: 0300-9572, 1873-1570. DOI: 10.1016/j.resuscitation.2012.12.016. (Visited on 2025-09-15) (cit. on p. 58).

[Smi+22] F. Jasmijn Smits et al. "Algorithm-Based Care versus Usual Care for the Early Recognition and Management of Complications after Pancreatic Resection in the Netherlands: An Open-Label, Nationwide, Stepped-Wedge Cluster-Randomised Trial". In: *The Lancet* 399.10338 (2022-05), pp. 1867–1875. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(22)00182-9. (Visited on 2024-08-15) (cit. on pp. 58, 93).

[SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. *Practical Bayesian Optimization of Machine Learning Algorithms*. 2012-08. DOI: 10.48550/arXiv.1206.2944. arXiv: 1206.2944 [cs, stat]. (Visited on 2023-04-18) (cit. on pp. 41, 47).

[25] *Som-Shahlab/Femr*. Dr. Nigam Shah's research group at Stanford Medicine. 2025-05. (Visited on 2025-05-15) (cit. on p. 132).

[Sou+20] Esteban T. D. Souwer, Esther Bastiaannet, Ewout W. Steyerberg, Jan-Willem T. Dekker, Frederiek van den Bos, and Johanna E. A. Portielje. "Risk Prediction Models for Postoperative Outcomes of Colorectal Cancer Surgery in the Older Population - a Systematic Review". In: *JOURNAL OF GERIATRIC ONCOLOGY* 11.8 (2020-11), pp. 1217–1228. ISSN: 1879-4068. DOI: 10.1016/j.jgo.2020.04.006 (cit. on p. 57).

[Spa72] Karen Sparck Jones. "A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL". In: *Journal of Documentation* 28.1 (1972-01), pp. 11–21. ISSN: 0022-0418. DOI: 10.1108/eb026526. (Visited on 2026-01-20) (cit. on p. 118).

[SW19] Aaron Springer and Steve Whittaker. "Progressive Disclosure: Empirically Motivated Approaches to Designing Effective Transparency". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI '19. New York, NY, USA: Association for Computing Machinery, 2019-03, pp. 107–120. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302322. (Visited on 2025-10-27) (cit. on p. 113).

[Sta+22] Wessel T Stam, Lotte K Goedknegt, Erik W Ingwersen, Linda J Schoonmade, Emma R J Bruns, and Freek Daams. "The Prediction of Surgical Complications Using Artificial Intelligence in Patients Undergoing Major Abdominal Surgery: A Systematic Review." In: *Surgery* 171.4 (2022), pp. 1014–1021. ISSN: 1532-7361. DOI: 10.1016/j.surg.2021.10.002 (cit. on pp. 24–26, 57).

[Ste24] Ethan Steinberg. *Using Foundation Models to Learn How to Represent Electronic Health Records*. Stanford University, 2024 (cit. on pp. 13, 21).

[Ste+23] Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. "MOTOR: A Time-to-Event Foundation Model For Structured Medical Records". In: *The Twelfth International Conference on Learning Representations*. 2023-10. (Visited on 2025-08-11) (cit. on pp. 20, 22, 106–108, 127, 130, 140, 151).

[Ste+21] Ethan Steinberg, Ken Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam H. Shah. "Language Models Are an Effective Representation Learning Technique for Electronic Health Record Data". In: *Journal of Biomedical Informatics* 113 (2021-01), p. 103637. ISSN: 15320464. DOI: 10.1016/j.jbi.2020.103637. (Visited on 2023-07-26) (cit. on pp. 22, 105–108, 110, 117, 129, 138).

[Ste+24] Ethan Steinberg, Michael Wornow, Suhana Bedi, Jason Alan Fries, Matthew B. A. McDermott, and Nigam H. Shah. *Meds_reader: A Fast and Efficient EHR Processing Library*. 2024-11. DOI: 10.48550/arXiv.2409.09095. arXiv: 2409.09095 [cs]. (Visited on 2025-11-15) (cit. on pp. 129, 142).

[Stu08] Student. "The Probable Error of a Mean". In: *Biometrika* 6.1 (1908), pp. 1–25. ISSN: 0006-3444. DOI: 10.2307/2331554. JSTOR: 2331554. (Visited on 2026-01-05) (cit. on p. 61).

[Stu+24] Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. *Jina-Embeddings-v3: Multilingual Embeddings With Task LoRA*. 2024-09. DOI: 10.48550/arXiv.2409.10173. arXiv: 2409.10173 [cs]. (Visited on 2025-07-25) (cit. on pp. 79, 80, 86).

[Su+24] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. "RoFormer: Enhanced Transformer with Rotary Position Embedding". In: *Neurocomputing* 568 (2024-02), p. 127063. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2023.127063. (Visited on 2026-01-14) (cit. on p. 107).

[Su+25] Xiaorui Su, Shvat Messica, Yepeng Huang, Ruth Johnson, Lukas Fesser, Shanghua Gao, Faryad Sahneh, and Marinka Zitnik. *Multimodal Medical Code Tokenizer*. 2025-06. DOI: 10.48550/arXiv.2502.04397. arXiv: 2502.04397 [cs]. (Visited on 2025-11-29) (cit. on p. 130).

[Sut+24] Thomas M. Sutter, Yang Meng, Andrea Agostini, Daphné Chopard, Norbert Fortin, Julia E. Vogt, Babak Shahbaba, and Stephan Mandt. "Unity by Diversity: Improved Representation Learning for Multimodal VAEs". In: *Advances in Neural Information Processing Systems* 37 (2024-12), pp. 74262–74297. (Visited on 2025-10-22) (cit. on p. 22).

[Sye+21a] M Syed, S Syed, K Sexton, HB Syeda, M Garza, and ... "Application of Machine Learning in Intensive Care Unit (ICU) Settings Using MIMIC Dataset: Systematic Review". In: *Informatics* (2021) (cit. on pp. 24–26).

[Sye+21b] Mahanazuddin Syed, Shorabuddin Syed, Kevin Sexton, Hafsa Bareen Syeda, Maryam Garza, Meredith Zozus, Farhanuddin Syed, Salma Begum, Abdullah Usama Syed, Joseph Sanford, and Fred Prior. "Application of Machine Learning in Intensive Care Unit (ICU) Settings Using MIMIC Dataset: Systematic Review". In: *Informatics* 8.1 (2021-03), p. 16. ISSN: 2227-9709. DOI: 10.3390/informatics8010016. (Visited on 2022-03-22) (cit. on pp. 25, 29, 30, 33, 46, 47).

[Tan+20] Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. "Democratizing EHR Analyses with FIDDLE: A Flexible Data-Driven Preprocessing Pipeline for Structured Clinical Data". In: *Journal of the American Medical Informatics Association* 27.12 (2020-12), pp. 1921–1934. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa139. (Visited on 2022-11-28) (cit. on pp. 29, 31, 102).

[Tas+21] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. "CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 24804–24816. (Visited on 2023-08-09) (cit. on pp. 207, 208).

[Ter+25] Sanne Ter Horst, Raymond J. van Wijk, Anna D. Schoonhoven, Anouk de Lange, Jan C. Ter Maaten, and Hjalmar R. Bouma. "Pulse Oximetry beyond Oxygen Saturation: Early Waveform Characteristics in Sepsis Patients with Adverse Outcomes - A Proof-of-Concept Study". In: *Journal of Critical Care* 89 (2025-10), p. 155115. ISSN: 1557-8615. DOI: 10.1016/j.jcrc.2025.155115 (cit. on p. 94).

[Ter19] Maksim A. Terpilowski. "Scikit-Posthocs: Pairwise Multiple Comparison Tests in Python". In: *Journal of Open Source Software* 4.36 (2019-04), p. 1169. ISSN: 2475-9066. DOI: 10.21105/joss.01169. (Visited on 2026-01-21) (cit. on p. 241).

[Tho+21] Patrick J. Thoral, Jan M. Peppink, Ronald H. Driessen, Eric J. G. Sijbrands, Erwin J. O. Kompanje, Lewis Kaplan, Heatherlee Bailey, Jozef Kesecioglu, Maurizio Cecconi, Matthew Churpek, Gilles Clermont, Mihaela van der Schaar, Ari Ercole, Armand R. J. Girbes, and Paul W. G. Elbers. "Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example*". In: *Critical Care Medicine* 49.6 (2021-06), e563–e577. ISSN: 0090-3493. DOI: 10.1097/CCM.0000000000004916. (Visited on 2022-04-29) (cit. on pp. 9, 16, 33, 34, 57, 103, 104, 199).

[Tom+19] Nenad Tomašev et al. "A Clinically Applicable Approach to Continuous Prediction of Future Acute Kidney Injury". In: *Nature* 572.7767 (2019-08), pp. 116–119. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1390-1. (Visited on 2023-05-30) (cit. on p. 46).

[TS24] Akiyoshi Tomihari and Issei Sato. "Understanding Linear Probing Then Fine-tuning Language Models from NTK Perspective". In: *Advances in Neural Information Processing Systems* 37 (2024-12), pp. 139786–139822. DOI: 10.52202/079017-4436. (Visited on 2025-12-08) (cit. on p. 132).

[Tur+25] Alisher Turubayev, Anna Shopova, Fabian Lange, Mahmut Kamalak, Paul Mattes, Victoria Ayvasky, Bert Arnrich, Bjarne Pfitzner, and Robin P. van de Water. *Closing Gaps: An Imputation Analysis of ICU Vital Signs*. 2025-10. DOI: 10.48550/arXiv.2510.24217. arXiv: 2510.24217 [cs]. (Visited on 2025-10-29) (cit. on pp. 4, 11, 27).

[vBoe+24] Anna M. van Boekel et al. "Systematic Evaluation of Machine Learning Models for Postoperative Surgical Site Infection Prediction". In: *PLOS ONE* 19.12 (2024-12), e0312968. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0312968. (Visited on 2025-01-15) (cit. on pp. 58, 93).

[vRos+23] Mathilde C. van Rossum, Robin E. M. Bekhuis, Ying Wang, Johannes H. Hegeman, Ellis C. Folbert, Miriam M. R. Vollenbroek-Hutten, Cornelis J. Kalkman, Ewout A. Kouwenhoven, and Hermie J. Hermens. "Early Warning Scores to Support Continuous Wireless Vital Sign Monitoring for Complication Prediction in Patients on Surgical Wards: Retrospective Observational Study". In: *JMIR Perioperative Medicine* 6.1 (2023-08), e44483. DOI: 10.2196/44483. (Visited on 2024-12-20) (cit. on pp. 58, 59, 93).

[van+25a] Wouter van Amsterdam, Michael Kamp, Rajesh Ranganath, Robin P. van de Water, Florian Markowetz, Evangelia Christodoulou, Yamuna Krishnamurthy, Christoph Lippert, Jeff Clark, Julia E. Vogt, Raul Santos-Rodriguez, Thomas Gärtner, Gilbert Koch, and Brett Beaulieu-Jones. "Missing Incentives, Missing Oversight: The Challenges of AI Monitoring in Clinical Practice". 2025 (cit. on pp. 11, 148).

[vdVan25] Robin van de Van de Water. *Rvandewater/Harmonizehr*. 2025-11. (Visited on 2026-01-19) (cit. on pp. 100, 103).

[vdWM25] Robin van de Water and Matthew McDermott. *Rvandewater/NWICU_MEDS: 0.0.11*. Zenodo. 2025-03. DOI: 10.5281/zenodo.15025635. (Visited on 2025-04-11) (cit. on p. 142).

[Van+23] L. F. Van de Water et al. "Effect of a Prediction Tool and Communication Skills Training on Communication of Treatment Outcomes: A Multicenter Stepped Wedge Clinical Trial (the SOURCE Trial)". In: *eClinicalMedicine* 64 (2023-10), p. 102244. ISSN: 2589-5370. DOI: 10.1016/j.eclinm.2023.102244. (Visited on 2025-12-03) (cit. on p. 150).

[Van25a] Robin Van de Water. *Rvandewater/FHIR_MEDS: 0.0.3*. Zenodo. 2025-07. DOI: 10.5281/ZENODO.16269018. (Visited on 2025-10-19) (cit. on pp. 100, 109, 113).

[Van25b] Robin Van de Water. *Rvandewater/INSPIRE_MEDS: Zenodo*. Zenodo. 2025-02. DOI: 10.5281/ZENODO.14891940. (Visited on 2025-10-16) (cit. on p. 142).

[VRR25] Robin Van de Water, Patrick Rockenschaub, and Pawel Renc. *Prockenschaub/AUMCdb_MEDS: 0.0.3*. Zenodo. 2025-11. DOI: 10.5281/ZENODO.17531825. (Visited on 2025-11-15) (cit. on pp. 9, 100, 104, 142).

[Van+24a] Robin Van de Water, Hendrik Nils Aurel Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. "Yet Another ICU Benchmark: A Flexible Multi-Center Framework for Clinical ML". In: *Proceedings of The Twelfth International Conference on Learning Representations*. 2024-05. DOI: 10.48550/arXiv.2306.05109. (Visited on 2025-01-31) (cit. on pp. xv, 3, 8, 27, 31, 44, 57, 84, 93, 100, 149).

[Van+22a] Robin Van de Water, Francesco Ventura, Z Kaoudi, J Quiane-Ruiz, and Volker Markl. "Farm Your ML-based Query Optimizer's Food!–Human-Guided Training Data Generation–". In: *Conference on Innovative Data Systems Research (CIDR)*. 2022 (cit. on pp. 8, 11).

[Van+22b] Robin Van de Water, Francesco Ventura, Zoi Kaoudi, Jorge-Arnulfo Quiane-Ruiz, and Volker Markl. "Farming Your ML-based Query Optimizer's Food". In: *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. Kuala Lumpur, Malaysia: IEEE, 2022-05, pp. 3186–3189. ISBN: 978-1-6654-0883-7. DOI: 10.1109/ICDE53745.2022.00294. (Visited on 2024-04-24) (cit. on pp. 8, 11, 119).

[Van+24b] Robin Van de Water, Axel Winter, Max M. Maurer, Felix August Treykorn, Bjarne Pfitzner, Igor M. Sauer, and Bert Arnrich. "Combining Time Series Modalities to Create Endpoint-driven Patient Records". In: *ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR): Harnessing Momentum for Science*. 2024-03. (Visited on 2024-04-03) (cit. on pp. xvi, 5, 8, 55).

[Van+24c] Robin Van de Water, Axel Winter, Max M. Maurer, Felix August Treykorn, Daniela Zuluaga, Bjarne Pfitzner, Igor M. Sauer, and Bert Arnrich. "Combining Hospital-grade Clinical Data and Wearable Vital Sign Monitoring to Predict Surgical Complications". In: *ICLR 2024 Workshop on Learning from Time Series For Health*. 2024-03. (Visited on 2024-04-03) (cit. on pp. xv, 5, 8, 55).

[VZM25] Robin Van de Water, Mingcheng Zhu, and Matthew McDermott. *Medical-Event-Data-Standard/eICU_-MEDS: 0.0.2*. Zenodo. 2025-11. DOI: 10.5281/ZENODO.17535692. (Visited on 2025-11-15) (cit. on pp. 9, 104, 142).

[vZM25] Robin van de Water, Mingcheng Zhu, and Matthew McDermott. *Medical-Event-Data-Standard/eICU_-MEDS: 0.0.2*. Zenodo. 2025-11. DOI: 10.5281/ZENODO.17535692. (Visited on 2026-01-15) (cit. on p. 100).

[Van25c] Robin P. Van de Water. "Scaling Up Clinical ML from Datasets to Entire Health Systems through the MEDS Ecosystem". In: *SBHD 2025 International Conference on Systems Biology of Human Diseases*. Berlin, 2025-06 (cit. on pp. xv, 7, 99).

[Van+25a] Robin P. Van de Water, Katharina Marie Alefs, Susanne Ibing, Stefan Kalabakov, Aadil Rasheed, Pia Francesca Rissom, Jan Carlo Schmid, Arkadiusz Kwasigroch, Eugenia Alleva, and Christoph Lippert. "FLAMES: EHR Foundation Models as a Resource for Health Systems". 2025 (cit. on pp. 7, 99, 100).

[Van+25b] Robin P. Van de Water, Christoph Riepe, Bjarne Pfitzner, Lara Faraj, Daniela Zuluaga, Christian Schineis, Katharina Beyer, Johann Pratschke, Igor Sauer, Axel Winter, and Max Maurer. "Machine Learning for 30-Day Mortality Prediction for High-Risk General Emergency Surgery". In: *In preparation* (2025) (cit. on pp. 55, 58, 59).

[VSR26] Robin P. Van de Water, Hendrik Schmidt, and Patrick Rockenschaub. "ReciPies: A Lightweight Data Transformation Pipeline for Reproducible ML". In: *Journal of Open Source Software* 11.117 (2026), p. 9261. DOI: 10.21105/joss.09261 (cit. on pp. xiv, 3, 8, 27, 43).

[Van+25c] Robin P. Van de Water, Axel Winter, Daniela Zuluaga Lotero, Bjarne Pfitzner, Lara Faraj, Bert Arnrich, Patrick Rockenschaub, Wenzel Schoning, Thomas Malinka, Christian Denecke, Johann Pratschke, Igor M. Sauer, and Max M. Maurer. *Continuous Multimodal AI with Wearable Vital Signs Predicts Postoperative Complications in the General Ward*. 2025-11. DOI: 10.1101/2025.11.25.25340950. (Visited on 2025-11-28) (cit. on pp. 4, 55, 59).

[Van25d] Robin Philippus Van de Water. *HIRID MEDS ETL*. Zenodo. 2025-10. DOI: 10.5281/zenodo.17370279. (Visited on 2025-10-18) (cit. on pp. 9, 100, 104, 142).

[Van25e] Robin Philippus Van de Water. *MEDS-Inspect*. 2025-02. DOI: 10.5281/zenodo.14891940. (Visited on 2025-04-04) (cit. on pp. 100, 112).

[van25a] Robin Philippus van de Water. *FHIR2MEDS ETL (MEDS on FHIR)*. 2025-07. DOI: 10.5281/zenodo.16269018. (Visited on 2026-01-15) (cit. on pp. 9, 104).

[van25b] Robin Philippus van de Water. *INSPIRE_MEDS ETL*. 2025-02. DOI: 10.5281/zenodo.14891940. (Visited on 2026-01-15) (cit. on pp. 9, 100, 104).

[van25c] Robin Philippus van de Water. *NWICU_MEDS ETL*. 2025-02. DOI: 10.5281/zenodo.14892134. (Visited on 2026-01-15) (cit. on pp. 9, 100, 104).

[van25d] Robin Philippus van de Water. *OMOP_MEDS ETL*. Zenodo. 2025-04. DOI: 10.5281/zenodo.15183099. (Visited on 2025-04-11) (cit. on pp. 9, 100, 104, 109, 142).

[van25e] Robin Philippus van de Water. *SICdb_MEDS ETL*. Zenodo. 2025-02. DOI: 10.5281/zenodo.14893939. (Visited on 2025-04-11) (cit. on pp. 9, 100, 104, 142).

[Van+25d] Robin Philippus Van de Water, Ethan Steinberg, Michael Wornow, Patrick Rockenschaub, and Matthew McDermott. *MIMIC-IV Demo Data in the Medical Event Data Standard (MEDS)*. 2025. DOI: 10.13026/T2Y8-EA41. (Visited on 2025-09-30) (cit. on pp. xiv, 9, 113).

[van+21] Tom van den Bosch, Anne-Loes K. Warps, Michael P. M. de Nerée tot Babberich, Christina Stamm, Bart F. Geerts, Louis Vermeulen, Michel W. J. M. Wouters, Jan Willem T. Dekker, Rob A. E. M. Tollenaar, Pieter J. Tanis, Daniël M. Miedema, and Dutch ColoRectal Audit. "Predictors of 30-Day Mortality Among Dutch Patients Undergoing Colorectal Cancer Surgery, 2011-2016". In: *JAMA Network Open* 4.4 (2021-04), e217737. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2021.7737. (Visited on 2026-01-05) (cit. on pp. 67, 68).

[van+25b] Siri L. van der Meijden et al. "Development and Validation of Artificial Intelligence Models for Early Detection of Postoperative Infections (PERISCOPE): A Multicentre Study Using Electronic Health Record Data". In: *The Lancet Regional Health - Europe* 49 (2025-02), p. 101163. ISSN: 2666-7762. DOI: 10.1016/j.lanepe.2024.101163. (Visited on 2025-01-15) (cit. on p. 96).

[van+23] Jonna A. van der Stam, Eveline H. J. Mestrom, Simon W. Nienhuijs, Ignace H. J. T. de Hingh, Arjen-Kars Boer, Natal A. W. van Riel, Koen T. J. de Groot, Wim Verhaegh, Volkher Scharnhorst, and R. Arthur Bouwman. "A Wearable Patch Based Remote Early Warning Score (REWS) in Major Abdominal Cancer Surgery Patients". In: *European Journal of Surgical Oncology* 49.1 (2023-01), pp. 278–284. ISSN: 0748-7983. DOI: 10.1016/j.ejso.2022.08.034. (Visited on 2024-12-13) (cit. on pp. 58, 59, 82, 85, 88, 89, 91).

[vdMH08] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data Using T-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. ISSN: 1533-7928. (Visited on 2025-10-03) (cit. on pp. 79, 90, 91, 137, 291).

[Var+24] Chris Varghese, Ewen M. Harrison, Greg O'Grady, and Eric J. Topol. "Artificial Intelligence in Surgery". In: *Nature Medicine* 30.5 (2024-05), pp. 1257–1268. ISSN: 1546-170X. DOI: 10.1038/s41591-024-02970-3. (Visited on 2024-10-30) (cit. on p. 56).

[Var24] Authors Various. "ACM REP '24: Proceedings of the 2nd ACM Conference on Reproducibility and Replicability". In: *ACM REP '24: Proceedings of the 2nd ACM Conference on Reproducibility and Replicability*. Rennes, France and New York, NY, USA: Association for Computing Machinery, 2024. ISBN: 979-8-4007-0530-4. DOI: 10.1145/3641525 (cit. on p. 44).

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need". In: *arXiv:1706.03762 [cs]* (2017-12). arXiv: 1706.03762 [cs]. (Visited on 2022-02-23) (cit. on pp. 21, 47, 82, 85, 94, 105, 107).

[Vav+16] George Vavoulas, Charikleia Chatzaki, Thodoris Malliotakis, Matthew Pediaditis, and Manolis Tsiknakis. *The MobiAct Dataset: Recognition of Activities of Daily Living Using Smartphones*. 2016-01, p. 151. DOI: 10.5220/0005792401430151 (cit. on p. 15).

[Ver+19] Sue Carol Verrillo, Maria Cvach, Krysia Warren Hudson, and Bradford D. Winters. "Using Continuous Vital Sign Monitoring to Detect Early Deterioration in Adult Postoperative Inpatients". In: *Journal of Nursing Care Quality* 34.2 (2019), pp. 107–113. ISSN: 1550-5065. DOI: 10.1097/NCQ.0000000000000350 (cit. on p. 93).

[Vin+24] Ritchie Vink et al. *Pola-Rs/Polars: Python Polars 1.0.0*. Zenodo. 2024-07. DOI: 10.5281/zenodo.12606903. (Visited on 2025-10-15) (cit. on pp. 42, 44, 113).

[Vir+20] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2 (cit. on p. 210).

[Voh16] Deepak Vohra. "Apache Parquet". In: *Practical Hadoop Ecosystem: A Definitive Guide to Hadoop-Related Frameworks and Tools*. Ed. by Deepak Vohra. Berkeley, CA: Apress, 2016, pp. 325–335. ISBN: 978-1-4842-2199-0. DOI: 10.1007/978-1-4842-2199-0_8. (Visited on 2023-05-23) (cit. on p. 198).

[Von+11] René Vonlanthen, Ksenija Slankamenac, Stefan Breitenstein, Milo A. Puhan, Markus K. Muller, Dieter Hahnloser, Dimitri Hauri, Rolf Graf, and Pierre-Alain Clavien. "The Impact of Complications on Costs of Major Surgical Procedures: A Cost Analysis of 1200 Patients". In: *Annals of Surgery* 254.6 (2011-12), pp. 907–913. ISSN: 1528-1140. DOI: 10.1097/SLA.0b013e31821d4a43 (cit. on p. 56).

[Vos+15] Erica A. Voss, Rupa Makadia, Amy Matcho, Qianli Ma, Chris Knoll, Martijn Schuemie, Frank J. DeFalco, Ajit Londhe, Vivienne Zhu, and Patrick B. Ryan. "Feasibility and Utility of Applications of the Common Data Model to Multiple, Disparate Observational Health Databases". In: *Journal of the American Medical Informatics Association: JAMIA* 22.3 (2015-05), pp. 553–564. ISSN: 1527-974X. DOI: 10.1093/jamia/ocu023 (cit. on pp. 9, 104, 117).

[Wan+20a] Hui Wang, Fuxing Deng, Buyao Zhang, and Shuangping Zhao. *Real-Time Prediction of AKI Among Middle-Aged and Older in ICU: A Retrospective and Machine Learning Study*. Preprint. In Review, 2020-08. DOI: 10.21203/rs.3.rs-57905/v1. (Visited on 2022-01-27) (cit. on p. 46).

[Wan+19] Shiqi Wang, Lei Xu, Quan Wang, Jipeng Li, Bin Bai, Zhengyan Li, Xiaoyong Wu, Pengfei Yu, Xuzhao Li, and Jichao Yin. "Postoperative Complications and Prognosis after Radical Gastrectomy for Gastric Cancer: A Systematic Review and Meta-Analysis of Observational Studies". In: *World Journal of Surgical Oncology* 17.1 (2019-03), p. 52. ISSN: 1477-7819. DOI: 10.1186/s12957-019-1593-9 (cit. on p. 56).

[Wan+20b] Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C. Hughes, and Tristan Naumann. "MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III". In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. Toronto Ontario Canada: ACM, 2020-04, pp. 222–235. ISBN: 978-1-4503-7046-2. DOI: 10.1145/3368555.3384469. (Visited on 2022-11-01) (cit. on pp. 29, 31, 102).

[Wan+17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017-07, pp. 3462–3471. DOI: 10.1109/CVPR.2017.369. arXiv: 1705.02315 [cs]. (Visited on 2025-08-01) (cit. on p. 15).

[Wan+22] Yuqing Wang, Yun Zhao, Rachael Callcut, and Linda Petzold. *Integrating Physiological Time Series and Clinical Notes with Transformer for Early Prediction of Sepsis*. 2022-03. arXiv: 2203.14469 [cs]. (Visited on 2022-06-13) (cit. on p. 46).

[War+25] Vincent D Warmerdam et al. *Koaning/Scikit-Lego: V0.9.5*. Zenodo. 2025-04. DOI: 10.5281/zenodo.15313097. (Visited on 2025-07-22) (cit. on p. 44).

[Wax+25] Shane Waxler et al. *Generative Medical Event Models Improve with Scale*. 2025-08. DOI: 10.48550/arXiv.2508.12104. arXiv: 2508.12104 [cs]. (Visited on 2025-08-22) (cit. on pp. 97, 107, 108, 141, 150, 151).

[Wee+19] Mariska Weenk, Mats Koeneman, Tom H. van de Belt, Lucien J. L. P. G. Engelen, Harry van Goor, and Sebastian J. H. Bredie. "Wireless and Continuous Monitoring of Vital Signs in Patients at the General Ward". In: *Resuscitation* 136 (2019-03), pp. 47–53. ISSN: 1873-1570. DOI: 10.1016/j.resuscitation.2019.01.017 (cit. on p. 93).

[WL14] Justin M. Weis and Paul C. Levy. "Copy, Paste, and Cloned Notes in Electronic Health Records". In: *Chest* 145.3 (2014-03), pp. 632–638. ISSN: 1931-3543. DOI: 10.1378/chest.13-0886 (cit. on p. 138).

[WFH18] Robert S. Weller, Kristina L. Foard, and Timothy N. Harwood. "Evaluation of a Wireless, Portable, Wearable Multi-Parameter Vital Signs Monitor in Hospitalized Neurological and Neurosurgical Patients". In: *Journal of Clinical Monitoring and Computing* 32.5 (2018-10), pp. 945–951. ISSN: 1573-2614. DOI: 10.1007/s10877-017-0085-0 (cit. on p. 93).

[WF89] L. E. Widman and G. L. Freeman. "A-to-D Conversion from Paper Records with a Desktop Scanner and a Microcomputer". In: *Computers and Biomedical Research, an International Journal* 22.4 (1989-08), pp. 393–404. ISSN: 0010-4809. DOI: 10.1016/0010-4809(89)90033-5 (cit. on p. 13).

[Win+25] Axel Winter, Bjarne Pfitzner, Robin P. van de Water, Lara Faraj, Christoph Riepe, Wolf-Heinrich Hahn, Felix Krenzien, Christian Schineis, Thomas Malinka, and Wenzel Schöning. "Overcoming the Data Barrier: Transfer Learning for 90-Day Mortality Prediction in General Surgery–a Retrospective Multicenter Development and Comparison Study". In: *International Journal of Surgery* (2025), pp. 10–1097. (Visited on 2025-11-23) (cit. on pp. xiv, 55, 58, 68, 117).

[Win+24] Axel Winter, Robin P. van de Water, Bjarne Pfitzner, Marius Ibach, Christoph Riepe, Robert Ahlborn, Lara Faraj, Felix Krenzien, Eva M. Dobrindt, and Jonas Raakow. "Enhancing Preoperative Outcome Prediction: A Comparative Retrospective Case–Control Study on Machine Learning versus the International Esodata Study Group Risk Model for Predicting 90-Day Mortality in Oncologic Esophagectomy". In: *Cancers* 16.17 (2024), p. 3000. (Visited on 2024-10-23) (cit. on pp. xvi, 8, 10, 55, 58).

[Wol+20] Thomas Wolf et al. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. 2020-07. DOI: 10.48550/arXiv.1910.03771. arXiv: 1910.03771 [cs]. (Visited on 2025-11-06) (cit. on pp. 79, 142, 149).

[Won+21] Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients". In: *JAMA internal medicine* 181.8 (2021-08), pp. 1065–1070. ISSN: 2168-6114. DOI: 10.1001/jamainternmed.2021.2626 (cit. on p. 148).

[Woo08] R. F. Woolson. "Wilcoxon Signed-Rank Test". In: *Wiley Encyclopedia of Clinical Trials*. John Wiley & Sons, Ltd, 2008, pp. 1–3. ISBN: 978-0-471-46242-2. DOI: 10.1002/9780471462422.eoct979. (Visited on 2026-01-21) (cit. on p. 87).

[Wor23] World Health Organization. *International Classification of Health Interventions (ICHI)*. Geneva, 2023 (cit. on p. 60).

[Wor+24] Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Re, Sanmi Koyejo, and Nigam Shah. "Context Clues: Evaluating Long Context Models for Clinical Prediction Tasks on EHR Data". In: *The Thirteenth International Conference on Learning Representations*. 2024-10. (Visited on 2025-04-11) (cit. on pp. 22, 106–108, 120, 127–131, 138–140).

[Wor+23a] Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Alan Fries, and Nigam Shah. "EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models". In: *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2023-11. (Visited on 2024-02-19) (cit. on pp. 9, 16, 21, 103, 104, 110, 111, 121, 123, 149).

[Wor+23b] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. "The Shaky Foundations of Large Language Models and Foundation Models for Electronic Health Records". In: *npj Digital Medicine* 6.1 (2023-07), p. 135. ISSN: 2398-6352. DOI: 10.1038/s41746-023-00879-8. (Visited on 2025-07-26) (cit. on pp. 13, 21, 100).

[XCS18] Cao Xiao, Edward Choi, and Jimeng Sun. "Opportunities and Challenges in Developing Deep Learning Models Using Electronic Health Records Data: A Systematic Review". In: *Journal of the American Medical Informatics Association* 25.10 (2018-10), pp. 1419–1428. ISSN: 1067-5027, 1527-974X. DOI: 10.1093/jamia/ocy068. (Visited on 2022-03-22) (cit. on pp. 24–26).

[Xie+22] Feng Xie, Han Yuan, Yilin Ning, Marcus Eng Hock Ong, Mengling Feng, Wynne Hsu, Bibhas Chakraborty, and Nan Liu. "Deep Learning for Temporal Data Representation in Electronic Health Records: A Systematic Review of Challenges and Methodologies". In: *Journal of Biomedical Informatics* 126 (2022-02), p. 103980. ISSN: 15320464. DOI: 10.1016/j.jbi.2021.103980. (Visited on 2022-03-09) (cit. on p. 26).

[Xu+24] Justin Xu, Jack Gallifant, Alistair Johnson, and Matthew B. A. McDermott. "ACES: Automatic Cohort Extraction System for Event-Stream Datasets". In: *The Thirteenth International Conference on Learning Representations*. 2024-10. (Visited on 2025-05-14) (cit. on pp. 114, 121, 127, 142).

[Xu+22] Ping Xu, Lin Chen, Yuanfang Zhu, Shuai Yu, Rangui Chen, Wenbin Huang, Fuli Wu, and Zhongheng Zhang. "Critical Care Database Comprising Patients With Infection". In: *Frontiers in Public Health* 10 (2022-03). ISSN: 2296-2565. DOI: 10.3389/fpubh.2022.852410. (Visited on 2026-01-17) (cit. on p. 16).

[Xu+14] Zhixiang Xu, Gao Huang, Kilian Q. Weinberger, and Alice X. Zheng. "Gradient Boosted Feature Selection". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '14. New York, NY, USA: Association for Computing Machinery, 2014-08, pp. 522–531. ISBN: 978-1-4503-2956-9. DOI: 10.1145/2623330.2623635. (Visited on 2025-10-04) (cit. on p. 94).

[Yad+24] Ambuj Yadav, Himanshu Dandu, Gaurav Parchani, Kumar Chokalingam, Pooja Kadambi, Rajesh Mishra, Ahsina Jahan, Jean-Louis Teboul, and Jos M. Latour. "Early Detection of Deteriorating Patients in General Wards through Continuous Contactless Vital Signs Monitoring". In: *Frontiers in Medical Technology* 6 (2024-08), p. 1436034. ISSN: 2673-3129. DOI: 10.3389/fmedt.2024.1436034. (Visited on 2025-08-14) (cit. on pp. 59, 92).

[Yan+24] An Yang et al. *Qwen2 Technical Report*. 2024-09. DOI: 10.48550/arXiv.2407.10671. arXiv: 2407.10671 [cs]. (Visited on 2025-11-15) (cit. on p. 108).

[Yan+23a] Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin P. Danek, and Jimeng Sun. "PyHealth: A Deep Learning Toolkit for Healthcare Applications". In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '23. New York, NY, USA: Association for Computing Machinery, 2023-08, pp. 5788–5789. ISBN: 979-8-4007-0103-0. DOI: 10.1145/3580305.3599178. (Visited on 2023-11-13) (cit. on pp. 29, 31, 32, 149).

[Yan+22] Xi Yang et al. *GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records*. 2022-12. DOI: 10.48550/arXiv.2203.03540. arXiv: 2203.03540 [cs]. (Visited on 2025-08-13) (cit. on pp. 22, 106, 107).

[Yan+23b] Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. "TransformEHR: Transformer-Based Encoder-Decoder Generative Model to Enhance Prediction of Disease Outcomes Using Electronic Health Records". In: *Nature Communications* 14.1 (2023-11), p. 7857. ISSN: 2041-1723. DOI: 10.1038/s41467-023-43715-z. (Visited on 2025-08-13) (cit. on pp. 22, 106, 107).

[Yèc+22] Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Rätsch. "HiRID-ICU-Benchmark – A Comprehensive Machine Learning Benchmark on High-resolution ICU Data". In: *arXiv:2111.08536 [cs]* (2022-01). arXiv: 2111.08536 [cs]. (Visited on 2022-03-18) (cit. on pp. 28, 31, 46, 47, 49, 104, 211, 212, 214).

[YJG03] Andy B. Yoo, Morris A. Jette, and Mark Grondona. "SLURM: Simple Linux Utility for Resource Management". In: *Job Scheduling Strategies for Parallel Processing*. Ed. by Dror Feitelson, Larry Rudolph, and Uwe Schwiegelshohn. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2003, pp. 44–60. ISBN: 978-3-540-39727-4. DOI: 10.1007/10968987_3 (cit. on p. 211).

[Yu+25] Junhan Yu, Zhunyi Feng, Junwei Lu, Tianxi Cai, and Doudou Zhou. *Time-Aware Attention for Enhanced Electronic Health Records Modeling*. 2025-07. DOI: 10.48550/arXiv.2507.14847. arXiv: 2507.14847 [cs]. (Visited on 2026-01-20) (cit. on p. 140).

[Zah+22] Nicole Zahradka, Sophie Geoghan, Hope Watson, Eli Goldberg, Adam Wolfberg, and Matt Wilkes. "Assessment of Remote Vital Sign Monitoring and Alarms in a Real-World Healthcare at Home Dataset". In: *Bioengineering (Basel, Switzerland)* 10.1 (2022-12), p. 37. ISSN: 2306-5354. DOI: 10.3390/bioengineering10010037 (cit. on pp. 58, 93).

[ZLL22] Xianlong Zeng, Simon L. Linwood, and Chang Liu. "Pretrained Transformer Framework on Pediatric Claims Data for Population Specific Tasks". In: *Scientific Reports* 12.1 (2022-03), p. 3651. ISSN: 2045-2322. DOI: 10.1038/s41598-022-07545-1. (Visited on 2025-10-24) (cit. on p. 107).

[Zha+22] Angela Zhang, Lei Xing, James Zou, and Joseph C. Wu. "Shifting Machine Learning for Healthcare from Development to Deployment and from Models to Data". In: *Nature Biomedical Engineering* 6.12 (2022-12), pp. 1330–1345. ISSN: 2157-846X. DOI: 10.1038/s41551-022-00898-y. (Visited on 2025-07-24) (cit. on p. 1).

[Zha+25a] Sheng Zhang, Qin Liu, Naoto Usuyama, Cliff Wong, Tristan Naumann, and Hoifung Poon. *Exploring Scaling Laws for EHR Foundation Models*. 2025-08. DOI: 10.48550/arXiv.2505.22964. arXiv: 2505.22964 [cs]. (Visited on 2025-10-14) (cit. on pp. 107, 148).

[ZZG24] Fei Zhao, Chengcui Zhang, and Baocheng Geng. "Deep Multimodal Data Fusion". In: *ACM Comput. Surv.* 56.9 (2024-04), 216:1–216:36. ISSN: 0360-0300. DOI: 10.1145/3649447. (Visited on 2025-10-16) (cit. on pp. 22, 76).

[Zha+21] Yue Zhao, Zhi Qiao, Cao Xiao, Lucas Glass, and Jimeng Sun. *PyHealth: A Python Library for Health Predictive Models*. 2021-01. arXiv: 2101.04209 [cs]. (Visited on 2023-05-02) (cit. on pp. 29, 102).

[Zha+25b] Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, and Dinggang Shen. "CLIP in Medical Imaging: A Survey". In: *Medical Image Analysis* 102 (2025-05), p. 103551. ISSN: 13618415. DOI: 10.1016/j.media.2025.103551. arXiv: 2312.07353 [cs]. (Visited on 2025-10-22) (cit. on p. 22).

[Zho+23] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. "Domain Generalization: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (2023-04), pp. 4396–4415. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2022.3195549. (Visited on 2025-07-30) (cit. on p. 24).

Appendix A

Disclosure

A.1 Funding and Affiliations

I was funded by the *"Gemeinsamer Bundesausschuss (G-BA) Innovationsausschuss" in the framework of "CASSANDRA - Clinical ASSist AND aleRt Algorithms"* (project number 01VSF20015) from 2021-2024 and in 2024-2025 I have received funding from the *European Commission under the Horizon 2020 project INTERVENE* (Grant agreement ID: 101016775).

My primary affiliation was with the *Hasso Plattner Institute* (HPI) at the University of Potsdam, Germany during the entirety of this PhD work. I have published while being a guest scientist (Gastwissenschaftler) at the *Charité - Universitätsmedizin Berlin* from 2021-2026. I have been working with data from the Mount Sinai Health System and therefore affiliated with the *Hasso Plattner Institute for Digital Health at Mount Sinai*, *Icahn School of Medicine at Mount Sinai*, *New York* from 2024-2026. In 2023 I was a visiting researcher at the University of California, Irvine (UCI).

A.2 Scientific Computing

Compute resources were provided and used depending on the data-privacy regulations:

- All work with the Mount Sinai Health System data was performed on secure on-premise Health Insurance Portability and Accountability Act (HIPAA) Minerva cluster for which the following acknowledgement applies: *This work was supported in part through the Minerva computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences.*
- All work with Charité - Universitätsmedizin Berlin data (Chapter 4) was performed on the Charité high-performance computing (HPC) cluster for which the following acknowledgement applies: *The authors acknowledge the Scientific Computing of the IT Division at the Charité - Universitätsmedizin Berlin for providing*

computational resources that have contributed to the research results reported in this paper.[1]

- All remaining compute work (Chapters 3 and 5) was performed on the Hasso Plattner Institute (HPI) Scientific Compute Infrastructure cluster for which I would like to acknowledge the support of the HPI DHCLab (dissolved) and HPI Scientific Compute Infrastructure Team.

A.3 Use of Generative AI

The author has used generative AI (Large Language Models) in the preparation of this thesis in accordance with the guidelines formulated in Porsdam Mann et al. [Por+24]. The following is a verbatim excerpt of these guidelines:

> 1. Human vetting and guaranteeing. At least one author must be able to guarantee and take responsibility for the accuracy and integrity of the article, including each substantive claim and supporting evidence.
>
> 2. Substantial human contribution. Each author must have provided a substantial contribution to (a) the conception or design of the work; (b) the acquisition, analysis or interpretation of the data or source material; or (c) the design of prompts, embeddings or fine-tuning procedures, as well as the selection and curation of LLM outputs, and the integration of those outputs into the manuscript.
>
> 3. Acknowledgement and transparency. Authors should acknowledge their use of LLMs in the manner and to the extent required for (a) understanding, verifying, replicating, or otherwise judging the credibility of the work (outcome goods); as well as (b) confirming that authorship credit and responsibility have been assigned appropriately (process goods), (c) in a way that is no more burdensome than is what is expected (within each field of research) for analogous works produced without the use of an LLM.
>
> *Porsdam Mann et al. [Por+24]*

The following statement summarizes the adherence to these guidelines in this thesis: *Any use of generative AI in this thesis adheres to ethical guidelines for use and acknowledgement of generative AI in academic research. All work presented represents a substantial intellectual contribution by the author, has been thoroughly vetted for accuracy, and the author assumes full responsibility for the integrity of all content.*

[1] https://www.charite.de/en/research/research_support_services/research_infrastructure/science_it/#c30646061[Accessed January 9, 2026]

Appendix B

Supplements Chapter 2

B.1 Literature Review

Our systematic literature review as performed with the databases of Google Scholar, PubMed, Web of Science, Cochrane Library, IEEE Explore, and the ACM Computing Library. To produce our results, we used the following query, separated by the "and" operator:
This produced 129 results. Duplicate removal left us with 42 entries. Title screening in addition to using snowballing (also called citation tracking) left 30 entries. Abstract screening then left 17 items, which have been fully read and analyzed.

Appendix C

Supplements Chapter 3

C.1 Extended Results

This Appendix contains results that were left out of the main text.

Table C.1: Comparing the use of dynamic feature generation (FG) to the baseline of ICU mortality prediction, AUROC (↑). Note that an otherwise identical experiment setup was used to obtain results for the "without feature generation" results.

	AUMCdb		HiRID		eICU		MIMIC-IV	
Preprocessing	w/ FG	w/o FG	w/ FG	w/o FG	w/ FG	w/o FG	w/ FG	w/o FG
LR	83.7±0.6	82.2±0.5	84.0±0.3	81.8±0.7	84.8±0.2	81.1±0.1	86.1±0.1	80.2±0.3
LGBM	**84.5±0.6**	83.5±0.4	**84.5±0.3**	82.5±0.6	**85.7±0.2**	83.5±0.3	**87.7±0.2**	85.9±0.1

Table C.2: Comparing the use of dynamic feature generation (FG) to the baseline of ICU mortality prediction, AUPRC (↑). Note that an otherwise identical experiment setup was used to obtain results for the "without feature generation" results.

	AUMCdb		HiRID		eICU		MIMIC-IV	
Preprocessing	w/ FG	w/o FG	w/ FG	w/o FG	w/ FG	w/o FG	w/ FG	w/o FG
LR	**52.9±1.2**	50.8±1.2	36.9±1.1	33.1±0.8	33.0±0.7	28.3±0.4	39.7±0.6	32.1±0.6
LGBM	**53.7±1.2**	50.9±1.0	**40.6±0.8**	35.3±0.9	**36.0±0.6**	32.5±0.9	**44.2±0.7**	40.1±0.7

Feature generation We compare the use of feature generation for classical ML models. The ReciPies package provides the functionality of assembling different preprocessing steps to be supplied by the user (i.e., a recipe). We show the results in Table C.1 and C.2.

The impact of static features We leveraged this customizable preprocessing to perform training prediction models without static features (i.e., age, sex, height, and weight). The AUROC results are found in Table 3.8 (the AUPRC in Table C.3). From these results, we can see that including the static data seems to result in better performance across all models and datasets.

Table C.3: Baseline AUPRC (↑) performance comparison of including static data for ICU Mortality Prediction.

	AUMCdb		HiRID		eICU		MIMIC-IV	
Inclusion	w/ static	w/o static	w/ static	w/o static	w/ static	w/o static	w/ static	w/o static
LR	52.9±1.2	**51.7±1.2**	36.9±1.1	34.0±1.1	33.0±0.7	32.2±0.6	39.7±0.6	36.9±0.6
LGBM	**53.7±1.2**	51.1±1.1	**40.6±0.8**	**38.9±1.5**	**36.0±0.6**	34.1±0.7	**44.2±0.7**	**41.0±0.7**
GRU	53.1±1.5	**52.9±1.2**	37.6±1.2	37.3±1.1	**36.1±0.9**	**35.3±0.8**	42.4±0.6	**41.5±0.7**
LSTM	**53.6±1.4**	51.0±1.0	37.8±1.0	36.2±1.4	35.7±0.8	**34.6±0.7**	41.0±0.7	40.2±0.8
TCN	**54.2±1.4**	**52.7±1.0**	39.2±1.3	**37.5±1.5**	34.3±0.6	**35.4±0.8**	41.4±0.8	**40.8±0.7**
TF	**54.4±1.1**	**52.7±0.9**	39.3±1.5	**38.4±1.5**	34.7±0.8	34.5±0.7	42.2±0.3	**41.3±0.8**

Table C.4: Baseline AUROC (↑), AUPRC (↑) performance of the included ML algorithms on the demo cohorts.

	eICU Demo		MIMIC-III Demo	
Algorithm	AUROC	AUPRC	AUROC	AUPRC
ICU Mortality				
LR	67.7±0.9	15.4±1.3	52.9±3.2	**37.6±2.3**
LGBM	**72.9±1.1**	**21.4±1.7**	**59.0±2.1**	33.0±2.4
GRU	71.5±1.3	**20.8±1.5**	50.5±3.6	33.9±3.5
LSTM	**71.3±1.6**	18.9±1.9	55.5±2.3	34.2±3.3
TCN	69.2±1.7	18.0±1.5	55.1±4.0	**38.8±3.6**
TF	71.6±1.3	18.7±1.7	53.3±3.5	**36.8±3.1**
AKI				
LR	61.3±0.6	16.8±0.4	52.9±2.3	16.8±2.0
LGBM	**72.6±0.3**	**23.8±0.4**	**60.7±1.7**	**22.4±1.8**
GRU	63.0±0.8	17.3±0.6	50.5±3.3	17.6±1.5
LSTM	61.9±0.9	16.2±0.6	53.5±2.5	15.4±1.3
TCN	64.5±1.0	17.6±0.6	53.7±2.8	19.6±2.0
TF	70.1±0.5	21.7±0.7	53.9±2.0	15.3±1.0
Sepsis				
LR	63.8±0.9	3.7±0.3	*	*
LGBM	53.5±1.5	2.8±0.2	*	*
GRU	64.7±1.3	4.1±0.3	*	*
LSTM	65.3±1.3	4.3±0.5	*	*
TCN	66.4±1.1	4.2±0.3	*	*
TF	**68.4±1.1**	**5.8±0.6**	*	*

* Our sepsis definition resulted in just one sepsis case for the MIMIC-III demo dataset. As a result, we could not use the 5-fold cross-validation approach to train a model reliably.

Demo datasets We offer out-of-the-box (i.e., executable straight after downloading the repository) experiment definitions with five tasks defined on two demo datasets: MIMIC-III demo and eICU demo. The results can be seen in Table C.4 and C.5. The traditional ml models perform better, most likely explained by the low number of samples. The kidney function task highlights the large difference in performance especially.

External validation (extended) The Length of Stay (LoS) results for ICU mortality prediction can be found in Figure C.1.

Table C.5: Baseline performance on the regression tasks. Results are reported in Mean Absolute Error (↓).

	Kidney Function		Length of Stay	
	eICU Demo	MIMIC-III Demo	eICU Demo	MIMIC-III Demo
EN	**0.30±0.00**	**0.33±0.03**	38.5±0.2	52.1±1.3
LGBM	0.87±0.01	0.86±0.04	**37.7±0.2**	**50.4±1.1**
GRU	0.54±0.02	3.23±0.19	39.7±0.6	54.7±1.0
LSTM	0.51±0.02	2.95±0.18	39.1±0.5	54.8±1.2
TCN	0.46±0.02	2.98±0.19	38.4±0.6	56.8±1.2
TF	0.60±0.03	3.22±0.18	38.9±1.2	57.1±1.2

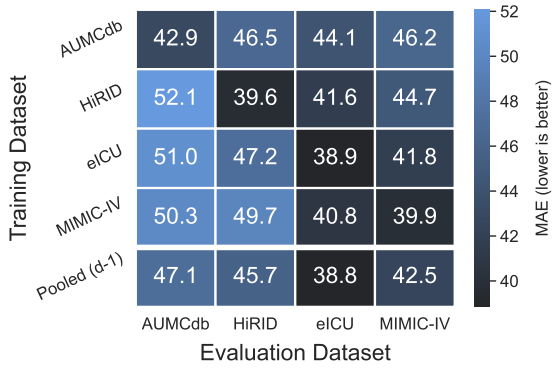


Figure C.1: Performance in Mean Absolute Error (MAE) of the Transformer model on *Length of Stay (LoS)*.

Fine-tuning (extended) The AUPRC results for our experiment with transfer learning can be found below and show a similar trend to Figure 3.8.

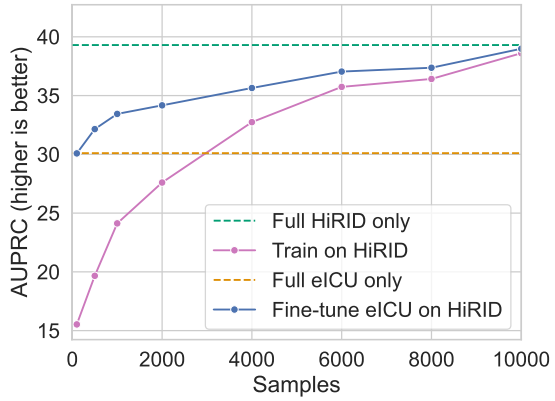


Figure C.2: AUPRC for fine-tuning an eICU GRU model for ICU mortality prediction on HiRID.

C.2 Datasets

This Appendix contains detailed description of the datasets and the preprocessing methodology.

C.2.1 Exclusion Criteria

We included all available ICU stays of adult patients in our analysis. For each stay, we applied the following exclusion criteria to ensure sufficient data volume and quality: remove any stays with **1)** an invalid admission or discharge time defined as a missing value or negative calculated length of stay, **2)** less than six hours spent in the ICU, **3)** less than four separate hours across the entire stay where at least one feature was measured, **4)** any time interval of ≥12 consecutive hours throughout the stay during which no feature was measured. Figure C.3 details the number of stays overall and by dataset excluded this way.

Additional exclusion criteria were applied based on the individual tasks; the details can be found schematically in Figure C.4 and C.5. For ICU mortality, we excluded all patients with a length of stay of fewer than 30 hours (either due to death or discharge). A minimum length of 30 hours was chosen to exclude any patients that were about to die (the sickest patients) or be discharged (the healthiest patients) at the time of prediction at 24 hours. For creatinine (Kidney Function), we excluded all patients with a length of stay of fewer than 48 hours or without a creatinine measurement between 24 and 48 hours (which was the outcome of interest). For AKI and sepsis, we excluded any stays where disease onset was outside the ICU or within the first six hours of the ICU stay. To account for differences in data recording across hospitals in eICU, we further excluded hospitals that did not have a single patient with AKI or sepsis to exclude hospitals with an insufficient recording of features necessary to define the outcome. Finally, for the AKI task, we excluded stays where the baseline creatinine, defined as the last creatinine measurement prior to ICU (if exists) or the earliest measurement in the ICU, was >4 mg/dL to exclude patients with preexisting renal insufficiency. For a numerical overview, please consult Table 4.10.

C.2.2 Preprocessing

A total of 52 features were used for model training (Table C.6), 4 of which were static and 48 that were dynamic. These features were selected as they are available across all datasets for most patients. Dynamic features primarily include vital signs (7 variables) and laboratory tests (39 variables), with two more variables that measure input (fraction of inspired oxygen) and output (urine). All variables were extracted via the `ricu` R package (version 0.5.3). The `ricu` name for each package is shown in Table C.6. The exact definition for each feature and how it was extracted from the individual databases can be found in the concept configuration file of the package's GitHub repository (commit 885bd0c). We also provided cohort definition code for this work, which can be run in both R and Python, in a github repository.

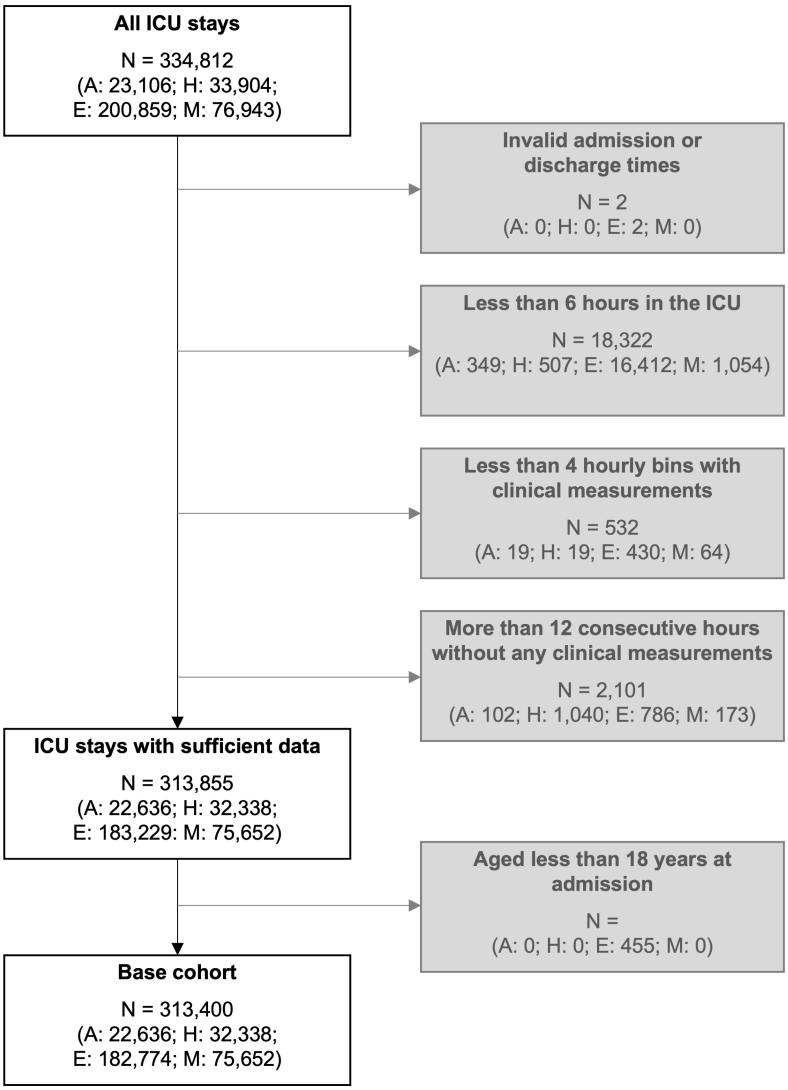


Figure C.3: Exclusion criteria applied to the base cohort. **N**: Total amount of cases. **A**: AUMCdb, **H**: HiRID, **E**: eICU, **M**: MIMIC-IV

Table C.6: Clinical concepts used as input to the prediction models.

Feature	`ricu`	*unit*
Static		
Age at hospital admission	`age`	*Years*
Female sex	`sex`	-
Patient height	`height`	*cm*
Patient weight	`weight`	*kg*
Time-varying		
Blood pressure (systolic)	`sbp`	*mmHg*
Blood pressure (diastolic)	`dbp`	*mmHg*

Table C.6: Clinical concepts used as input to the prediction models (continued)

Feature	`ricu`	*unit*
Heart rate	`hr`	*beats/minute*
Mean arterial pressure	`map`	*mmHg*
Oxygen saturation	`o2sat`	*%*
Respiratory rate	`resp`	*breaths/minute*
Temperature	`temp`	$°C$
Albumin	`alb`	*g/dL*
Alkaline phosphatase	`alp`	*IU/L*
Alanine aminotransferase	`alt`	*IU/L*
Aspartate aminotransferase	`ast`	*IU/L*
Base excess	`be`	*mmol/L*
Bicarbonate	`bicar`	*mmol/L*
Bilirubin (total)	`bili`	*mg/dL*
Bilirubin (direct)	`bili_dir`	*mg/dL*
Band form neutrophils	`bnd`	*%*
Blood urea nitrogen	`bun`	*mg/dL*
Calcium	`ca`	*mg/dL*
Calcium ionized	`cai`	*mmol/L*
Creatinine	`crea`	*mg/dL*
Creatinine kinase	`ck`	*IU/L*
Creatinine kinase MB	`ckmb`	*ng/mL*
Chloride	`cl`	*mmol/L*
CO^2 partial pressure	`pco2`	*mmHg*
C-reactive protein	`crp`	*mg/L*
Fibrinogen	`fgn`	*mg/dL*
Glucose	`glu`	*mg/dL*
Haemoglobin	`hgb`	*g/dL*
International normalised ratio (INR)	`inr_pt`	-
Lactate	`lact`	*mmol/L*
Lymphocytes	`lymph`	*%*
Mean cell haemoglobin	`mch`	*pg*
Mean corpuscular haemoglobin concentration	`mchc`	*%*
Mean corpuscular volume	`mcv`	*fL*
Methaemoglobin	`methb`	*%*
Magnesium	`mg`	*mg/dL*
Neutrophils	`neut`	*%*
O^2 partial pressure	`po2`	*mmHg*
Partial thromboplastin time	`ptt`	*sec*
pH of blood	`ph`	-
Phosphate	`phos`	*mg/dL*
Platelets	`plt`	*1,000 / µL*
Potassium	`k`	*mmol/L*
Sodium	`na`	*mmol/L*
Troponin T	`tnt`	*ng/mL*
White blood cells	`wbc`	*1,000 / µL*
Fraction of inspired oxygen	`fio2`	*%*
Urine output	`urine`	*mL*

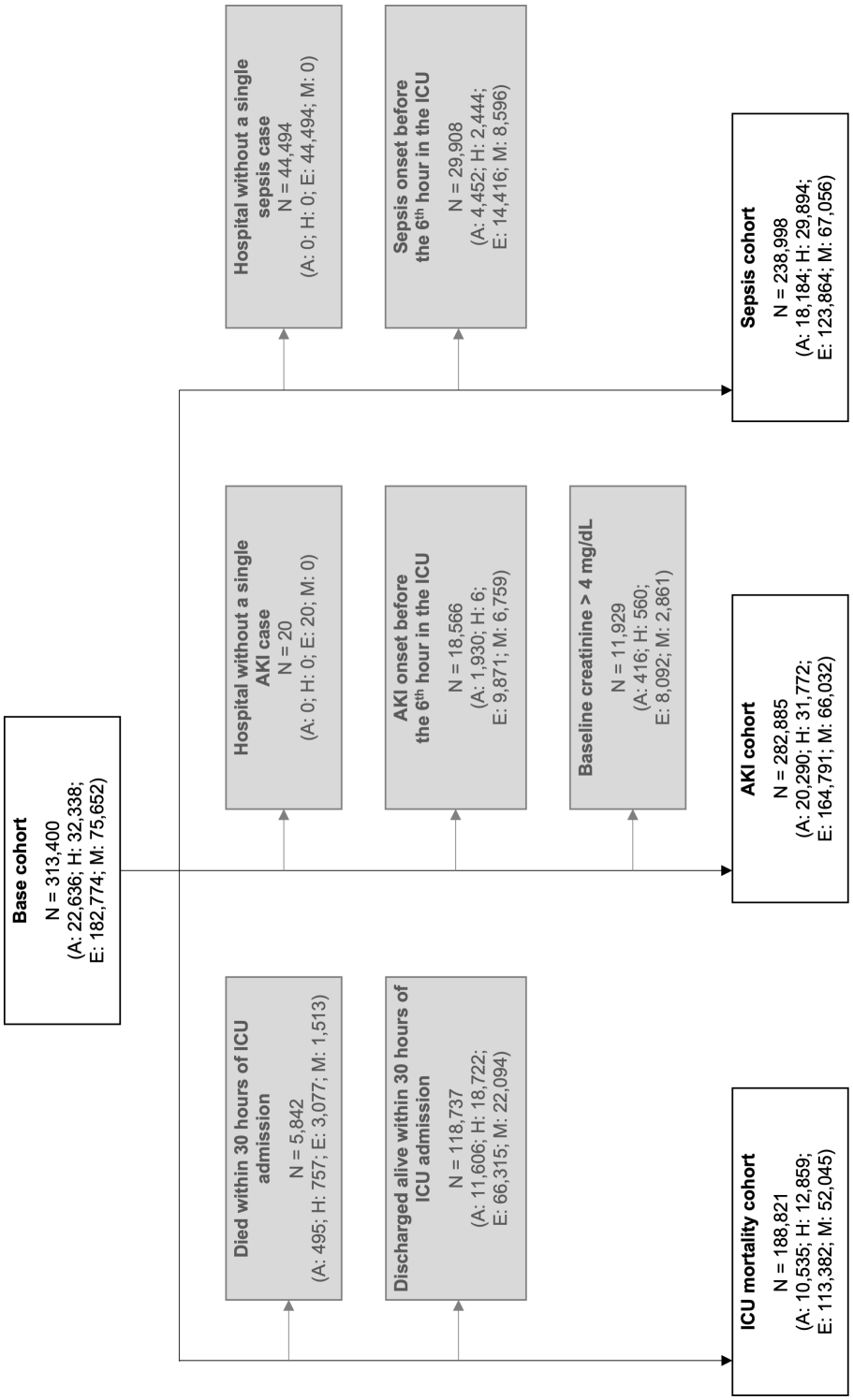


Figure C.4: Additional exclusion criteria applied for the classification tasks.

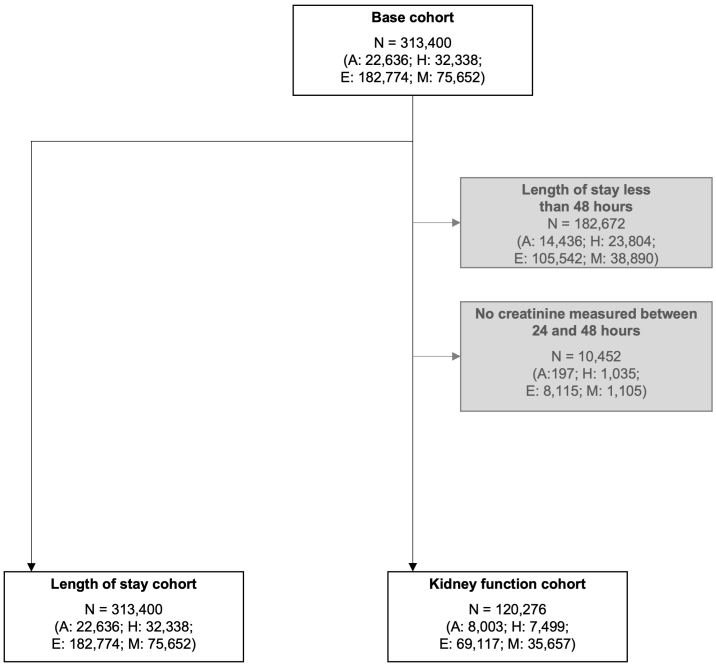


Figure C.5: Additional exclusion criteria applied for the regression tasks.

Additional features Furthermore, we consulted clinical experts to identify which features might be missing from our prediction setup. Several clinical features are currently missing from this setup, which could potentially improve prediction performance: *Glasgow coma scale score*, *Intubation*, *Ventilator settings*, *Renal replacement therapy*, *and Vasopressors*. We expect to be able to integrate more concepts as we collaborate with authors of datasets to make them available.

C.3 Outcome Definitions

The outcome definitions per task for each dataset are detailed in this Appendix.

C.3.1 ICU Mortality

ICU mortality was defined as death while in the ICU. This was generally ascertained via the recorded discharge status or discharge destination. Note that our definition of ICU mortality differs from the definition of `death` in the `ricu` R package, which describes hospital mortality that is unavailable for some included datasets.

AUMCdb Death was inferred from the `destination` column of the `admissions` table. A destination of "Overleden" (Dutch for "passed away") was treated as a death in the ICU. Since the date of death was recorded outside of the ICU and may therefore be imprecise, the recorded ICU discharge date was used as a more precise proxy for the time of death.

HiRID Death was inferred from the column `discharge_status` in table `general`. The status of "dead" was treated as a death in the ICU. Time of death was inferred as the last measurement of IDs 110 (mean arterial blood pressure) or 200 (heart rate) in column `variableid` of table `observations`.

eICU Death was inferred from the column `unitdischargestatus` in table `patient`. The status of "Expired" was treated as a death in the ICU. The recorded ICU discharge date was used as a proxy for the time of death.

MIMIC IV Death was inferred from the column `hospital_expire_flag` in table `admissions`. Since MIMIC IV only records a joint ICU/hospital expiration flag, ward transfers were analyzed to ascertain the location of death. If the last ward was the ICU, the death was considered ICU mortality.

Table C.7: Staging of AKI according to Kidney Disease Improving Global Outcomes (KDIGO) [KDI12]

Stage	Serum creatinine	Urine output
1	1.5–1.9 times baseline OR ≥0.3 mg/dl (≥26.5 μmol/l) increase within 48 hours	<0.5 ml/kg/h for 6–12 hours
2	2.0–2.9 times baseline	<0.5 ml/kg/h for ≥12 hours
3	3.0 times baseline (prior 7 days) OR Increase in serum creatinine to ≥4.0 mg/dl (≥353.6 μmol/l) within 48 hours OR Initiation of renal replacement therapy	<0.3 ml/kg/h for ≥24 hours OR Anuria for ≥12 hours

AKI, acute kidney injury; KDIGO, Kidney Disease Improving Global Outcomes.

C.3.2 Acute Kidney Injury

AKI was defined as KDIGO stage ≥1, either due to an increase in serum creatinine or low urine output (Table C.7) [KDI12]. Baseline creatinine was defined as the lowest creatinine measurement over the last 7 days. Urine rate was calculated as the amount of urine output in ml divided by the number of hours since the last urine output measurement (for a max gap of 24h), except for HiRID, in which urine rate was recorded

directly. The earliest urine output was divided by 1. The rate per kg was calculated based on the admission weight. If weight was missing, a weight of 75 kg was assumed instead.

AUMCdb Creatinine was defined via the standard `ricu` concept of serum creatinine as IDs 6836, 9941, or 14216 in column `itemid` of table `numericitems`. Urine output was defined as IDs 8794, 8796, 8798, 8800, 8803 in column `itemid` of table `numericitems` (note that this includes more items than those included in the standard `ricu` concept of urine output).

HiRID Creatinine was defined via the standard `ricu` concept of serum creatinine as ID 20000600 in column `variableid` of table `observations`. Urine rate was defined as ID 10020000 in column `variableid` of table `observations`.

eICU Creatinine was defined via the standard `ricu` concept of serum creatinine as IDs "creatinine" in column `labname` of table `lab`. Urine output was defined via the standard `ricu` concept of urine output as IDs "Urine" and "URINE CATHETER" in column `celllabel` of table `intakeoutput`.

MIMIC IV Creatinine was defined via the standard `ricu` concept of serum creatinine as ID 50912 in column `itemid` of table `labevents`. Urine output was defined via the standard `ricu` concept of urine output as IDs 226557, 226558, 226559, 226560, 226561, 226563, 226564, 226565, 226566, 226567, 226584, 227510 in column `itemid` of table `outputevents`.

C.3.3 Sepsis

The onset of sepsis was defined using the Sepsis-3 criteria [Sin+16], which defines sepsis as organ dysfunction due to infection. Following guidance from the original authors of Sepsis-3 [Sey+16], organ dysfunction was defined as an increase in SOFA score $\geq$2 points compared to the lowest value over the last 24 hours. Suspicion of infection was defined as the simultaneous use of antibiotics and culture of body fluids. The time of sepsis onset was defined as the first time of organ dysfunction within 48 hours before and 24 hours after suspicion of infection. Time of suspicion was defined as the earlier antibiotic initiation or culture request. Antibiotics and culture were considered concomitant if the culture was requested $\leq$24 hours after antibiotic initiation or if antibiotics were started $\leq$72 hours after the culture was sent to the lab. Where available, antibiotic treatment was inferred from administration records; otherwise, we used prescription data. To exclude prophylactic antibiotics, we required that antibiotics were administered continuously for $\geq$3 days [Rey+19]. Antibiotic treatment was considered continuous if an antibiotic was administered once every 24 hours for 3 days (or until death) or was prescribed for the entire time spent in the ICU. HiRID and eICU did not contain microbiological information. For these datasets, we followed [Moo+21a] and defined suspicion of infection through antibiotics alone. Note, however, that the sepsis prevalence in our study was considerably lower than theirs, which was as high as 37% in HiRID. We suspect this is because they did not require treatment for $\geq$3 days. For comparison and to contextualize the results of our experiments, we have benchmarked other sepsis definitions. See Table 3.9 for the number of stays and incidence for each cohort in this experiment performed on MIMIC-IV.

AUMCdb The SOFA score, microbiological cultures, and antibiotic treatment were defined via the standard `ricu` concepts `sofa`, `abx`, and `samp` (see the `ricu` package for more details).

HiRID The SOFA score and antibiotic treatment were defined via the standard `ricu` concept `sofa` and `abx` (see the `ricu` package for more details). No microbiology data were available in HiRID.

eICU The SOFA score and antibiotic treatment were defined via the standard `ricu` concept `sofa` and `abx` (see the `ricu` package for more details). Microbiology data in eICU was not reliable [Moo+21a] and therefore omitted.

MIMIC-IV The SOFA score and microbiological cultures were defined via the standard `ricu` concepts `sofa` and `samp` (see the `ricu` package for more details). Antibiotics were defined based on table `inputevents`. This differs from the standard `ricu` `abx` concept, which also considers the `prescriptions` table.

C.3.4 Kidney Function

The median creatinine level over the course of the second day of a patient's ICU stay was defined as the target of interest. Given the available datasets, this was chosen as the most suitable proxy of kidney function.

AUMCdb Creatinine was defined via the standard `ricu` concept of serum creatinine as IDs 6836, 9941, or 14216 in column `itemid` of table `numericitems`.

HiRID Creatinine was defined via the standard `ricu` concept of serum creatinine as ID 20000600 in column `variableid` of table `observations`.

eICU Creatinine was defined via the standard `ricu` concept of serum creatinine as IDs "creatinine" in column `labname` of table `lab`.

MIMIC IV Creatinine was defined via the standard `ricu` concept of serum creatinine as ID 50912 in column `itemid` of table `labevents`.

C.3.5 Remaining Length of Stay

At each hour, the remaining length of stay in the ICU was calculated in hours until discharge. A maximum forecasting window of 7 days was chosen, as forecasts beyond this interval were judged extremely difficult and of lesser clinical relevance. Arguably, an even shorter window of 3 days could be chosen to support short-term capacity planning. This is left for future investigation.

AUMCdb Length of stay was calculated via the standard `ricu` concept using the columns `admittedat` and `dischargedat` of table `admissions`.

HiRID Length of stay was calculated via the standard `ricu` concept using the column `admissiontime` of table `general` as well as the last observation in table `observations`.

eICU Length of stay was calculated via the standard `ricu` concept using the columsn `unitadmitoffset` and `unitdischargeoffset` of table `patient`.

MIMIC IV Length of stay was calculated via the standard `ricu` concept using the columsn `intime` and `outtime` of table `icustays`.

C.4 Usage of YAIB and Implementation

We describe the most important practical aspects of using *YAIB* in research in this Appendix. Please note that there is a *wiki*, dedicated to usage and development, available at: `https://github.com/rvandewater/YAIB/wiki`.

C.4.1 Use of Existing Code Repositories

As mentioned in the main text, we used parts of the HiRID-Benchmark code and heavily modified and extended the code to support our extra features and extensibility, mentioned in Table 3.1. HiRID-Benchmark is available at GitHub and makes use of an MIT license, as does our code repository.

C.4.2 Adding a Data Source

Adding a new dataset type, to use within *YAIB*, can be easily done by providing it in a `.gin` task definition file, see Listing C.1. Note, however, that any datasets formatted in the default way do not require any changes to be used by *YAIB*. By default, we have chosen to work with the Apache parquet [Voh16] file format, which is a modern, open-source column-oriented format that does not require a lot of storage due to efficient data compression[1]. We separate the data into three separate files: `DYNAMIC`, `STATIC`, and `OUTCOME`; this is defined for dynamic variables (that change during the stay), constant parameters, and the prediction task label respectively. Our cohort definition code produces the files in exactly this format. Furthermore, we see the concept of `roles` with the definition of the `vars` dictionary. These roles are assigned as defined in RecipIES , the preprocessing package developed alongside *YAIB*. The `GROUP` variable defines which internal dataset variable should be used to "group by" for, e.g., aggregating patient vital signs. The `SEQUENCE` variable defines the sequential dimension of the dataset (in the common case, this would be time). The other keys in this dictionary define the feature columns and outcome variables for prediction.

```python
@gin.configurable("base_classification_preprocessor")
class DefaultClassificationPreprocessor(Preprocessor):
    def __init__(self, generate_features: bool = True, scaling: bool = True, use_static_features: bool = True):
        """
        Args:
            generate_features: Generate features for dynamic data.
            scaling: Scaling of dynamic and static data.
            use_static_features: Use static features.
        Returns:
            Preprocessed data.
        """


    def apply(self, data, vars):
        """
        Args:
            data: Train, validation and test data dictionary. Further divided in static, dynamic, and outcome.
            vars: Variables for static, dynamic, outcome.
        Returns:
            Preprocessed data.
        """
        ...
        return data

    def _process_static(self, data, vars):
        ...
        return data
```

[1] `https://parquet.apache.org/`

```
    def _process_dynamic(self, data, vars):
        ...
        return data

    def _dynamic_feature_generation(self, data, dynamic_vars):
        ...
        return data
```

Listing C.1: Example preprocessing pipeline structure.

C.4.3 Creating a Preprocessing Pipeline

Our preprocessing pipeline is set up to be as general as possible and allows for custom implementations, defined as subclass from the `Preprocessor` class and passed as a command-line argument. For our tasks, we have defined a default preprocessing pipeline for both classification and regression tasks. Listing C.2 shows the class structure of the default classification preprocessor. In the private methods of this class, ReciPies is used to apply feature generation steps (which differ for ml and dl models). The abstract `Preprocessor` has two functions that need to be implemented: `__init__()` (which initializes the preprocessor and configures the settings) and `apply(data)` (which returns the preprocessed data dictionary of features and labels for each of the train, validate, and test splits)

C.4.4 Example: Training a Mortality Prediction LSTM Model

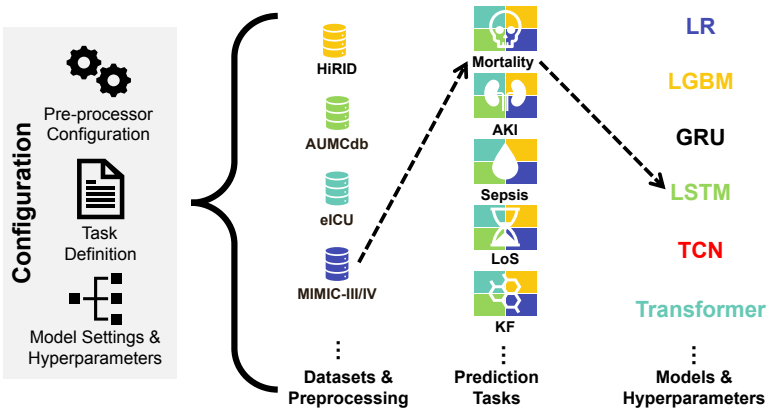


Figure C.6: Experiment definition schematic for YAIB. The fundamental experiment configuration of the benchmark contains three basic elements, **1)** the dataset, **2)** the prediction task, and **3)** the model and (list of) hyperparameters. Each element can be combined in different ways. Additionally, we provide an interface for extending each element (Datasets & Preprocessing, Prediction Tasks, Models & Hyperparameters) in the process. Provided is an example of an experiment configuration: predicting Sepsis on MIMIC [Tho+21] with an LSTM [HS97] model.

We demonstrate the complete process of training a Long Short-Term Memory (LSTM) model to predict sepsis on the MIMIC-III demo dataset with *YAIB*. This is shown schematically in Figure C.6. In Listing C.2, the basic task setup for mortality prediction after 24 hours is shown. We define the dataset files, by default split into 3 parquet files with the corresponding names. The listing describes three dataset components: dynamic data, outcome definitions, and static data. Below, one sees the variables and different *"roles"* assigned to concrete strings (see Appendix C.4.2 for detail). In this

```
# COMMON IMPORTS
include "configs/tasks/common/Imports.gin"

# MODE SETTINGS
Run.mode = "Classification"
NUM_CLASSES = 2 # Binary classification
HORIZON = 24
train_common.weight = "balanced"

# DEEP LEARNING
DLPredictionWrapper.loss = @cross_entropy
# DATASET AND PREPROCESSING
preprocess.file_names = {
    "DYNAMIC": "dyn.parquet",
    "OUTCOME": "outc.parquet",
    "STATIC": "sta.parquet",
}
vars = {
    "GROUP": "stay_id",
    "LABEL": "label",
    "SEQUENCE": "time",
    "DYNAMIC": ["alb", "alp", "alt", "ast", "be", "bicar", "bili", "bili_dir", "bnd", "bun", "ca", "cai", "ck", "ckmb", "cl",
        "crea", "crp", "dbp", "fgn", "fio2", "glu", "hgb", "hr", "inr_pt", "k", "lact", "lymph", "map", "mch", "mchc", "mcv",
        "methb", "mg", "na", "neut", "o2sat", "pco2", "ph", "phos", "plt", "po2", "ptt", "resp", "sbp", "temp", "tnt",
        "urine", "wbc"],
    "STATIC": ["age", "sex", "height", "weight"],
}

# SELECTING PREPROCESSOR
preprocess.preprocessor = @base_classification_preprocessor
preprocess.vars = %vars
preprocess.use_static = True

# SELECTING DATASET
PredictionDataset.vars = %vars

# CROSS VALIDATION
execute_repeated_cv.cv_repetitions = 5
execute_repeated_cv.cv_folds = 5
```

Listing C.2: Example configuration for a task definition configuration. We identify 4 main categories: the prediction type (classification or regression), the deep learning loss, the dataset variables, and preprocessing settings, and the cross-validation settings. Note that the dataset is fully configurable here.

listing, we also pass the vars to the preprocessing and dataset class. Finally, we see the definition of the cross-validation folds and iterations.

In Listing C.3 we see the configuration for the LSTM. We first define the generating features from dynamic data (relevant for traditional ml). Then we bind the LSTM model with a gin flag. After this, the hyperparameters are specified. The optimizer and encoder parameters are then specified. Note that we can specify ranges of hyperparameters to be tuned by the hyperparameter optimizer. Settings for this can be found in the bottom cluster of code. Listing C.4 shows how to train our LSTM model on the `mimic_demo` dataset (included in our repository).

```
import gin.torch.external_configurables
import icu_benchmarks.models.wrappers
import icu_benchmarks.models.encoders

default_preprocessor.generate_features = False

# Train params
train_common.model = @DLWrapper()

DLWrapper.encoder = @LSTMNet()
DLWrapper.optimizer_fn = @Adam
DLWrapper.train.epochs = 1000
DLWrapper.train.batch_size = 64
DLWrapper.train.patience = 10
DLWrapper.train.min_delta = 1e-4

# Optimizer params
optimizer/hyperparameter.class_to_tune = @Adam
optimizer/hyperparameter.weight_decay = 1e-6
optimizer/hyperparameter.lr = (1e-5, 3e-4)

# Encoder params
model/hyperparameter.class_to_tune = @LSTMNet
model/hyperparameter.num_classes = %NUM_CLASSES
model/hyperparameter.hidden_dim = (32, 256, "log-uniform", 2)
model/hyperparameter.layer_dim = (1, 3)

# Hyperparamter tuning
tune_hyperparameters.scopes = ["model", "optimizer"]
tune_hyperparameters.n_initial_points = 5
tune_hyperparameters.n_calls = 30
tune_hyperparameters.folds_to_tune_on = 2
```

Listing C.3: Example configuration for hyperparameters of LSTM. Tunable hyperparameter ranges can be floating points, integers, and categorical values.

```
#!
icu-benchmarks  train \
    -d demo_data/mortality24/mimic_demo \
    -n mimic_demo \
    -t BinaryClassification \
    -tn Mortality24 \
    --log-dir ../yaib_logs/ \
    -m LSTM \
    -gc \
    -lc \
    -s 2222 \
    -l ../yaib_logs/ \
    --tune
```

Listing C.4: Running YAIB. We train the LSTM model on the MIMIC-III demo dataset for the Mortality24 task.

C.4.5 Options

Table C.8: Options for the `train` command.

Flag	Required	Description
`-reproducible`	No	Make torch reproducible. (default: True)
`-hp, -hyperparams`	No	Hyperparameters for model.
`-tune`	No	Find best hyperparameters. (default: False)
`-checkpoint`	No	Use previous checkpoint.

We specify the command line options that YAIB provides for benchmarking prediction tasks. One can use the `icu-benchmarks` command with either `train` or `evaluate`. Table C.8 shows the arguments that can be used with `train`. Table C.9 contains the options to use with `evaluate` (i.e., for evaluation with a pre-trained model). In Table C.10, the general options for *YAIB* are shown.

Table C.9: Options for the `evaluate` command.

Flag	Required	Description
`-sn -source-name`	Yes	Name of the source dataset.
`-source-dir`	Yes	Directory containing gin and model weights.

Table C.10: General arguments for the use of *YAIB*.

Flag	Required	Description
`-d, -data-dir`	Yes	Path to the parquet data directory.
`-n, -name`	Yes	Name of the (target) dataset.
`-t, -task`	Yes	Name of the task gin.
`-m, -model`	No	Name of the model gin)Default.
`-tn, -task-name`	No	Name of the task, used for naming experiments.
`-e, -experiment`	No	Name of the experiment gin.
`-l, -log-dir`	No	Log directory with model weights.
`-s, -seed`	No	Random seed for processing, tuning, and training.
`-v, -verbose`	No	Whether to use verbose logging. (default: True)
`-cpu`	No	Set to use CPU. (default: False)
`-db, -debug`	No	Set to load less data. (default: False)
`-lc, -load_cache`	No	Set to load generated data cache. (default: False)
`-gc, -generate_cache`	No	Set to generate data cache. (default: False)
`-p, -preprocessor`	No	Load custom preprocessor from file.
`-pl, -plot`	No	Generate common plots. (default: False)
`-wd, -wandb-sweep`	No	Activates Weights and Biases hyper parameter sweep. (default: False)
`-imp, -pretrained-imputation`	No	Path to pre trained imputation model.

C.5 Extending YAIB

YAIB is built to adapt to your needs with as little effort as possible. *YAIB* allows you to change any part of your pipeline with ease: add a new dataset, define additional clinical concepts, adapt a task, implement imputation algorithms, use additional models, or evaluate custom metrics. This section provides examples on *YAIB* may be extended with respect to each of the above.

C.5.1 Add A New Dataset: Salzburg Intensive Care Database (SICdb)

SICdb is a recently published open source ICU dataset [Rod+23a]. The dataset includes 27,000 admissions to the ICU at University Hospital Salzburg (Austria) from 2013 to 2021. SICdb provides highly granular data with up to minute level resolution. To make SICdb available within *YAIB*, it must be interfaced to `ricu`. This ensures that `ricu` knows how to access and extract data from the dataset. Interfacing a new dataset requires three key steps:

1. Define the table structure and possible ID types as a JSON configuration.
2. Define how all ID systems and their origin times relate to each other.

```
{
   "name": "sic",
   "id_cfg": {
     "patient": {
       "id": "patientid",
       "position": 1,
       "start": "firstadmission",
       "end": "offsetofdeath",
       "table": "cases"
     },
     "icustay": {
       "id": "caseid",
       "position": 2,
       "start": "offsetafterfirstadmission",
       "end": "timeofstay",
       "table": "cases"
     }
   },
   "tables": {
     "cases": {
       "files": "cases.csv.gz",
       "defaults": {
         "index_var": "offsetafterfirstadmission",
         "time_vars": ["offsetafterfirstadmission", "offsetofdeath"]
       },
       "cols": {
         "caseid": {
           "name": "CaseID",
           "spec": "col_integer"
         },
         "patientid": {
           "name": "PatientID",
           "spec": "col_integer"
         },
         "admissionyear": {
           "name": "AdmissionYear",
           "spec": "col_integer"
         },
         ...
       }
     },
     "d_references": {
       ...
     },
     ...
   }
}
```

Listing C.5: ID and table configuration for the Salzburg Intensive Care Database Database in JSON.

Define ID types and table structure First, `ricu` needs to know what tables and columns exist within the data source. This is specified via the JSON configuration file partially shown in Listing C.5. Tables are defined under the `tables` element. The definition of each table contains the name of its source file, usually provided in `.csv` or `.csv.gz` format, and a list of all columns and their data types. In addition, default roles can be defined for certain columns in the table. These usually include the time index (if present), all other time columns, and a column that is considered to contain the value of interest.

In addition to the available tables, `ricu` also expects information about the main ID types used in the dataset. Each piece of information in ICU datasets is usually linked to a certain unit of observation, most commonly the patient (`patient`), the hospital admission (`hospadm`), or the specific ICU stay (`icustay`). By knowing what IDs a piece of information is measured for, `ricu` is able to temporally relate all information within the dataset. For example, `labevents` in MIMIC IV are recorded for hospital admissions, whereas `chartevents` are recorded for ICU stays. Defining how these two ID systems relate to each other allows them to be mapped to a common time scale (e.g., time since ICU admission). At the same time, knowing that an ICU stay is detailed in `icustays`

in MIMIC and in `cases` in SICdb allows to define the same semantic reference point in both databases.

The two ID types available in SICdb are the `patient` and the `icustay`. There is no separate demarcation of the `hospadm`. The observation time for a `patient` ranges from their first observed admission to their death (if it occurred). `icustays` range from the current admission to the ICU until the end of the stay.

Calculate origin times for each ID type After `ricu` has been told which ID systems exist in the dataset, it also needs to know when they start and end. Much of this process is automated. For SICdb, all origin times are already provided in a format suitable for use with `ricu`, and thus the default behavior is appropriate for them. However, minor adjustments are often necessary. For example, discharge times in SICdb are not provided as absolute times but in seconds since the start of admission. Such adjustments can be made on a case-by-case basis by subtyping the respective functions (in this case `id_win_helper`) and overwriting the default behavior (Listing C.6). Since SICdb provides time in seconds since admission, `ricu` must further be told to work with relative times in seconds. This can be conveniently achieved through existing helper functions (Listing C.6).

Following the steps above makes SICdb available and fully usable within `ricu`. While some further helper functions may be necessary to enable optional functionality such as automatic determination of measurement units (SICdb stores units in a separate reference table that needs to be merged at runtime), these are not essential to the main functionality of `ricu`. Note that the above only interfaces SICdb. It does *not* automatically map all existing clinical concepts for SICdb. Defining a clinical concept for SICdb still requires manual mapping of the concept to SICdb data items, for example via the appropriate measurement IDs (see also the next session on adding a clinical concept). We do not expect that this process can ever be fully avoided (unless it was already performed prior, for example by mapping to and providing the data in OMOP format). However, we found that the framework and helper functions that `ricu` provides greatly simplify this process. Additional information on `ricu` and its design principles can be found in [Ben+23].

C.5.2 Define a Clinical Concept: Potassium Chloride

YAIB comes with a wide range of pre-defined clinical concepts. However, it is likely that new tasks and applications require additional variables. An area of particular interest in this respect are medications. Using the exemplar of Potassium Chloride (a fluid frequently administered in the ICU), we demonstrate how new drugs can be added to *YAIB* with minimal effort. For many concepts, all that is required is a JSON dict that describes the correct measurement IDs within each dataset (Listing C.7). The JSON snippet can then be appended to the existing `ricu` concept file (for additions that should become part of the main package) or added to the search path of `load_dictionary` via the `cfg_dirs` parameter. The definition of complicated transformations such as calculation of hourly rates is also supported by existing helper functions. If custom calculations are necessary, they can be provided as user-defined functions via the `callback` element.

```r
id_win_helper.sic_env <- function(x) {
  # return a mapping between two ID systems (e.g., ICU stay ID and patient ID),
  # including relative start and end times
  cfg <- sort(as_id_cfg(x), decreasing = TRUE)

  ids <- field(cfg, "id")
  sta <- field(cfg, "start")
  end <- field(cfg, "end")

  tbl <- as_src_tbl(x, unique(field(cfg, "table")))

  mis <- setdiff(sta, colnames(tbl))

  res <- load_src(tbl, cols = c(ids, intersect(sta, colnames(tbl)), end))

  assert_that(length(mis) == 1L)
  res[, firstadmission := 0L]

  res <- res[, c(sta, end) := lapply(.SD, s_as_mins), .SDcols = c(sta, end)]
  res[, timeofstay := offsetafterfirstadmission + timeofstay] # convert to absolute discharge time

  res <- setcolorder(res, c(ids, sta, end))
  res <- rename_cols(res, c(ids, paste0(ids, "_start"),
                        paste0(ids, "_end")), by_ref = TRUE)

  as_id_tbl(res, ids[2L], by_ref = TRUE)
}

load_difftime.sic_tbl <- function(x, rows, cols = colnames(x),
                              id_hint = id_vars(x),
                              time_vars = ricu::time_vars(x), ...) {
  # Load time differences in SICdb by treating each time variable as
  # the relative time since admission in seconds
  load_as_relative_time(x, {{ rows }}, cols, id_hint, time_vars, s_as_mins)
}
```

Listing C.6: Helper functions that calculate the origin and temporal relation between events in the Salzburg Intensive Care Database.

C.5.3 Adapt a Task: KDIGO Stage Prediction

Cohorts created in *YAIB* are built to be easily adaptable. For example, predicting ordinal KDIGO stage rather than binary presence/absence of AKI is as simple as changing the outcome variable from `aki` to `kdigo`. More substantial changes to the task are also supported. If the general setup (i.e., covariates, exclusion criteria, etc.) remains the same but a novel outcome should be predicted, the outcome should be created as a clinical concept in `ricu` which can then simply be used as the outcome variable. If covariates, exclusion criteria, or prediction times change, existing helper functions can be utilized. See the existing task definitions for further examples.

C.5.4 Adding a New Prediction Model: Random Forest, RNN, and Temporal Fusion Transformer

We allow prediction models to be easily added and integrated into a Pytorch-Lightning (PL) [Ft23] module. This incorporates advanced logging and debugging capabilities, as well as built-in parallelism. Our interface derives from the PL `BaseModule`[2].

For standard Scikit-Learn type ML models (e.g., Light Gradient Boosting Machine (LGBM) [Ke+17]), one can implement the `MLWrapper`, incorporating the model steps in the process. Note that this class is also derived from the PL `BaseModule`; this leads to minimal code overhead. See Listing C.8 for details for the implementation of a random

[2] `https://lightning.ai/docs/pytorch/stable/common/lightning_module.html`

```
"kcl_dur": {
    "description": "potassium chloride duration",
    "category": "medications",
    "aggregate": "max",
    "sources": {
     "aumc": [
       {
         "ids": 9001,
         "table": "drugitems",
         "sub_var": "itemid",
         "stop_var": "stop",
         "grp_var": "orderid",
         "callback": "aumc_dur"
       }
     ],
     "eicu": [
       {
         "regex": "^potassium chloride",
         "table": "infusiondrug",
         "sub_var": "drugname",
         "callback": "eicu_duration(gap_length = hours(5L))",
         "class": "rgx_itm"
       }
     ],
     "hirid": [
       {
         "ids": 1000396,
         "table": "pharma",
         "sub_var": "pharmaid",
         "grp_var": "infusionid",
         "callback": "hirid_duration"
       }
     ],
     "miiv": [
       {
         "ids": [225166, 227522],
         "table": "inputevents",
         "sub_var": "itemid",
         "stop_var": "endtime",
         "grp_var": "linkorderid",
         "callback": "mimic_dur_inmv"
       }
     ]
    }
  }
```

Listing C.7: JSON definition for rate of potassium chloride across included datasets.

forest model. The only needed code here is the hyperparameter configuration and the initialization of the superclass.

```
@gin.configurable
class RFClassifier(MLWrapper):
    _supported_run_modes = [RunMode.classification]

    def __init__(self, *args, **kwargs):
        self.model = self.set_model_args(ensemble.RandomForestClassifier, *args, **kwargs)
        super().__init__(*args, **kwargs)
```

Listing C.8: Example ML model definition. We create a Random Forest classifier that implements the default *YAIB* prediction model interface

The definition of DL models can be done by creating a subclass from the `DLPredictionWrapper`; this inherits the standard methods needed for training DL learning models. Again, our implementation using Pytorch-Lightning significantly reduces the code overhead and complexity. See Listing C.9 for the example of a simple RNN model. We can then create a gin configuration file for this model such as that in Listing C.11 to specify default parameters and hyperparameter ranges for hp-tuning.

More advanced, or state-of-the-art, models are also easily implemented. One of *YAIB*'s users has implemented a Temporal Fusion Transformer architecture [Lim+21].

```python
@gin.configurable
class RNNet(DLPredictionWrapper):
    """Torch standard RNN model"""

    _supported_run_modes = [RunMode.classification, RunMode.regression]

    def __init__(self, input_size, hidden_dim, layer_dim, num_classes, *args, **kwargs):
        super().__init__(
            input_size=input_size, hidden_dim=hidden_dim, layer_dim=layer_dim, num_classes=num_classes, *args, **kwargs
        )
        self.hidden_dim = hidden_dim
        self.layer_dim = layer_dim
        self.rnn = nn.RNN(input_size[2], hidden_dim, layer_dim, batch_first=True)
        self.logit = nn.Linear(hidden_dim, num_classes)

    def init_hidden(self, x):
        h0 = x.new_zeros(self.layer_dim, x.size(0), self.hidden_dim)
        return h0

    def forward(self, x):
        h0 = self.init_hidden(x)
        out, hn = self.rnn(x, h0)
        pred = self.logit(out)
        return pred
```

Listing C.9: Example DL model definition. We use the inbuilt Torch RNN layers to built a Recurrent Neural Network.

This model provides good performance on multi-horizon time series forecasting, as well as interpretable insights. Specifically, [Lim+21] describe the TFT as follows: *"(1) examining the importance of each input variable in prediction, (2) visualizing persistent temporal patterns, and (3) identifying any regimes or events that lead to significant changes in temporal dynamics"*. See Listing C.10 for implementation details. This implementation is based upon the NVIDIA PyTorch implementation[3]. To train and optimize this model from a choice of hyperparameters, we need to specify a GIN file to bind the parameters, see Listing C.11. Note that we can use modifiers for the optimizer (e.g, Adam optimizer) and ranges that we can specify in rounded brackets "()". Square brackets, "[]", result in a random choice where the variable is uniformly sampled.

C.5.5 Adding a New Imputation Method: CSDI

We have added a range of imputation methods to *YAIB*, including interfaces to existing imputation libraries [Du23; Jar+22]. Here, we describe the addition of a recently introduced method that uses conditional score-based diffusion models conditioned on observed data: the CSDI[Tas+21]. To make the process of implementing these models easier, we have created the `ImputationWrapper` class that extends the pre-existing `DLWrapper` (itself a subclass of the `LightningModule` of Pytorch-lightning) with extra functionality.

The CSDI model is a diffusion model that follows the general architecture of conditional diffusion models[HJA20]; It introduces noise into a subset of time series data used as conditional observations to later denoise the data and predict accurate values for the imputation targets. CSDI is based on a U-Net architecture[RFB15] including residual connections.

[Tas+21] included two additional features into their model, which are inspired by DiffWave [Kon+21b]: an attention mechanism and the ability to input side information. The attention mechanism uses transformer layers, as shown in Figure C.7. An

[3] `https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/Forecasting/TFT`

```python
class TFT(DLPredictionWrapper):
    """
    Implementation of Temporal Fusion Transformer, https://arxiv.org/abs/1912.09363.
    """


    _supported_run_modes = [RunMode.classification, RunMode.regression]
    def __init__(
        self,
        num_classes, # Classes for multiclass classification
        encoder_length, # Determines interval to use for prediction
        hidden, # Amount of hidden layers
        dropout, # Dropout layers
        n_heads, # Attention heads
        dropout_att,
        example_length, # Determines interval to predict
        quantiles=[0.1, 0.5, 0.9], # quantiles to produce
        static_categorical_inp_size=[2], # Number of categories
        temporal_known_categorical_inp_size=[],
        temporal_observed_categorical_inp_size=[48], # Number of categorical observed variables
        static_continuous_inp_size=3, # Number of static continuous variables
        temporal_known_continuous_inp_size=0,
        temporal_observed_continuous_inp_size=48,
        temporal_target_size=1, # Number of target variables
        **kwargs,
    ):

        #derived variables
        num_static_vars = len(static_categorical_inp_size) + static_continuous_inp_size
        num_future_vars = len(temporal_known_categorical_inp_size) + temporal_known_continuous_inp_size
        num_historic_vars = sum([num_future_vars, temporal_observed_continuous_inp_size, temporal_target_size, len(
            temporal_observed_categorical_inp_size),])

        super().__init__(num_classes=num_classes, encoder_length=encoder_length, hidden=hidden,
                n_heads=n_heads, dropout_att=dropout_att, example_length=example_length, quantiles=quantiles, num_static_vars=
                    num_static_vars, num_future_vars=num_future_vars, num_historic_vars=num_historic_vars, *args,
                    static_categorical_inp_size=1, temporal_known_categorical_inp_size=0, temporal_observed_categorical_inp_size
                    =48, static_continuous_inp_size=3, temporal_known_continuous_inp_size=0, temporal_observed_continuous_inp_size
                    =48, temporal_target_size=1, **kwargs)



        self.encoder_length = encoder_length # Determines from how distant past we want to use data from

        self.embedding = LazyEmbedding(static_categorical_inp_size, temporal_known_categorical_inp_size,
            temporal_observed_categorical_inp_size, static_continuous_inp_size, temporal_known_continuous_inp_size,
            temporal_observed_continuous_inp_size, temporal_target_size,hidden) # embeddings for all variables

        self.static_encoder = StaticCovariateEncoder(num_static_vars, hidden, dropout) # encoding for static variables
        self.TFTback = TFTBack(encoder_length, num_historic_vars, hidden, dropout, num_future_vars, n_heads, dropout_att,
            example_length,quantiles)
        self.logit = nn.Linear(len(quantiles), num_classes) # Linear layer to output to the number of classes and allow modification
            by predictionwrapper.
```

Listing C.10: Temporal Fusion Transformer model definition. Note that this implementation is similar to existing code and can use existing methods and the original dataloader of *YAIB* which avoids code duplication.

input with K features, L length, and C channels is reshaped first to apply temporal attention and later reshaped again to apply feature attention. The second additional feature allows side information to be used as input to the model by a categorical feature embedding [Tas+21].

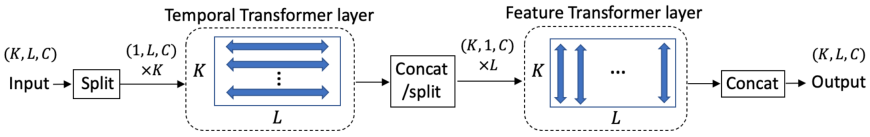


Figure C.7: The attention mechanism of CSDI adapted from [Tas+21].

See Listing C.12 for the most important implementation code: the model initialization. We note that of this code, very little has been adapted from the original code repository[4] included in the original publication [Tas+21].

[4] https://github.com/ermongroup/CSDI/tree/main

```
# Hyperparameters for TFT model.

# Common settings for DL models
include "configs/prediction_models/common/DLCommon.gin"

# Optimizer params
train_common.model = @TFT

optimizer/hyperparameter.class_to_tune = @Adam
optimizer/hyperparameter.weight_decay = 1e-6
optimizer/hyperparameter.lr = (1e-5, 3e-4)

# Encoder params
model/hyperparameter.class_to_tune = @TFT
model/hyperparameter.encoder_length = 24
model/hyperparameter.hidden = 256
model/hyperparameter.num_classes = %NUM_CLASSES
model/hyperparameter.dropout = (0.0, 0.4)
model/hyperparameter.dropout_att = (0.0, 0.4)
model/hyperparameter.n_heads =4
model/hyperparameter.example_length=25
```

Listing C.11: Temporal Fusion Transformer parameter configuration.

```
{
    def __init__(
        self, input_size, time_step_embedding_size, feature_embedding_size, unconditional, target_strategy, num_diffusion_steps,
              diffusion_step_embedding_dim, n_attention_heads, num_residual_layers, noise_schedule, beta_start, beta_end, n_samples,
               conv_channels, *args, **kwargs,
    ):
        super().__init__(...)
        self.target_dim = input_size[2]
        self.n_samples = n_samples

        self.emb_time_dim = time_step_embedding_size
        self.emb_feature_dim = feature_embedding_size
        self.is_unconditional = unconditional
        self.target_strategy = target_strategy

        self.emb_total_dim = self.emb_time_dim + self.emb_feature_dim
        if not self.is_unconditional:
            self.emb_total_dim += 1 # for conditional mask
        self.embed_layer = nn.Embedding(num_embeddings=self.target_dim, embedding_dim=self.emb_feature_dim)

        input_dim = 1 if self.is_unconditional else 2
        self.diffmodel = diff_CSDI(
            conv_channels,
            num_diffusion_steps,
            diffusion_step_embedding_dim,
            self.emb_total_dim,
            n_attention_heads,
            num_residual_layers,
            input_dim,
        )

        # parameters for diffusion models
        self.num_steps = num_diffusion_steps
        if noise_schedule == "quad":
            self.beta = np.linspace(beta_start**0.5, beta_end**0.5, self.num_steps) ** 2
        elif noise_schedule == "linear":
            self.beta = np.linspace(beta_start, beta_end, self.num_steps)

        self.alpha_hat = 1 - self.beta
        self.alpha = np.cumprod(self.alpha_hat)
        self.alpha_torch = torch.tensor(self.alpha).float().unsqueeze(1).unsqueeze(1)

}
```

Listing C.12: Implementing the CSDI architecture in YAIB. Note that our implementation is very similar to the original github repository, which demonstrates the flexibility of implementing new models in YAIB.

C.5.6 Adding an Evaluation Metric: Jensen Shannon Divergence and Binary Fairness

We support adding multiple types of evaluation metrics for benchmarking DL or ML models. We additionally support three common metric libraries: TorchMetrics [Nic+22], Ignite [Fom+20], and Scikit-Learn [Ped+11a]. Adding a metric is a straightforward pro-

cedure. We added the Jensen Shannon Divergence (JSD) with the help of the SciPy library [Vir+20]. See Listing C.13 for details.

```python
class JSD(EpochMetric):
    def __init__(self, output_transform: Callable = lambda x: x, check_compute_fn: bool = False) -> None:
        super(JSD, self).__init__(lambda x, y: JSD_fn(x, y), output_transform=output_transform, check_compute_fn=check_compute_fn)

        def JSD_fn(y_preds: torch.Tensor, y_targets: torch.Tensor):
            return jensenshannon(abs(y_preds).flatten(), abs(y_targets).flatten()) ** 2
```

Listing C.13: Implementing JSD using SciPy in YAIB. In this case we used the Ignite interface, but users can also choose to extend from the TorchMetric or SK-Learn interface.

One can then add the metric to be evaluated for a particular, see Listing C.15.

In order to asses fairness within ML prediction, a common metric to check is group fairness. This is computed through the ratio between positivity rates and true positives rates for different groups. Two types of these metrics are demographic parity [CKP09] and equal opportunity ratio [Har+16]. The TorchMetrics [Nic+22] library includes the group fairness module interface which we can adapt for use in *YAIB*. We use a wrapper, that extends the TorchMetrics implementation and extracts a "group tensor" that indicates to which group the sample belongs. See Listing C.14 to achieve the desired result. After this, we pass the data and feature names in the training step, as the function

```python
class BinaryFairnessWrapper(BinaryFairness):
    """
        This class is a wrapper for the BinaryFairness metric from TorchMetrics.
    """
    group_name = None
    def __init__(self, group_name = "sex", *args, **kwargs) -> None:
        self.group_name = group_name
        super().__init__(*args, **kwargs)
    def update(self, preds, target, data, feature_names) -> None:
        """" Standard metric update function"""
        groups = data[:, :, feature_names.index(self.group_name)]
        group_per_id = groups[:, 0]
        return super().update(preds=preds.cpu(),
                              target=target.cpu(),
                              groups=group_per_id.long().cpu())
```

Listing C.14: Adding a wrapper for Group fairness metric.

requires information about the assigned groups of the dataset; here, we track demographic parity by default, i.e., we use "sex" as the group name. In our pipeline, this has been one-hot encoded to 0 or 1. Then, adding this to the constants file, as seen in Listing C.15, automatically calculates this metric during the training process.

```python
class DLMetrics:
    BINARY_CLASSIFICATION = {
        "AUC": AUROC(task="binary"),
        "PR" : AveragePrecision(task="binary"),
        "F1": F1Score(task="binary", num_classes=2),
        "Calibration_Error": CalibrationError(task="binary",n_bins=10)
        "Calibration_Curve": CalibrationCurve,
        "PR_Curve": PrecisionRecallCurve,
        "RO_Curve": RocCurve,
        "JSD": JSD,
        "Binary_Fairness": BinaryFairnessWrapper(num_groups=2, task='demographic_parity', group_name="sex")
        }
```

Listing C.15: Metrics recorded for binary classification. Adding the metrics to this dictionary results in automatic logging. See the whole file for more details.

C.6 Experimental Setup and Reproducibility

To reproduce the results obtained in this paper, we have detailed our methodology in this Appendix.

C.6.1 Infrastructure and Hardware

We used a high-performance computing cluster to perform our experiments, which is owned by the Digital Health Cluster of the Hasso Plattner Institute in Potsdam. No data was transferred to any external parties in this process. This computing cluster is run with renewable energy and can be considered climate-neutral. The cluster is running the SLURM [YJG03] management tool on Ubuntu 20.04.6 LTS. with a number of Nvidia A100, A40, RTX 2080TI, and RTX Titan GPUs. The traditional machine learning algorithms were trained with Intel Xeon Platinum 8160 and AMD EPYC 7643 CPU resources.

Table C.11: Average estimated duration of training tasks with YAIB.

	Once per stay classification	Hourly classification	Hourly regression
Machine Learning Models	3 minutes	10 minutes	10 minutes
Deep Learning Models	30 minutes	60 minutes	90 minutes

C.6.2 Libraries

A full list of libraries is available in the *YAIB* repository; please use the Conda environment manager to install these. We have aimed to use the most recent library versions (while maintaining compatibility) to improve efficiency and reduce errors. At time of writing, the most recent version of Python that supports the used libraries was used: `Python 3.10`. The most important libraries are: `Gin-config 0.5.0`, `Pytorch 2.0` (in combination with Cuda 11.8.0), `Pytorch-lightning 2.0`, `Scikit-learn 1.2.2`, `Lightgbm 3.3.5`, `Pandas 2.0.0`.

C.6.3 Complexity

As mentioned in [Yèc+22], the transformer memory complexity concerning the sequence length is quadratic. With our hardware, training deep learning models overall takes less than 3 hours. Training ml models takes less than 1 hour. See Table C.11 for average training durations. Note that there was large variability between the cohorts due to the size difference of datasets (i.e., eICU Collaborative Research Database contains almost eight times the amount of stays as Amsterdam University Medical Center Database). The algorithmic complexity is specified in the implementation details of the Scikit-learn [Ped+11a] (logistic regression (LR) and elastic net (EN)), LightGBM [Ke+17] (Light Gradient Boosting Machine (LGBM)), and PyTorch library [Ft23] (Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU)). Furthermore, [BKK18] (Temporal Convolutional Network (TCN)) and

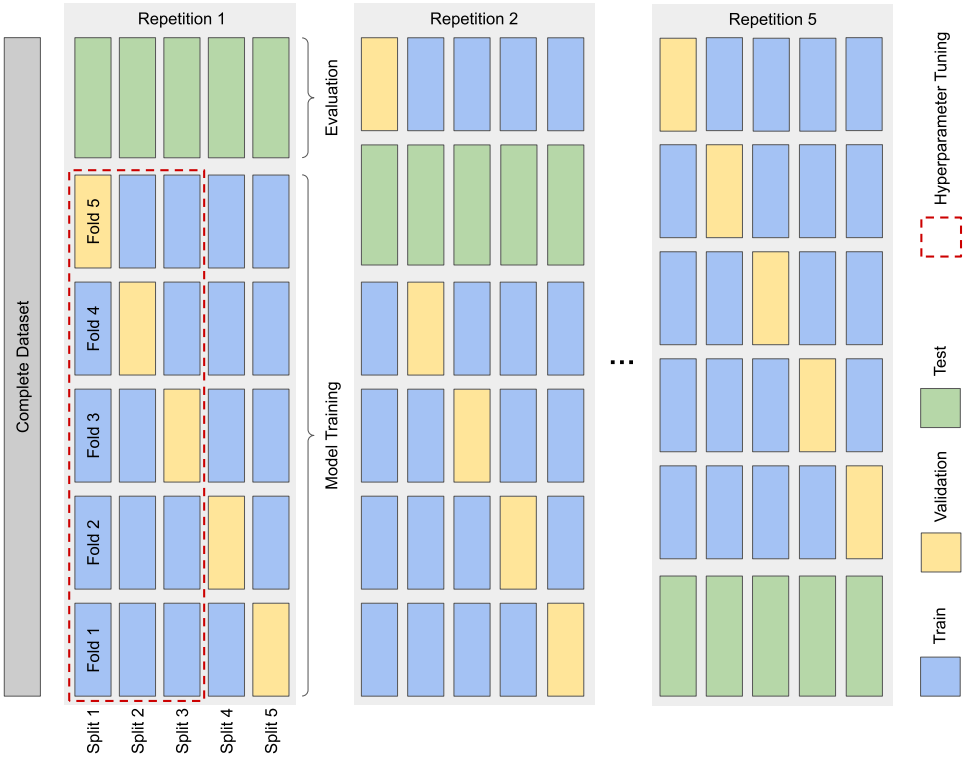


Figure C.8: Schematic overview of 5 times repeated 5-fold cross-validation. This example uses only the first three splits of the first repetition for hyperparameter tuning. The repetitions and amount of folds and the folds to tune on can be easily adjusted (see Listing C.3).

[Yèc+22] (transformer) are implementations are also detailed in their respective works. Figure C.8 shows our repeated cross-validation training method. Note that in order to obtain our final results, we do 5 repetitions of 5 fold cross-validation with an excluded test set.

C.6.4 Reproduciblity

For a detailed description to reproduce our experiments, we refer to the Paper reproducibility file (included in the repository). We have followed the standards specified by official ML reproducibility guidelines by Papers with Code[5].

C.6.5 External validation

All external validation validation models have been trained with an 80/20 train/val split to use as much of the dataset as possible. The test splits are the same as used for the same dataset experiments (diagonal in Figure 3.6) For the external validation pooled results (d-1), we used subsets of 10,000 stays for each dataset to simulate a setting where the datasets have a similar sample size. We wanted to ensure the size of the datasets, which differs significantly between datasets, had no undue influence on the training.

[5] `https://github.com/paperswithcode/releasing-research-code`

C.6.6 Fine-tuning

For our fine-tuning experiment, shown in Figure 3.8 we used an ADAM optimizer with a starting learning rate of 0.00001 and an exponential learning rate scheduler to reduce learning rates gradually. The rest of the hyperparameters are exactly as in the original source models.

C.7 Hyperparameters

Here, we detail the tuning setup and hyperparameters used in our experiments.

C.7.1 Tuning approach

In Table C.12 and C.13, we specify the hyperparameters used for hyperparameter-tuning for the baseline experiments for deep learning and machine learning models, respectively. We incorporated different sampling methods for hyperparameter selection. Hyperparameters were chosen to be mostly identical to [Yèc+22], to improve comparability and for reproducibility reasons. However, we have chosen to allow for continuous ranges of hyperparameters in some cases, to improve the performance and functionality of *YAIB*. Log-uniform means that the parameters are sampled according to the reciprocal distribution:

$$f(x;a,b) = \frac{1}{x[\log_e(b) - \log_e(a)]} \quad \text{for } a \leq x \leq b \text{ and } a > 0. \tag{C.1}$$

Uniform means that the parameters are sampled according to the uniform distribution:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \tag{C.2}$$

Table C.12: Model hyperparameters for deep learning models., default values for all models (above), and the distributions for the DL models (below), considered during Bayesian hyperparameter optimization.

Category	Parameter	Value
Hyperparameter search	Initial points	5
	Calls	30
	Folds to tune on	2
General Parameters	Epochs	1000
	Min delta	1E-4
	Patience	10
	Batch size	64
	Weight	Balanced
Loss	Regression task	Mean Squared Error
	Classification task	Cross-entropy
Optimizer Parameters	Optimizer	Adam
	Weight decay	1e-6
	Learning rate	Uniform([1E-5, 3E-4])
Model	**Parameter**	**Value**
LSTM	Hidden dimension	Log-uniform([32, 128])
	Hidden dimension	RandomInt(1, 3)
	Dropout probability	Uniform([0.0, 0.4])
GRU	Hidden dimension	Log-uniform([32, 128])
	Number of layers	RandomInt(1, 3)
TCN	Hidden dimension	Log-uniform([32, 128])
	Number of layers	Log-uniform([32, 256])
	Kernel size	Log-uniform([2, 32])
	Horizon	24
Transformer	Hidden dimension	Log-Uniform([32, 128])
	Number of layers	RandomInt(1, 10)
	Number of heads	Log-uniform(1, 8)
	Depth	Uniform([1,3])
	Kernel size	Log-uniform([2, 32])
	Dropout probability	Uniform([0.0, 0.4])
	Dropout attention	Uniform([0, 0.4])
	L1 regularization	0.0
	Hidden multiplication	2

Table C.13: Model hyperparameters for machine learning models., default values for all models (above), and distributions for the ML models (below), considered during Bayesian hyperparameter optimization.

Category	Parameter	Value
Hyperparameter search	Initial points	10
	Calls	50
	Folds to tune on	3
General Parameters	Patience	10
	Jobs	8
Loss	Regression task	Logloss
	Classification task	Cross-entropy
Model	**Parameter**	**Value**
LR	C	Log-uniform([1E-3, 1E1])
	Penalty	Choice(l1, l2, elasticnet)
	L1 Ratio	Uniform([0.0, 1.0])
	Solver	saga
	Max iterations	100000
EN	Alpha	Log-uniform([1E-2, 1E1])
	Tol	Log-uniform([1E-5, 1E-1])
	Hidden dimension	RandomInt(1, 3)
	Dropout probability	Uniform([0.0, 0.4])
	L1 Ratio	Uniform([0.0, 1.0])
	Solver	saga
	Max iterations	10000
LGBM	Column sample	Uniform([0.33, 1.0])
	Sub sample	Uniform([0.33, 1.0])
	Leaves	Log-uniform([8, 128])
	Max depth	RandomInt(3, 7)
	Hidden dimension	RandomInt(1, 3)
	Estimators	10000
	Min child samples	1000
	Subsample frequency	1

C.7.2 Deep Learning Models

We detail the hyperparameters that have been chosen using our Bayesian hyperparameter optimization approach.

Gated Recurrent Unit (GRU) The range of hyperparameters considered for the GRU model are found in Table C.14.

Table C.14: Chosen hyperparameters for Gated Recurrent Unit (GRU).

Dataset	Learning Rate	Layer Dimension	Hidden Dimension
Mortality			
AUMC	3.00E-04	3	48
HiRID	2.37E-04	2	52
eICU	3.00E-04	1	135
MIMIC-IV	1.43E-04	2	77
AKI			
AUMC	2.81E-04	3	256
HiRID	2.06E-04	3	115
eICU	1.43E-04	3	240
MIMIC-IV	2.82E-04	3	139
Sepsis			
AUMC	2.28E-04	2	77
HiRID	3.00E-04	3	59
eICU	8.51E-05	2	77
MIMIC-IV	2.39E-04	3	52
KF			
AUMC	3.00E-04	3	93
HiRID	1.11E-04	3	196
eICU	6.35E-05	3	196
MIMIC-IV	1.11E-04	1	148
LoS			
AUMC	5.45E-05	3	158
HiRID	1.00E-05	1	117
eICU	1.03E-05	3	254
MIMIC-IV	1.57E-05	2	237

Long Short-Term Memory (LSTM) The range of hyperparameters considered for the LSTM model are found in Table C.15.

Table C.15: Chosen hyperparameters for Long Short-Term Memory (LSTM).

Dataset	Learning Rate	Layer Dimension	Hidden Dimension
Mortality			
AUMC	1.87E-04	1	145
HiRID	1.54E-04	2	256
eICU	3.00E-04	3	149
MIMIC-IV	3.00E-04	2	185
AKI			
AUMC	2.62E-04	3	57
HiRID	3.00E-04	3	54
eICU	3.00E-04	3	70
MIMIC-IV	3.00E-04	3	256
Sepsis			
AUMC	2.10E-04	1	153
HiRID	2.48E-04	1	139
eICU	1.12E-04	2	40
MIMIC-IV	2.46E-04	1	161
KF			
AUMC	2.79E-04	2	81
HiRID	3.00E-04	1	256
eICU	1.75E-04	3	33
MIMIC-IV	2.49E-04	1	256
LoS			
AUMC	3.24E-05	1	62
HiRID	6.65E-05	3	255
eICU	2.86E-05	3	215
MIMIC-IV	1.80E-05	3	253

Temporal Convolutional Network (TCN) The range of hyperparameters considered for the TCN model are found in Table C.16.

Table C.16: Chosen hyperparameters for Temporal Convolutional Network (TCN).

Dataset	Learning Rate	Dropout	Kernel	Number of Channels
Mortality				
AUMC	5.84E-05	2.71E-01	6	92
HiRID	2.14E-04	1.12E-01	6	80
eICU	2.35E-05	1.15E-02	23	100
MIMIC-IV	1.81E-05	3.50E-01	3	130
AKI				
AUMC	3.00E-04	4.00E-01	3	144
HiRID	2.56E-04	2.56E-01	12	168
eICU	3.00E-04	1.23E-01	3	81
MIMIC-IV	2.98E-04	7.61E-02	3	249
Sepsis				
AUMC	3.00E-04	0.00E+00	2	32
HiRID	3.00E-04	0.00E+00	2	256
eICU	1.93E-04	3.98E-01	2	61
MIMIC-IV	2.23E-04	1.06E-01	4	78
KF				
AUMC	2.56E-04	2.78E-01	6	169
HiRID	1.75E-04	2.33E-01	3	34
eICU	2.15E-04	1.92E-01	5	138
MIMIC-IV	1.88E-05	2.15E-01	2	33
LoS				
AUMC	3.00E-04	1.92E-01	2	32
HiRID	2.74E-04	1.71E-01	29	43
eICU	3.00E-04	2.91E-01	10	32
MIMIC-IV	1.57E-04	1.67E-01	12	44

Transformer The range of hyperparameters considered for the transformer model are found in Table C.17.

Table C.17: Chosen hyperparameters for Transformer.

Dataset	Learning Rate	Dropout	Heads	Hidden Dimension	Depth
Mortality					
AUMC	1.29E-04	1.32E-01	2	95	2
HiRID	1.58E-04	0.00E+00	1	247	1
eICU	1.00E-05	4.00E-01	3	256	1
MIMIC-IV	6.18E-05	1.65E-01	1	48	3
AKI					
AUMC	1.18E-04	9.75E-03	8	52	3
HiRID	3.00E-04	1.50E-01	1	154	2
eICU	1.28E-04	1.33E-01	2	96	2
MIMIC-IV	1.22E-04	4.13E-02	2	72	3
Sepsis					
AUMC	2.61E-04	4.39E-02	1	32	3
HiRID	3.00E-04	0.00E+00	1	32	1
eICU	2.76E-05	1.05E-02	2	211	2
MIMIC-IV	3.53E-05	3.67E-01	1	98	3
KF					
AUMC	1.95E-04	4.56E-02	1	51	2
HiRID	2.48E-04	9.34E-02	7	160	3
eICU	2.62E-04	2.82E-02	1	52	1
MIMIC-IV	1.53E-04	8.62E-02	5	160	2
LoS					
AUMC	1.13E-04	4.11E-02	3	76	2
HiRID	1.88E-05	1.88E-05	4	102	1
eICU	3.96E-05	5.67E-02	3	172	2
MIMIC-IV	1.13E-04	4.11E-02	3	76	2

C.7.3 Machine Learning Models

Logistic regression (LR) The range of hyperparameters considered for the LR model are found in Table C.18.

Table C.18: Chosen hyperparameters for logistic regression (LR).

Dataset	C	Penalty	L1 Ratio
Mortality			
AUMC	3.63E-02	elasticnet	1.00E+00
HiRID	3.45E-02	l2	6.63E-02
eICU	2.78E-02	elasticnet	1.00E+00
MIMIC-IV	2.05E-01	elasticnet	1.00E+00
AKI			
AUMC	1.77E-02	l1	1.00E+00
HiRID	1.00E+01	l1	4.16E-01
eICU	2.52E-02	l1	5.99E-01
MIMIC-IV	1.28E-01	l1	2.98E-01
Sepsis			
AUMC	4.87E-02	l1	2.21E-01
HiRID	2.59E-03	l2	3.74E-01
eICU	1.98E-03	elasticnet	6.20E-01
MIMIC-IV	2.20E-03	l1	1.81E-02

Elastic net (EN) The range of hyperparameters considered for the EN model are found in Table C.19.

Table C.19: Chosen hyperparameters for elastic net (EN).

Dataset	Alpha	Tol	L1 Ratio
KF			
AUMC	1.04E-02	3.60E-02	1.22E-03
HiRID	1.04E-02	3.60E-02	1.22E-03
eICU	1.05E-02	9.57E-04	8.09E-03
MIMIC-IV	1.00E-02	2.45E-03	0.00E+00
LoS			
AUMC	1.00E-02	1.67E-05	7.66E-02
HiRID	1.00E-02	1.02E-05	2.40E-02
eICU	1.00E-02	4.82E-02	0.00E+00
MIMIC-IV	1.00E-02	7.42E-02	0.00E+00

Light Gradient Boosting Machine (LGBM) The range of hyperparameters considered for the LGBM model are found in Table C.20.

Table C.20: Chosen hyperparameters for Light Gradient Boosting Machine (LGBM).

Dataset	Depth	Column Sample	Leaves	Subsample
Mortality				
AUMC	6	9.89E-01	117	9.90E-01
HiRID	5	1.00E+00	8	1.00E+00
eICU	7	5.41E-01	128	1.00E+00
MIMIC-IV	7	1.00E+00	28	1.00E+00
AKI				
AUMC	7	1.00E+00	110	8.69E-01
HiRID	7	1.00E+00	79	8.69E-01
eICU	7	9.97E-01	117	8.35E-01
MIMIC-IV	7	1.00E+00	128	1.00E+00
Sepsis				
AUMC	7	1.00E+00	128	7.83E-01
HiRID	4	6.18E-01	17	8.22E-01
eICU	7	8.97E-01	52	6.04E-01
MIMIC-IV	7	1.00E+00	128	5.84E-01
KF				
AUMC	7	7.01E-01	15	9.99E-01
HiRID	3	9.58E-01	16	9.95E-01
eICU	7	1.00E+00	47	1.00E+00
MIMIC-IV	7	9.30E-01	33	9.96E-01
LoS				
AUMC	7	3.30E-01	83	1.00E+00
HiRID	6	3.30E-01	49	5.01E-01
eICU	7	3.30E-01	51	3.30E-01
MIMIC-IV	7	3.30E-01	128	3.30E-01

C.8 Machine Learning Reproducibility Checklist

This checklist (version 2.0) was found to be the most recent version at the time of writing. It can be found at: `www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist-v2.0.pdf`. For all models and algorithms presented, check if you include:

- A clear description of the mathematical setting, algorithm, and/or model. [Yes] *See section C.6 for an overview of the used infrastructure, libraries, complexity, reproducibility. See section C.7 for the hyperparameter tuning setup.*
- A clear explanation of any assumptions. [Yes] *We make limited assumptions about the data we use. Mainly, we assume stays are i.i.d. as is usual in clinical machine learning.*
- An analysis of the complexity (time, space, sample size) of any algorithm. [Yes] *See section C.6.*

For any theoretical claim, check if you include:

- A clear statement of the claim. [N/A] *We did not make theoretical claims.*
- A complete proof of the claim.[N/A]

For all datasets used, check if you include:

- The relevant statistics, such as number of examples. [Yes] *See section C.2 and section C.3 for details of the data.*
- The details of train / validation / test splits. [Yes] *See subsection 3.5.2.*
- An explanation of any data that were excluded, and all pre-processing step. [Yes] *See subsection 3.5.2, section C.2, and section C.3.*
- A link to a downloadable version of the dataset or simulation environment. [Yes] *See the included `yaib-cohorts` code in order to generate the datasets (additionally, the YAIB repository contains preprocessed demo datasets)*
- For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control. [N/A] *We did not collect new data.*

For all shared code related to this work, check if you include:

- Specification of dependencies. [Yes] *This can be found in section C.6 and at our GitHub repository (below).*
- Training code. [Yes] *This can be found in the main YAIB code (`cross-validation.py` and `train.py`. Details for the used parameters can be found in section C.6*
- Evaluation code. [Yes] *See answer above.*
- (Pre-)trained model(s). [Yes] *Pretrained models for each task are publicly available and can be provided upon request.*
- README file includes table of results accompanied by precise command to run to produce those results.[Yes] *We include the results in the main text and section C.1. Additionally, we have a reproducibility document (`PAPER.md`) in our repository that follows common guidelines.*

For all reported experimental results, check if you include:

- The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results.[Yes] *For the method see subsection 3.3.7, for the hyperparameter configuration, see: section C.7. For the individual hyperparameter configurations per model, dataset, and task, see subsection C.7.2 and subsection C.7.3 for ml and dl models, respectively.*
- The exact number of training and evaluation runs. [Yes] *See subsection 3.5.2.*

- A clear definition of the specific measure or statistics used to report results.[Yes] *See subsection 3.5.2.*
- A description of results with central tendency (e.g. mean) & variation (e.g. error bars). [Yes] *We report the standard deviation of all results, to accurately reflect the precision of our results.*
- The average runtime for each result, or estimated energy cost.[Yes] *See section C.6.*
- A description of the computing infrastructure used.[Yes] *See section C.6.*

Appendix D

Supplements Chapter 4

D.1 Retrospective Multicenter Prediction of Postoperative Complications

D.1.1 Inclusion Criteria

This retrospective cohort study enrolled patients of legal age undergoing major elective visceral surgery at two independent, high-volume, tertiary, university-affiliated surgical centers between January 2014 and December 2022 in either of the two participating medical centers. The study was registered at the DRKS (*Deutsches Register Klinischer Studien* or German Clinical Trials Register):`https://drks.de/search/de/trial/DRKS00033658`
Inclusion criteria categorized by surgical site were as follows:
Esophagus, 4.94 % of total cases:

- 5-423 Partial esophageal resection without restoration of continuity
- 5-424 Partial esophageal resection with restoration of continuity
- 5-425 (Total) esophagectomy without restoration of continuity
- 5-426 (Total) esophagectomy with restoration of continuity
- 5-427 Reconstruction of the esophageal passage (as an independent procedure)

Stomach 12.55 % of total cases:

- 5-434 Atypical partial gastric resection
- 5-435 Partial gastric resection (2/3 resection)
- 5-436 Subtotal gastric resection (4/5 resection)
- 5-437 (Total) gastrectomy
- 5-438 (Total) gastrectomy with esophageal resection

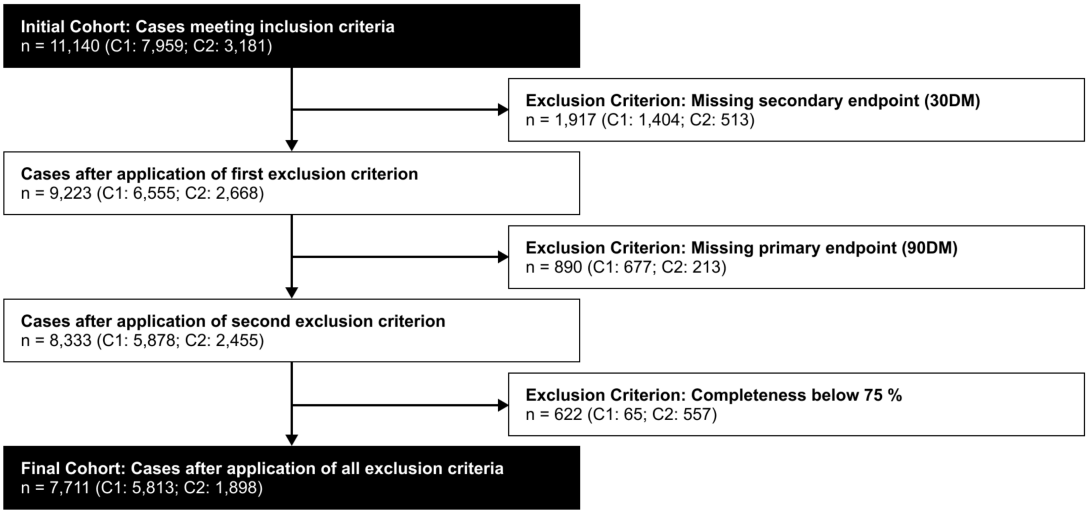


Figure D.1: Schematic representation of the inclusion and exclusion criteria. C1 = Center 1, C2 = Center 2, 30DM = 30-Day Mortality, 90DM = 90-Day Mortality. Used under CC BY-NC 4.0 license from the supplements of Riepe et al. [Rie+25b].

Intestine 48.49 % of total cases:

- 5-454 Resection of the small intestine
- 5-455 Partial resection of the colon
- 5-456 (Total) colectomy and proctocolectomy
- 5-465 Repositioning of a double-barreled enterostoma
- 5-466 Restoration of the continuity of the intestine with terminal enterostomata
- 5-484 Rectal resection with sphincter preservation
- 5-485 Rectal resection without sphincter preservation

Liver 21.96 % of total cases:

- 5-501 Local excision and destruction of diseased liver tissue (atypical liver resection)
- 5-502 Anatomical (typical) liver resection

Pancreas 12.06 % of total cases:

- 5-524 Partial resection of the pancreas
- 5-525 (Total) pancreatectomy

Table D.1: Eliminated features during preprocessing due to completeness below 50%. TSH = Thyroid Stimulating Hormone, AST = Aspartate Transaminase, ALT = Alanine Transaminase, HbA1c = Glycated Hemoglobin, pH = potential Hydrogen, CA19-9 = Carbohydrate Antigen 19-9, CA125 = Cancer Antigen 125, CEA = Carcinoembryonic Antigen. Used under CC BY-NC 4.0 license from the supplements of Riepe et al. [Rie+25b].

#	Feature	Completeness (%)
1	Albumin	36.07
2	Bilirubin (Direct)	2.39
3	Bilirubin (Indirect)	0.59
4	TSH	20.27
5	ALT	22.65
6	AST	24.09
7	Lactate	38.79
8	HbA1c	8.12
9	Prothrombin Time	32.04
10	pH	39.70
11	Base Excess	39.67
12	Bicarbonate	1.06
13	Alpha Fetoprotein	2.50
14	CA 19-9	6.58
15	CA 125	1.30
16	CAE	5.70

Table D.2: Statistical analysis of the 8 patient characteristics and statistical comparison of these between the two included surgical centers. Numerical features are given as an average with standard deviation. C1 = Center 1, C2 = Center 2, SD = Standard Deviation, BMI = Body Mass Index, ASA = American Society of Anesthesiologists, ECOG = Eastern Cooperative Oncology Group, rCCI = revised Charlson Comorbidity Index. Used under CC BY-NC 4.0 license from the supplements of Riepe et al. [Rie+25b].

#	Feature	Complete (%)	Combined (C1 + C2)	Internal (C1)	External (C2)	p
1	Gender	94.71				0.27
	Male	54.88% (n = 4232)	54.24% (n = 3153)	56.85% (n = 1079)		
	Female	39.83% (n = 3071)	38.74% (n = 2252)	43.15% (n = 819)		
2	Age (y)	100	57.93 (SD = 15.69)	58.82 (SD = 14.83)	55.21 (SD = 17.79)	<0.01
3	Height (cm)	85.25	172.91 (SD = 10.09)	172.91 (SD = 10.17)	172.89 (SD = 9.86)	0.39
4	Weight (kg)	85.44	79.85 (SD = 25.09)	81.97 (SD = 26.9)	73.52 (SD = 17.2)	<0.01
5	BMI (kg/m^2)	85.25	26.62 (SD = 7.71)	27.32 (SD = 8.3)	24.52 (SD = 5.04)	<0.01
6	ASA Score	88.33				<0.01
	1	4.92% (n = 379)	4.27% (n = 248)	6.90% (n = 131)		
	2	45.18% (n = 3484)	41.96% (n = 2439)	55.06% (n = 1045)		
	3	36.77% (n = 2835)	40.63% (n = 2362)	24.92% (n = 473)		
	4	1.45% (n = 112)	1.60% (n = 93)	1.00% (n = 19)		
	5	0.01% (n = 1)	0.02% (n = 1)	0.00% (n = 0)		
7	ECOG Score	68.03				<0.01
	0	52.26% (n = 4030)	44.52% (n = 2588)	75.97% (n = 1442)		
	1	14.75% (n = 1137)	19.01% (n = 1105)	1.69% (n = 32)		
	2	1.02% (n = 79)	1.29% (n = 75)	0.21% (n = 4)		
8	rCCI	100				<0.01
	0	13.94% (n = 1075)	10.27% (n = 597)	25.18% (n = 478)		
	1	5.76% (n = 444)	4.63% (n = 269)	9.22% (n = 175)		
	2	8.74% (n = 674)	8.27% (n = 481)	10.17% (n = 193)		
	3	9.66% (n = 745)	9.65% (n = 561)	9.69% (n = 184)		
	4	9.45% (n = 729)	10.10% (n = 587)	7.48% (n = 142)		
	5	8.66% (n = 668)	8.76% (n = 509)	8.38% (n = 159)		
	6	5.62% (n = 433)	5.78% (n = 336)	5.11% (n = 97)		
	7	3.53% (n = 272)	3.53% (n = 205)	3.53% (n = 67)		
	8	6.60% (n = 509)	7.41% (n = 431)	4.11% (n = 78)		
	9	7.86% (n = 606)	9.24% (n = 537)	3.64% (n = 69)		
	10	7.69% (n = 593)	8.64% (n = 502)	4.79% (n = 91)		
	11	6.89% (n = 531)	7.69% (n = 447)	4.43% (n = 84)		
	12	2.93% (n = 226)	3.23% (n = 188)	2.00% (n = 38)		
	13	1.56% (n = 120)	1.53% (n = 89)	1.63% (n = 31)		
	14	0.75% (n = 58)	0.84% (n = 49)	0.47% (n = 9)		
	15	0.26% (n = 20)	0.29% (n = 17)	0.16% (n = 3)		
	16	0.09% (n = 7)	0.12% (n = 7)	0.00% (n = 0)		
	17	0.01% (n = 1)	0.02% (n = 1)	0.00% (n = 0)		

Table D.3: Statistical analysis of the 37 pre-existing conditions and statistical comparison of these between the two included surgical centers. Comp. = Completeness (%), C1 = Center 1, C2 = Center 2, AIDS = Acquired Immunodeficiency Syndrome. Used under CC BY-NC 4.0 license from the supplements of Riepe et al. [Rie+25b].

#	Feature	Comp.	Combined (C1 + C2)	Internal (C1)	External (C2)	p
9	Myocardial Infarction	100	2.55% (n = 197)	2.08% (n = 121)	4.00% (n = 76)	<0.01
10	Congestive Heart Failure	100	2.35% (n = 181)	1.96% (n = 114)	3.53% (n = 67)	<0.01
11	Peripheral Vascular Disease	100	4.10% (n = 316)	3.99% (n = 232)	4.43% (n = 84)	0.45
12	Cerebrovascular Disease	100	0.93% (n = 72)	0.81% (n = 47)	1.32% (n = 25)	0.06
13	Dementia	100	0.41% (n = 32)	0.40% (n = 23)	0.47% (n = 9)	0.80
14	Chronic Pulmonary Disease	100	9.18% (n = 708)	9.15% (n = 532)	9.27% (n = 176)	0.91
15	Rheumatic Disease	100	0.82% (n = 63)	0.76% (n = 44)	1.00% (n = 19)	0.38
16	Peptic Ulcer Disease	100	1.01% (n = 78)	1.10% (n = 64)	0.74% (n = 14)	0.21
17	Liver Disease (Mild)	100	12.77% (n = 985)	14.59% (n = 848)	7.22% (n = 137)	<0.01
18	Liver Disease (Moderate to Severe)	100	1.19% (n = 92)	1.36% (n = 79)	0.68% (n = 13)	0.03
19	Diabetes (w/o chronic Complications)	100	18.04% (n = 1391)	21.35% (n = 1241)	7.90% (n = 150)	<0.01
20	Diabetes (w/chronic Complications)	100	1.98% (n = 153)	1.63% (n = 95)	3.06% (n = 58)	<0.01
21	Hemiplegia or Paraplegia	100	0.97% (n = 75)	0.77% (n = 45)	1.58% (n = 30)	<0.01
22	Renal Disease	100	9.45% (n = 729)	9.48% (n = 551)	9.38% (n = 178)	0.93
23	Malignancy	100	65.02% (n = 5014)	71.01% (n = 4128)	46.68% (n = 886)	<0.01
24	Metastatic Solid Tumor	100	32.46% (n = 2503)	36.73% (n = 2135)	19.39% (n = 368)	<0.01
25	AIDS	100	0.16% (n = 12)	0.14% (n = 8)	0.21% (n = 4)	0.71
26	Cardiac Arrhythmia	100	6.45% (n = 497)	5.87% (n = 341)	8.22% (n = 156)	<0.01
27	Valvular Disease	100	2.15% (n = 166)	1.57% (n = 91)	3.95% (n = 75)	<0.01
28	Pulmonary Circulatory Disorder	100	0.21% (n = 16)	0.17% (n = 10)	0.32% (n = 6)	0.36
29	Arterial Hypertension	100	42.85% (n = 3304)	45.6% (n = 2651)	34.4% (n = 653)	<0.01
30	Other Neurological Disorders	100	1.00% (n = 77)	0.88% (n = 51)	1.37% (n = 26)	0.08
31	Hypothyroidism	100	6.26% (n = 483)	6.24% (n = 363)	6.32% (n = 120)	0.95
32	Coagulopathy	100	3.13% (n = 241)	2.44% (n = 142)	5.22% (n = 99)	<0.01
33	Obesity	100	15.46% (n = 1192)	18.17% (n = 1056)	7.17% (n = 136)	<0.01
34	Weight Loss	100	1.35% (n = 104)	1.32% (n = 77)	1.42% (n = 27)	0.84
35	Fluid and Electrolyte Disorders	100	15.56% (n = 1200)	6.43% (n = 374)	43.52% (n = 826)	<0.01
36	Blood Loss Anemia	100	0.29% (n = 22)	0.34% (n = 20)	0.11% (n = 2)	0.15
37	Deficiency Anemia	100	1.30% (n = 100)	0.76% (n = 44)	2.95% (n = 56)	<0.01
38	Alcohol Abuse	100	1.97% (n = 152)	2.32% (n = 135)	0.90% (n = 17)	<0.01
39	Drug Abuse	100	0.83% (n = 64)	1.00% (n = 58)	0.32% (n = 6)	<0.01
40	Psychoses	100	0.18% (n = 14)	0.22% (n = 13)	0.05% (n = 1)	0.23
41	Depression	100	2.14% (n = 165)	1.67% (n = 97)	3.58% (n = 68)	<0.01
42	Coronary Heart Disease	100	7.43% (n = 573)	7.40% (n = 430)	7.53% (n = 143)	0.88
43	Chronic Pancreatitis	100	3.19% (n = 246)	3.82% (n = 222)	1.26% (n = 24)	<0.01
44	Ulcerative Colitis	100	6.69% (n = 516)	2.96% (n = 172)	18.12% (n = 344)	<0.01
45	Crohn's Disease	100	12.85% (n = 991)	4.28% (n = 249)	39.09% (n = 742)	<0.01

D.1.2 Data Description and Comparison of Centers

Table D.4: Statistical analysis of the 17 intervention characteristics and statistical comparison of these between the two included surgical centers. Numerical features are given as an average with standard deviation. C1 = Center 1, C2 = Center 2, SD = Standard Deviation. Used under CC BY-NC 4.0 license from the supplements of Riepe et al. [Rie+25b].

#	Feature	Complete (%)	Combined (C1 + C2)	Internal (C1)	External (C2)	
46	Primary System	100				
	Esophagus		4.94% (n = 381)	4.94% (n = 287)	4.95% (n = 94)	
	Stomach		12.55% (n = 968)	14.83% (n = 862)	5.58% (n = 106)	
	Intestine		48.49% (n = 3739)	39.26% (n = 2282)	76.77% (n = 1457)	
	Liver		21.96% (n = 1693)	26.84% (n = 1560)	7.01% (n = 133)	
	Pancreas		12.06% (n = 930)	14.14% (n = 822)	5.69% (n = 108)	
47	Esophagus System	79.63	6.07% (n = 373)	6.40% (n = 274)	5.33% (n = 99)	
48	Stomach System	79.63	6.17% (n = 379)	6.28% (n = 269)	5.93% (n = 110)	
	Intestine System	79.63	58.32% (n = 3581)	49.65% (n = 2127)	78.34% (n = 1454)	<0.01
	Liver System	79.63	22.23% (n = 1365)	28.13% (n = 1205)	8.62% (n = 160)	<0.01
	Pancreas System	79.63	14.35% (n = 881)	17.88% (n = 766)	6.20% (n = 115)	<0.01
52	System Count (n)	79.63	1.07 (SD = 0.27)	1.08 (SD = 0.29)	1.04 (SD = 0.23)	
53	Urgency	100				
	N4		2.98% (n = 230)	2.39% (n = 139)	4.79% (n = 91)	
	N5		97.02% (n = 7481)	97.61% (n = 5674)	95.21% (n = 1807)	
	Resurgery	100	0.73% (n = 56)	0.64% (n = 37)	1.00% (n = 19)	0.14

Table D.4: (continued) Statistical analysis of the 17 intervention characteristics and statistical comparison of these between the two included surgical centers.

#	Feature	Complete (%)	Combined (C1 + C2)	Internal (C1)	External (C2)	P
55	Month	100				
	January		8.48% (n = 654)	8.58% (n = 499)	8.17% (n = 155)	
	February		8.77% (n = 676)	8.69% (n = 505)	9.01% (n = 171)	
	March		9.27% (n = 715)	9.03% (n = 525)	10.01% (n = 190)	
	April		8.26% (n = 637)	8.43% (n = 490)	7.74% (n = 147)	
	May		8.59% (n = 662)	8.64% (n = 502)	8.43% (n = 160)	
	June		8.66% (n = 668)	8.46% (n = 492)	9.27% (n = 176)	
	July		8.42% (n = 649)	8.58% (n = 499)	7.90% (n = 150)	
	August		8.40% (n = 648)	8.27% (n = 481)	8.80% (n = 167)	
	September		8.69% (n = 670)	8.70% (n = 506)	8.64% (n = 164)	
	October		7.96% (n = 614)	7.98% (n = 464)	7.90% (n = 150)	
	November		8.34% (n = 643)	8.39% (n = 488)	8.17% (n = 155)	
	December		6.16% (n = 475)	6.23% (n = 362)	5.95% (n = 113)	
56	Weekday	100				
	Monday		18.65% (n = 1438)	18.39% (n = 1069)	19.44% (n = 369)	
	Tuesday		22.28% (n = 1718)	23.64% (n = 1374)	18.12% (n = 344)	
	Wednesday		19.60% (n = 1511)	19.16% (n = 1114)	20.92% (n = 397)	
	Thursday		20.49% (n = 1580)	19.34% (n = 1124)	24.03% (n = 456)	
	Friday		18.52% (n = 1428)	19.27% (n = 1120)	16.23% (n = 308)	
	Saturday		0.27% (n = 21)	0.07% (n = 4)	0.90% (n = 17)	
	Sunday		0.19% (n = 15)	0.14% (n = 8)	0.37% (n = 7)	
57	Daytime	100				<0.01
	Morning		90.83% (n = 7004)	93.31% (n = 5424)	83.25% (n = 1580)	
	Afternoon		8.94% (n = 689)	6.45% (n = 375)	16.54% (n = 314)	
	Night		0.23% (n = 18)	0.24% (n = 14)	0.21% (n = 4)	
58	Hour	100				<0.01
	0		0.09% (n = 7)	0.10% (n = 6)	0.05% (n = 1)	
	1		0.01% (n = 1)	0.02% (n = 1)	0.00% (n = 0)	
	2		0.03% (n = 2)	0.00% (n = 0)	0.11% (n = 2)	
	4		0.03% (n = 2)	0.03% (n = 2)	0.00% (n = 0)	
	5		0.01% (n = 1)	0.00% (n = 0)	0.05% (n = 1)	
	6		0.01% (n = 1)	0.02% (n = 1)	0.00% (n = 0)	
	7		0.29% (n = 22)	0.36% (n = 21)	0.05% (n = 1)	
	8		46.05% (n = 3551)	48.24% (n = 2804)	39.36% (n = 747)	
	9		13.31% (n = 1026)	15.05% (n = 875)	7.96% (n = 151)	
	10		8.77% (n = 676)	8.65% (n = 503)	9.11% (n = 173)	
	11		9.34% (n = 720)	9.34% (n = 543)	9.33% (n = 177)	
	12		7.33% (n = 565)	6.49% (n = 377)	9.91% (n = 188)	
	13		5.75% (n = 443)	5.16% (n = 300)	7.53% (n = 143)	
	14		4.21% (n = 325)	3.34% (n = 194)	6.90% (n = 131)	
	15		1.78% (n = 137)	1.29% (n = 75)	3.27% (n = 62)	
	16		1.22% (n = 94)	0.84% (n = 49)	2.37% (n = 45)	
	17		0.89% (n = 69)	0.48% (n = 28)	2.16% (n = 41)	
	18		0.54% (n = 42)	0.34% (n = 20)	1.16% (n = 22)	
	19		0.16% (n = 12)	0.10% (n = 6)	0.32% (n = 6)	
	20		0.09% (n = 7)	0.03% (n = 2)	0.26% (n = 5)	
	21		0.04% (n = 3)	0.02% (n = 1)	0.11% (n = 2)	
	22		0.04% (n = 3)	0.05% (n = 3)	0.00% (n = 0)	
	23		0.03% (n = 2)	0.03% (n = 2)	0.00% (n = 0)	
59	Mean Monthly Temp. (°C)	100	11.43 (SD = 6.83)	11.40 (SD = 6.81)	11.53 (SD = 6.91)	0.43
60	Minimum Monthly Temp. (°C)	100	1.33 (SD = 6.55)	1.28 (SD = 6.56)	1.45 (SD = 6.52)	0.53
61	Maximum Monthly Temp. (°C)	100	23.69 (SD = 8.58)	23.64 (SD = 8.55)	23.82 (SD = 8.68)	0.26
62	Monthly Sunshine H. (h)	98.64	152.76 (SD = 84.82)	151.85 (SD = 84.24)	155.58 (SD = 86.56)	0.06

Table D.5: Statistical analysis of the 16 laboratory values and statistical comparison of these between the two included surgical centers. Numerical features are given as an average with standard deviation. C1 = Center 1, C2 = Center 2, SD = Standard Deviation, gGT = gamma-Glutamyl Transpeptidase, CRP = C-Reactive Protein, WBC = White Blood Count, INR = International Normalized Ratio, aPTT= activated Partial Thromboplastin Time. Used under CC BY-NC 4.0 license from the supplements of Riepe et al. [Rie+25b].

#	Feature	Complete (%)	Combined (C1 + C2)	Internal (C1)	External (C2)	
63	Sodium (mmol / L)	96.98	139.53 (SD = 2.90)	139.53 (SD = 2.89)	139.54 (SD = 2.91)	0.7
64	Potassium (mmol / L)	99.31	3.99 (SD = 0.45)	4.00 (SD = 0.45)	3.99 (SD = 0.45)	0.56
65	Bilirubin (mg / dL)	63.18	0.83 (SD = 1.83)	0.83 (SD = 1.78)	0.87 (SD = 2.14)	0.15
66	Urea (mg / dL)	72.75	32.55 (SD = 18.48)	32.67 (SD = 18.25)	32.09 (SD = 19.27)	<0.01
67	gGT (U / L)	72.83	129.16 (SD = 265.02)	136.20 (SD = 273.50)	104.53 (SD = 231.38)	<0.01
68	Lipase (U / L)	64.83	60.95 (SD = 159.35)	62.65 (SD = 167.50)	54.02 (SD = 120.11)	<0.01
69	CRP (mg / L)	63.05	19.19 (SD = 40.67)	17.15 (SD = 38.30)	27.73 (SD = 48.46)	<0.01
70	Hemoglobin (g / dL)	99.47	12.53 (SD = 2.09)	12.60 (SD = 2.05)	12.29 (SD = 2.20)	<0.01
71	WBC Count (cells / nL)	99.34	7.59 (SD = 2.92)	7.53 (SD = 2.91)	7.77 (SD = 2.96)	<0.01
72	Platelets (cells / nL)	99.31	275.09 (SD = 105.01)	267.80 (SD = 100.97)	297.41 (SD = 113.69)	<0.01
73	Hematocrit (L / L)	46.95	0.37 (SD = 0.06)	0.37 (SD = 0.06)	0.35 (SD = 0.07)	<0.01
74	INR (-)	99.01	1.03 (SD = 0.11)	1.03 (SD = 0.12)	1.04 (SD = 0.10)	0.07
75	aPTT (sec)	84.27	34.57 (SD = 5.61)	34.64 (SD = 5.49)	34.34 (SD = 6.00)	<0.01
76	Erythrocytes (cells / pL)	99.31	4.39 (SD = 0.66)	4.40 (SD = 0.65)	4.35 (SD = 0.69)	0.03
77	Creatinine (mg / dL)	79.03	0.98 (SD = 0.89)	0.98 (SD = 0.91)	0.98 (SD = 0.66)	0.01
78	Glucose (mg / dL)	50.03	123.11 (SD = 43.32)	123.22 (SD = 42.83)	122.66 (SD = 45.40)	0.72

Table D.6: Model performance for both endpoints during internal testing, shown as mean and standard deviation across all folds and seeds. The best metrics are **emboldened**. XGBoost = eXtreme Gradient Boosting [CG16], AUROC = Area Under the Receiver Operating Characteristic, AUPRC = Area Under the Precision Recall Curve, MCC = Matthews correlation coefficient. Used under CC BY-NC 4.0 license from the supplements of Riepe et al. [Rie+25b].

Endpoint	Classifier	AUROC	AUPRC	MCC	F1-Score
90DM	XGBoost	**0.85 (0.03)**	0.23 (0.04)	0.29 (0.04)	0.31 (0.04)
	Logistic Regression	**0.85 (0.02)**	**0.24 (0.05)**	**0.31 (0.05)**	**0.34 (0.05)**
	Balanced Random Forest	**0.85 (0.02)**	0.23 (0.04)	0.28 (0.03)	0.3 (0.03)
	Chance Level	0.50	0.04	0.00	0.04
30DM	XGBoost	0.85 (0.03)	0.14 (0.03)	0.22 (0.05)	0.22 (0.05)
	Logistic Regression	**0.87 (0.03)**	**0.17 (0.06)**	**0.25 (0.06)**	**0.25 (0.06)**
	Balanced Random Forest	0.86 (0.03)	0.14 (0.04)	0.23 (0.05)	0.23 (0.04)
	Chance Level	0.50	0.02	0.00	0.02

Table D.7: Software libraries and respective versions used within this study. Used under CC BY-NC 4.0 license from the supplements of Riepe et al. [Rie+25b].

#	Library	Version
1	graphviz	0.20.1
2	imbalanced-learn	0.10.1
3	lightgbm	3.3.5
4	matplotlib	3.5.2
5	numba	0.56.4
6	numpy	1.22.4
7	pandas	2.1.4
8	scikit-learn	1.3.2
9	SciPy	1.14.0
10	seaborn	0.12.2
11	shap	0.41.0
12	xgboost	1.7.6

D.2 Detailed Inclusion Criteria

Table D.8: Patient inclusion flowchart.

Group	Count
Consented patients	1285
Excluded from study	94
No major operation performed	69
Withdrawn (no wearables used)	18
Implausible outcome definition	4
Patients for analysis	1194

Table D.9: Patient complication groups.

Group	Count
Patients for analysis	1194
No complications	813
Only minor complications	123
At least one major complication	258
Target major complication	121
SSI-3*	47
Bile/BDA† leak	38
POPF‡	36
Other major complication	137
Gastrointestinal	27
Bleeding	24
Biliary	21
Wound healing	18
Pulmonary	17
Cardiovascular	9
Renal	9
Infectious	6
Neurological	3
Pancreatic	3

* SSI-3 = Deep surgical site infection
† BDA = Biliodigestive anastomosis
‡ POPF = Postoperative Pancreatic Fistula

Patients were included in the study with informed consent. Their age was 18 years or more. The groups included in the study are: Visceral Surgical Major Resections in the Organ Systems of Liver, Pancreas, Upper and Lower Gastrointestinal Tract according to several OPS (*Operationen- und Prozedurenschlüssel* [Bun25]) groups. Intestine

- Small Intestine Resections: 5-454.0 to 5-545.y
- Reversal of a Double-Barreled Enterostomy: 5-465.0 to 5-465.y
- Restoration of Intestinal Continuity in Terminal Enterostomies: 5-466.0 to 5-466.y

- Partial Resection of the Colon: 5-455.0 to 5-455.y
- Total Colectomy, Variations: 5-456.0 to 5-456.y
- Rectum Resection with Sphincter Preservation: 5-484.0 to 5-484.y
- Rectum Resection without Sphincter Preservation: 5-485.0 to 5-485.y

Upper Gastrointestinal Tract:

- Partial Esophagectomy without Restoration of Continuity: 5-423.0 to 5-423.y
- Partial Esophagectomy with Restoration of Continuity: 5-424.0 to 5-424.y
- (Total) Esophagectomy without Restoration of Continuity: 5-425.0 to 5-425.y
- (Total) Esophagectomy with Restoration of Continuity: 5-426.0 to 5-426.y
- Atypical Partial Gastrectomy: 5-434.0 to 5-434.y
- Partial Gastrectomy (2/3 Resection): 5-435.0 to 5-435.y
- Subtotal Gastrectomy (4/5 Resection): 5-436.0 to 5-436.y
- (Total) Gastrectomy: With Esophagojejunostomy: 5-437.0 to 5-437.y
- (Total) Gastrectomy with Subtotal Esophagectomy: 5-438.0 to 5-438.y

Pancreas Surgery:

- Partial Resection of the Pancreas: 5-524.0 to 5-524.y
- (Total) Pancreatectomy: 5-525.0 to 5-525.y

Liver Surgery:

- Anatomical (Typical) Liver Resection: 5-502.0 to 5-502.y
- Atypical Resections: 5-501.0 to 5-501.y

The following groups of patients are excluded:

- Patients not capable of giving consent
- Patients who have undergone an organ transplant during the same stay
- Performance of an additional intraoperative hyperthermic chemotherapy (HiPEC)

Table D.10: Major organ system groups included in the study.

Amount	Surgery type
Liver	491
Colorectal	350
Pancreas	300
Upper GI	53
Other	1

Table D.11: Numeric target statistics for the Prospective Cohort

	median
Organ System	2.0 [1.9, 2.1]
Nutrition Risk Screening (NRS) 2002 Score	2.0 [1.9, 2.1]
Height	174.0 [173.4, 174.6]
Weight	76.0 [75.0, 77.0]
Age At Surgery	63.0 [62.2, 63.8]
Diabetes Without Chronic Complications	0.0 [0.0, 0.0]
AIDS (HIV)	0.0 [0.0, 0.0]
Psychoses	0.0 [0.0, 0.0]
Jones	2.0 [1.9, 2.1]
Eastern Cooperative Oncology Group (ECOG)	1.0 [1.0, 1.0]
Charlson Comorbidity Index (CCI)	7.0 [6.8, 7.2]
Operation Duration	267.0 [258.8, 275.2]
Pancreas Transplantation	0.0 [0.0, 0.0]
Lung Resection	0.0 [0.0, 0.0]
Kidney Resection	0.0 [0.0, 0.0]
Simultaneous Hernia Repair	0.0 [0.0, 0.0]
Simultaneous Ablation	0.0 [0.0, 0.0]

D.3 Detailed Data Characteristics

Table D.14: Lab values and their amounts

Index	Full Name	Name	Unit	Count
0	Alanin Aminotransferase	alat	U/l	36,671
1	Albumin	alb	g/l	12,811
2	Alkaline Phosphatase	alp	U/l	31,109
3	Amylase	ams	U/l	2,099
4	Partial Thromboplastin Time	aptt	sec	35,452
5	Aspartate Aminotransferase	asat	U/l	36,604
6	Antithrombin	at	%	4,240
7	Basophils	baso	/nl	11,853
8	Basophils Relative	baso_rel	%	11,684
9	Base Excess	be	mmol/l	54,687
10	Calcium	ca	mmol/l	62,873
11	Creatine Kinase	ck	U/l	2,556
12	Creatine Kinase MB	ck_mb	U/l	932
13	Chloride	cl	mmol/l	57,260
14	Carboxyhemoglobin	cohb	%	55,047
15	Creatinine	cr	mg/dl	42,560
16	C Reactive Protein	crp	mg/l	33,283
17	Cystatin C	cys_c	mg/l	291
18	D Dimer	d_dim	mg/l	168

Continued on next page

Table D.14: Lab values and their counts (continued)

Index	Full Name	Name	Unit	Count
19	Direct Bilirubin	dbil	mg/dl	3,575
20	Erythroblasts	ebl	/nl	7,496
21	Erythroblasts Relative	ebl_rel	%	208
22	Eosinophils	eos	/nl	11,871
23	Eosinophils Relative	eos_rel	%	11,729
24	Iron	fe	µmol/l	3,351
25	Ferritin	fer	µg/l	2,977
26	Fibrinogen	fg	g/l	5,349
27	Fraction of Inspired Oxygen	fio2	%	56,033
28	Glutamate Dehydrogenase	gdh	U/l	1,952
29	Gamma Glutamyltransferase	ggt	U/l	35,398
30	Glucose	glu	mg/dl	78,908
31	Hemoglobin	hb	g/dl	100,512
32	Glycated Hemoglobin	hba1c	%	1,899
33	Bicarbonate	hco3	mmol/l	110,467
34	Hematocrit	hct	%	55,154
35	High Density Lipoprotein	hdl	mg/dl	1,766
36	Deoxyhemoglobin	hhb	%	54,550
37	Haptoglobin	hp	g/l	347
38	Indirect Bilirubin	ibil	mg/dl	691
39	Immature Granulocytes	ig	/nl	11,752
40	Immature Granulocytes Relative	ig_rel	%	10,889
41	Immunoglobulin A	iga	g/l	391
42	Immunoglobulin E	ige	kU/l	58
43	Immunoglobulin G	igg	g/l	435
44	Immunoglobulin M	igm	g/l	393
45	International Normalized Ratio	inr	-	33,745
46	I/T Ratio	it_ratio	-	11,408
47	Potassium	k	mmol/l	88,764
48	Lactate	lac	mg/dl	54,220
49	Lactate Dehydrogenase	ldh	U/l	10,725
50	Low Density Lipoprotein	ldl	mg/dl	1,743
51	Lipase	lps	U/l	26,592
52	Lipase drain	lps_drain	U/l	12,265
53	Lymphocytes	lym	/nl	11,875
54	Lymphocytes Relative	lym_rel	%	11,751
55	Mean Corpuscular Hemoglobin	mch	pg	45,326
56	Mean Corpuscular Hemoglobin Concentration	mchc	g/dl	45,286
57	Mean Corpuscular Volume	mcv	fl	45,308
58	Methemoglobin	methb	%	55,059
59	Magnesium	mg	mmol/l	6,581
60	Monocytes	mono	/nl	11,890
61	Monocytes Relative	mono_rel	%	11,745
62	Mean Platelet Volume	mpv	fl	44,341
63	Myelocytes	myelo	%	344
64	Sodium	na	mmol/l	88,283
65	Ammonia	nh3	µmol/l	211
66	N Terminal Pro B Type Natriuretic Peptide	nt_probnp	ng/l	800
67	Oxyhemoglobin	o2hb	%	55,146
68	Phosphorus	p	mmol/l	168
69	Pseudocholinesterase	pche	kU/l	1,487
70	Carbon Dioxide Partial Pressure	pco2	mmHg	110,054
71	Procalcitonin	pct	µg/l	4,010
72	Potential of Hydrogen	ph	-	57,796
73	Platelets	plt	/nl	45,271
74	Neutrophils	pmn	/nl	11,850
75	Neutrophils Relative	pmn_rel	%	10,876
76	Oxygen Partial Pressure	po2	mmHg	55,069
77	Phosphate	po4	mmol/l	8,859
78	Protein	pro	g/l	3,218
79	Quick Value	quick	%	33,677
80	Erythrocytes	rbc	/pl	45,286
81	Red Cell Distribution Width	rdw	%	45,177
82	Reticulocytes	rtic	/nl	4,438
83	Schistocytes	schisto	%	122
84	Oxygen Saturation	so2	%	55,625

Continued on next page

Table D.14: Lab values and their counts (continued)

Index	Full Name	Name	Unit	Count
85	Total Bilirubin	tbil	mg/dl	60,499
86	Total Bilirubin Drain	tbil_drain	mg/dl	10,077
87	Total Cholesterol	tc	mg/dl	2,581
88	Temperature	temp	°C	56,684
89	Total Triglycerides	tg	mg/dl	2,915
90	Transferrin	trans	g/l	3,260
91	Transferrin Saturation	ts	%	3,178
92	Thyroid Stimulating Hormone	tsh	mU/l	10,048
93	Uric Acid	ua	mg/dl	6,882
94	Urea	urea	mg/dl	35,332
95	25 OH Vitamin D3	vd25	nmol/l	876
96	Leukocytes	wbc	/nl	45,271

Table D.15: ICU observations and their amounts

	Full Name	Unit	Count	Count per ID	Std Dev per ID
0	Bilirubin SOFA	mg/dL	4,560	3.0	11.0
1	Creatinine SOFA	mg/dL	4,560	3.0	11.0
2	Fraction of Inspired Oxygen	%	395,704	207.0	401.0
3	Glasgow Coma Scale SOFA	score	4,560	3.0	11.0
4	Height	cm	1,542	1.0	1.0
5	Heart Rate	bpm	1,088,887	586.0	922.0
6	Hypotension SOFA	score	4,560	3.0	11.0
7	Partial Pressure of Oxygen SOFA	mmHg	4,560	3.0	11.0
8	Respiratory Rate	breaths/min	266,882	101.0	416.0
9	Oxygen Saturation	%	550,592	296.0	459.0
10	Sequential Organ Failure Assessment	score	4,560	3.0	11.0
11	Temperature	°C	344,061	184.0	347.0
12	Thrombocytes SOFA	count	4,560	3.0	11.0
13	Invasive Blood Pressure	mmHg	1,160,328	631.5	1,175.0
14	Non-Invasive Blood Pressure	mmHg	652,467	282.0	747.0
15	Weight	kg	3,832	2.0	4.0

Table D.16: ICU scores and their amounts

	Full Name	Count	Count per ID	Std Dev per ID
0	APACHE II	849	1.0	1.0
1	CAM-ICU	17,133	8.0	28.0
2	Numeric Rating Scale	85	1.0	2.0
3	Frailty Score	279	1.0	0.0
4	Glasgow Coma Scale	20,826	9.0	34.0
5	NIH Stroke Scale	101	7.0	9.0
6	BSAS	6	1.5	1.0
7	DDS8	645	2.0	3.0
8	ICDSC	909	2.0	6.0
9	RASS	25,099	11.0	37.0
10	SAPS II	4,583	3.0	11.0
11	SOFA	4,560	3.0	11.0
12	BPS	2,530	2.0	23.0
13	Delirium Score	7,872	4.0	6.0
14	FAS	2,238	1.0	4.0
15	Neurostatus	22	11.0	7.0
16	TISS-10	4,755	3.0	11.0

D.4 Model Evaluation

A double-nested cross-validation setup was used for all experiments. The hyperparameters are tuned separately for each dataset with at least 50 iterations. The model evaluation code is available at: `https://github.com/rvandewater/YAIB`. The following hyperparameters were used for *CASHEWS+* (rounded where appropriate):

Table D.12: Binary target statistics for the Prospective Cohort

	amount
Myocardial Infarction	95
Congestive Heart Failure	30
Peripheral Vascular Disease	73
Cerebrovascular Disease	9
Dementia	63
Chronic Pulmonary Disease	486
Rheumatic Disease	7
Peptic Ulcer Disease	18
Mild Liver Disease	108
Diabetes With Chronic Complications	428
Hemiplegia or Paraplegia	11
Renal Disease	88
Any Malignancy	479
Moderate or Severe Liver Disease	104
Metastatic Solid Tumor	497
Cardiac Arrhythmias	30
Valvular Disease	10
Pulmonary Circulation Disorder	1
Arterial Hypertension	608
Other Neurological Disorders	10
Hypothyroidism	31
Coagulopathy	21
Obesity	172
Weight Loss	6
Fluid and Electrolyte Disorders	201
Blood Loss Anemia	12
Deficiency Anemia	13
Alcohol Abuse	82
Drug Abuse	48
Depression	10
Coronary Heart Disease (KHK)	95
Chronic Pancreatitis	204
Ulcerative Colitis	25
Crohn's Disease	173
Pancreas Reconstruction	92
Small Bowel Anastomosis	98
Colon Anastomosis	162
Rectum Anastomosis	106
Stoma Reversal	80
Reconnection Surgery	23
Cholecystectomy	50
Cholecystectomy in Context	351
Vascular Resection	43
Biliary Drainage (BDA)	257
Splenectomy	102
Stoma Placement	257
Hyperthermic Intraperitoneal Chemotherapy (HIPEC)	2
Other Procedures	30
Previous Operation	6

- `n_estimators`: 2000
- `max_depth`: 3
- `num_leaves`: 8
- `min_child_samples`: 10
- `min_child_weight`: 10.0
- `learning_rate`: 0.018
- `reg_alpha`: 0.000 001

Table D.13: Categorical target statistics for the Prospective Cohort

Category	Subcategory	Amount
Sex	Male	716
Sex	Female	479
Functional Status	Partially Dependent	689
Functional Status	Independent	360
Pancreas Resection	None	287
Pancreas Resection	Pancreatectomy	106
Pancreas Resection	Left Resection	89
Pancreas Resection	Head Resection	84
Pancreas Resection	Segment Resection	5
Pancreas Resection	Beger Operation	2
Pancreas Resection	NaN	1
Liver Resection	None	501
Liver Resection	Major	330
Liver Resection	Minor	171
Esophagus Resection	None	40
Esophagus Resection	Total	31
Esophagus Resection	Transhiatal Extended Total Gastrectomy + Distal Esophagus Resection	5
Esophagus Resection	Abdominal Thoracic	3
Esophagus Resection	NaN	1
Esophagus Anastomosis	None	39
Esophagus Anastomosis	Thoracic	33
Esophagus Anastomosis	Transhiatal	6
Stomach Resection	None	34
Stomach Resection	Total	17
Stomach Resection	Atypical	15
Stomach Resection	Partial	2
Stomach Reconstruction	None	17
Stomach Reconstruction	Esophagojejunostomy	15
Stomach Reconstruction	Gastrojejunostomy	2
Small Bowel Resection	None	34
Small Bowel Resection	Ileum	23
Small Bowel Resection	Jejunum	9
Small Bowel Resection	Duodenum	2
Colon Resection	None	135
Colon Resection	Segment Resection	13
Colon Resection	Right Hemicolectomy	48
Colon Resection	Sigmoid Resection	32
Colon Resection	Colectomy	12
Colon Resection	Left Hemicolectomy	12
Colon Resection	Ileocecal Resection	9
Colon Resection	Proctocolectomy	7
Colon Resection	Transverse Resection	1
Colon Resection	Cecum Resection	1
Rectum Resection	None	125
Rectum Resection	Total Mesorectal Excision (TME)	97
Rectum Resection	Rectum Amputation	19
Rectum Resection	Transanal Total Mesorectal Excision (TaTME)	9
Liver Transplantation	None	0
Kidney Transplantation	None	0

- `reg_lambda`: 1.0
- `subsample`: 0.4
- `subsample_freq`: 1
- `colsample_bytree`: 0.4
- `colsample_bynode`: 0.4
- `boosting_type`: dart

- drop_rate: 0.1
- max_drop: 10
- skip_drop: 0.1
- top_rate: 0.1
- other_rate: 0.2
- feature_fraction: 0.4
- bagging_fraction: 0.4
- bagging_freq: 0
- min_split_gain: 1.0
- cat_smooth: 100.0
- cat_l2: 2.94
- eval_metric: binary_error
- is_unbalance: True
- scale_pos_weight: 0.34

```python
from statistics import stdev
from scipy.stats import t
from math import sqrt

def corrected_dependent_ttest(data1, data2, n_training_samples, n_test_samples, alpha):
    n = len(data1)
    differences = [(data1[i]-data2[i]) for i in range(n)]
    sd = stdev(differences)
    divisor = 1 / n * sum(differences)
    test_training_ratio = n_test_samples / n_training_samples
    denominator = sqrt(1 / n + test_training_ratio) * sd
    t_stat = divisor / denominator
    # degrees of freedom
    df = n - 1
    # calculate the critical value
    cv = t.ppf(1.0 - alpha, df)
    # calculate the p-value
    p = (1.0 - t.cdf(abs(t_stat), df)) * 2.0
    # return everything
    return t_stat, df, cv, p
```

Listing D.1: Implementation of the Nadeau and Bengio correction of dependent Student's t-test [NB03]. Inspired by an open-source implementation [De 21].

D.5 Detailed Results

Table D.17: The p-values for the AUROC of leaving out specific data modalities by comparing eight model configurations using the Friedman [Fri40] test. This indicated significant performance differences across modalities ($\chi^2 = 107.28$, $p < 0.0001$). Methodology was inspired by Rainio et al. [RTK24] and Demšar [Dem06]. Post-hoc pairwise correction (Dunn's test with Benjamini-Hochberg correction from scikit-posthocs [Ter19]).

Missing modality	All features	Pre-operative	Intra-operative	ICU monitors	Notes and meds	Time of day	Laboratory tests	GW wearables
All features	1.00	0.50	0.90	0.96	0.90	0.05	0.01	<0.01
Pre-operative	0.50	1.00	0.37	0.48	0.36	0.30	0.09	<0.01
Intra-operative	0.90	0.37	1.00	0.91	0.96	0.03	<0.01	<0.01
ICU monitors	0.96	0.48	0.91	1.00	0.90	0.04	0.01	<0.01
Notes and meds	0.90	0.36	0.96	0.90	1.00	0.03	<0.01	<0.01
Time of day	0.05	0.30	0.03	0.04	0.03	1.00	0.70	<0.01
Laboratory tests	0.01	0.09	<0.01	0.01	<0.01	0.70	1.00	<0.01
GW wearables	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1.00

Table D.18: Horizon ablation results.

Experiment	AUROC	AUPRC	Brier Score	AUROC CI	AUPRC CI	Brier CI
24h	0.814 ± 0.031	0.110 ± 0.048	0.015 ± 0.001	[0.802, 0.827]	[0.090, 0.130]	[0.015, 0.015]
12h	0.854 ± 0.022	0.153 ± 0.063	0.008 ± 0.000	[0.845, 0.863]	[0.127, 0.179]	[0.008, 0.008]
9h	0.886 ± 0.018	0.236 ± 0.066	0.006 ± 0.000	[0.879, 0.894]	[0.208, 0.263]	[0.006, 0.006]
6h	0.888 ± 0.014	0.292 ± 0.066	0.004 ± 0.000	[0.882, 0.894]	[0.265, 0.319]	[0.004, 0.004]
3h	0.899 ± 0.016	0.335 ± 0.059	0.002 ± 0.000	[0.892, 0.906]	[0.311, 0.360]	[0.002, 0.002]
1h	0.890 ± 0.034	0.306 ± 0.083	0.001 ± 0.000	[0.876, 0.905]	[0.271, 0.340]	[0.001, 0.001]

Table D.19: Main results.

Experiment	AUROC	AUPRC	Brier Score	AUROC CI	AUPRC CI	Brier CI
Pre-OP	0.492 ± 0.066	0.149 ± 0.038	0.114 ± 0.002	[0.464, 0.519]	[0.133, 0.165]	[0.113, 0.114]
+Intra-OP	0.491 ± 0.064	0.130 ± 0.027	0.102 ± 0.002	[0.465, 0.517]	[0.119, 0.141]	[0.102, 0.103]
+ICU	0.504 ± 0.057	0.134 ± 0.029	0.102 ± 0.002	[0.481, 0.527]	[0.122, 0.146]	[0.102, 0.103]
CASHEWS	0.815 ± 0.032	0.122 ± 0.038	0.006 ± 0.000	[0.802, 0.828]	[0.106, 0.138]	[0.006, 0.006]
CASHEWS+	0.886 ± 0.018	0.236 ± 0.066	0.006 ± 0.000	[0.879, 0.894]	[0.208, 0.263]	[0.006, 0.006]

Table D.20: Model variation results.

Experiment	AUROC	AUPRC	Brier Score	AUROC CI	AUPRC CI	Brier CI
ICU & GW	0.886 ± 0.018	0.236 ± 0.066	0.006 ± 0.000	[0.879, 0.894]	[0.208, 0.263]	[0.006, 0.006]
GW only	0.797 ± 0.033	0.137 ± 0.045	0.006 ± 0.000	[0.784, 0.811]	[0.119, 0.155]	[0.006, 0.006]
ICU only	0.948 ± 0.033	0.305 ± 0.117	0.005 ± 0.002	[0.934, 0.962]	[0.257, 0.353]	[0.004, 0.005]

Table D.21: Feature ablation results.

Experiment	AUROC	AUPRC	Brier Score	AUROC CI	AUPRC CI	Brier CI
All features	0.848 ± 0.019	0.183 ± 0.054	0.006 ± 0.000	[0.840, 0.855]	[0.161, 0.205]	[0.006, 0.006]
Pre-operative	0.842 ± 0.020	0.181 ± 0.053	0.006 ± 0.000	[0.834, 0.850]	[0.160, 0.203]	[0.006, 0.006]
Intra-operative	0.850 ± 0.019	0.188 ± 0.054	0.006 ± 0.000	[0.842, 0.857]	[0.165, 0.210]	[0.006, 0.006]
ICU monitors	0.848 ± 0.020	0.171 ± 0.055	0.006 ± 0.000	[0.840, 0.856]	[0.148, 0.194]	[0.006, 0.006]
Notes and meds	0.850 ± 0.019	0.177 ± 0.048	0.006 ± 0.000	[0.842, 0.858]	[0.157, 0.197]	[0.006, 0.006]
Time of day	0.833 ± 0.021	0.114 ± 0.039	0.006 ± 0.000	[0.824, 0.842]	[0.098, 0.130]	[0.006, 0.006]
Laboratory tests	0.829 ± 0.021	0.178 ± 0.054	0.006 ± 0.000	[0.820, 0.837]	[0.156, 0.200]	[0.006, 0.006]
GW wearables	0.773 ± 0.030	0.110 ± 0.059	0.006 ± 0.000	[0.761, 0.785]	[0.086, 0.135]	[0.006, 0.006]

Table D.22: Wearable feature ablation results.

Experiment	AUROC	AUPRC	Brier Score	AUROC CI	AUPRC CI	Brier CI
All features	0.848 ± 0.019	0.183 ± 0.054	0.006 ± 0.000	[0.840, 0.855]	[0.161, 0.205]	[0.006, 0.006]
w/o GW wearables	0.773 ± 0.030	0.110 ± 0.059	0.006 ± 0.000	[0.761, 0.785]	[0.086, 0.135]	[0.006, 0.006]
PPG embeddings	0.831 ± 0.026	0.154 ± 0.066	0.006 ± 0.000	[0.820, 0.841]	[0.126, 0.181]	[0.006, 0.006]
Vital signs	0.827 ± 0.014	0.164 ± 0.051	0.006 ± 0.000	[0.821, 0.833]	[0.143, 0.185]	[0.006, 0.006]
HRV features	0.829 ± 0.020	0.161 ± 0.069	0.006 ± 0.000	[0.820, 0.837]	[0.132, 0.189]	[0.006, 0.006]
Core temperature	0.829 ± 0.018	0.138 ± 0.051	0.006 ± 0.000	[0.821, 0.836]	[0.117, 0.159]	[0.006, 0.006]

D.6 TRIPOD+AI Statement for Clinical Prediction Models

This chapter adheres to the TRIPOD+AI guidelines for reporting clinical prediction models developed using machine learning techniques [Col+24]. The checklist is adapted to this specific work and is presented below for transparency and completeness.

Contents:

Title

Abstract

Methods

Result

Discussion

D.6.1 Continuous Multimodal AI with Wearable Vital Signs Predicts Postoperative Complications in the General Ward

1 Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted The target population is participants undergoing visceral surgery. The primary endpoint is every postoperative intra-abdominal fluid collection that requires intervention. In this context, intervention is an operation, endoscopy with treatment measures, or radiologic intervention measures. The end goal is the development and Evaluation of a Multivariable Model Combining Hospital-Grade Clinical Data and Wearable Vital Sign Monitoring for Continuous Prediction of Surgical Complications.

D.6.2 Abstract

From the TRIPOD+AI for abstracts: https://www.tripod-statement.org/wp-content/uploads/2019/12/TRIPODAI-for-Abstracts.pdf

1. Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted

"Continuous Multimodal AI with Wearable Vital Signs Predicts Postoperative Complications in the General Ward" — This title identifies the study as developing/evaluating a multivariable prediction model (AI with wearable vital signs), the target population (postoperative patients in the general ward), and the outcome (postoperative complications).

Background

2. Provide a brief explanation of the healthcare context and rationale for developing or evaluating the performance of all models

The document explains that while artificial intelligence has successfully predicted severe adverse events in ICUs (such as sepsis and mortality), most postoperative complications occur on general wards, where staffing and technical monitoring constraints impede real-time detection. Despite >25% of major abdominal surgery patients experiencing life-threatening events, physiological patterns are subtle, nonspecific, and time-varying, making early detection challenging. The rationale is to develop an AI-based

system that combines routine clinical data with continuous high-resolution vital signs from wearable telemetric photopletysmography (PPG) sensors to enable early identification of complications on general wards where intermittent bedside assessments are standard practice.

Objectives

3. Specify the study objectives, including whether the study describes model development, evaluation, or both

The study describes **both development and evaluation**. The primary objectives were: (1) to prospectively develop multimodal AI prediction models for surgical ward patients that integrate routine clinical data with high-resolution telemetric vital signs, (2) to create accurate AI-driven digital patient representations for early identification of clinically relevant complications (deep surgical site infections, bile leakages, pancreatic fistulas), and (3) to translate the model into an actionable alarm framework for medical staff.

Methods

4. Describe the sources of data

Data sources included:**Admission data**: RedCap clinical trial database. **Clinical information**: I.S.H.Med system (stay information and laboratory results). **ICU vital signs**: COPRA system. **Wearable data**: Corsano Band (PPG) and GreenTEG CORE (core temperature) devices collected via DeviceHub. **Clinical documentation**: Physicians' and nurses' free-text notes. The dataset comprised 5.4 million ICU measurements, 2.5 million laboratory values, and 13.8 billion PPG data points, all collected continuously at a rate of 32 Hz.

5. Describe the eligibility criteria and setting where the data were collected

Setting: Tertiary, high-volume surgical center (Charité University Hospital, Berlin, Germany) between 2022-2025 as part of the CASSANDRA (Clinical ASSist and Alert Algorithms) study.

Eligibility: Patients undergoing major planned esophageal, gastric, liver, pancreatic, and colorectal surgery who provided written informed consent. **Exclusions**: (1) major surgery not performed (n=69), (2) withdrawal from study (n=18), (3) outcome adjudication ambiguity (n=4), (4) lacking PPG signal quality or amount (n=58). **Final cohort**: 1,136 patients for machine learning development (from 1,285 initially enrolled).

6. Specify the outcome to be predicted by the model, including time horizon of predictions in case of prognostic models

The primary outcome is clinically relevant intra-abdominal complications requiring intervention (surgical, radiologic, or endoscopic), specifically deep surgical site infections (SSI-III), bile leakages, and postoperative pancreatic fistulas (combined as they all represent intra-abdominal pathologies requiring intervention). A time horizon of 9 hours was selected as the primary prediction horizon, balancing good performance with sufficient time to intervene. Evaluation also examined prediction horizons of 6 hours, 12 hours, and 24 hours. Complications occurring on general wards were the focus (86.5% of SSI-III cases, with a median of 6.9 days after surgery).

7. Specify the type of model, a summary of the model-building steps, and the method for internal validation

Model type: **LightGBM** (Gradient Boosted Trees) was selected after benchmarking against Logistic Regression, Random Forest, XGBoost, CatBoost, and deep learning architectures (Transformer, GRU, TCN).

Model-building steps:

- Early multimodal fusion of heterogeneous data (demographics, comorbidities, procedural data, ICU vitals, medications, laboratory results, clinical notes, wearable data)
- Data segmented into 30-minute intervals
- 7,280 total features extracted per segment (including PPG embeddings via MiniRocket, HRV features, temperature entropy, laboratory values, clinical note embeddings)
- Two models developed: CASHEWS (routine clinical data only) and CASHEWS+ (with wearable monitoring)
- Feature selection to top 500 features based on SHAP values

Internal validation: **5-times repeated 5-fold stratified cross-validation** with patient timelines split as a whole to prevent data leakage. Test set: 20% of subjects, validation set: 20% of training subset.

8. Specify the measures used to assess model performance (e.g., discrimination, calibration, clinical utility)

Discrimination: Area Under the ROC Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC) — used given low event prevalence (0.006). Alarm burden (alarms per patient per day), Precision and recall at various operating points and comparison with Remote Early Warning Score (REWS).

Alarm system optimization: Evaluated with mechanisms including 8-hour blackout period, persistence (3 activations within 6 hours), and 8-hour silencing period to reduce alarm fatigue.

Results

9. Report the number of participants and outcome events

- Total participants: 1,136 patients in ML cohort
- Outcome events (SSI-III and related complications): 111 total (9.8% of cohort), with 96 (86.5%) occurring on general ward
  - SSI-III: 47 cases
  - Bile/biliodigestive anastomosis leak: 39 cases
  - Postoperative pancreatic fistula: 37 cases
- No complications: 809 patients (71.2%)

- Only minor complications: 124 patients (10.9%)
- Other major complications: 138 patients (12.1%)

10. Summarize the predictors in the final model

Feature count by modality: 7,280 total features, including 2,017 PPG embeddings, 1,998 laboratory values, 972 intraoperative aggregated features, 576 wearable vital signs, 432 ICU scores, and others.

11. Report model performance estimates (with confidence intervals)

Primary endpoint (9-hour prediction horizon): CASHEWS+ AUROC: 0.89 (0.88-0.89) — after feature selection to 500 features. CASHEWS+ AUPRC: 0.11 (achieved 109% improvement over CASHEWS without wearables). Median AUROC (raw model): 0.90 (0.89-0.91) Performance across prediction horizons: 3 hours before detection: AUROC 0.90 ± 0.01, 6 hours before detection: AUROC 0.90 ± 0.03, and 24 hours before detection: AUROC 0.81 ± 0.03.

Discussion

12. Give an overall interpretation of the main results

The study successfully demonstrated that multimodal, real-time AI models, combining continuous wearable vital signs with routine clinical data, can achieve high performance in the early detection of postoperative intra-abdominal complications on general wards (AUROC 0.90, with predictions made 9 hours in advance). Continuous telemetric PPG monitoring increased AUROC by 8% and AUPRC by 109%.

Limitations acknowledged: Single-center study with standardized ERAS protocols (may limit generalizability); largest continuous monitoring cohort reported but could benefit from larger external validation; model utility requires demonstration through patient-centered outcomes (length of stay, complication rates, mortality) rather than technical performance alone.

Registration

13. Give the registration number and name of the registry or repository

- **Study ID**: DRKS00034973
- **URL**: https://drks.de/search/de/trial/DRKS00034973
- **Ethics approval**:
  - **Committee**: Ethics Committee of Charité University Hospital
  - **Approval number**: EA4/198/21
  - **Approval date**: 20 October 2021
- **Funding**: "Gemeinsamer Bundesausschuss (G-BA) Innovationsausschuss" framework — CASSANDRA project (Grant number: 01VSF20015)
- **Code/Data availability**:
  - **CASHEWS repository**: https://github.com/rvandewater/CASHEWS

  - **YAIB benchmark framework**: https://github.com/rvandewater/YAIB

- **Data availability**: Summary statistics only; full dataset cannot be released due to an ethics agreement, but code and preprocessing pipelines are available.

D.6.3 Introduction

Background

3a Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models The model is designed for diagnostic use to help identify postsurgical complications (deep surgical site infections). An early diagnosis of complications after surgery can improve patient outcomes, reduce healthcare costs, and optimise hospital resource use. In current standard practice, patients in the regular ward are not continuously monitored. There is a limited number of healthcare personnel responsible for monitoring a large number of patients. Moreover, patients need space and time to rest, and intensive monitoring in ICUs can be mentally overwhelming and physically restrictive, meaning that an unneeded prolonged ICU stay is undesirable for the quality of life and time of recovery of the patient.

Multiple studies assessing the use of continuous vital signs monitoring in general wards have shown that it is feasible within the workflow of healthcare workers and is also accepted by patients.

Developmental challenges

Conventional models rely solely on static preoperative or postoperative data without real-time updates. Our model incorporates real-time vital sign data. There are several challenges to implementing our approach into clinical practice:

- Many wearable device vendors have proprietary algorithms with different features and data output capabilities. It is often unclear what the source signal was and how it has been altered.
- Continuous support is necessary for extracting data from wearables.
- Setting up the infrastructure in the regular ward, added tasks in the clinical workflow (to assign, change, and clean the devices because of the battery).
- Development of an alarm model based on a wide array of features that must be checked and preprocessed.
- Establishing alarm thresholds or actions to be taken if abnormalities in vital signs are detected, and who is responsible for assessing them.

3b Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (eg, healthcare professionals, patients, public) Clinicians (doctors) in the surgical ward are the intended target audience for the developed tool. The intended purpose of the model is to identify

earlier complications. Several decisions can be made based on this model, including the probability of the patient having an SSI-III at a given point, which may necessitate additional tests to rule out the possibility of a false positive. If the suspicion of a false alarm has not been ruled out, they should order a CT scan. The target population consists of patients undergoing visceral surgery, including the types of surgery described below.

3c Describe any known health inequalities between sociodemographic groups Our study population was intentionally inclusive, with no restrictions on age (adults aged 18 and above), gender, socioeconomic status, or nationality, ensuring diverse representation across demographic groups.

The types of surgery included (abdominal visceral resections) are more frequently performed in older adults, with a median age of 63.0 [53.0, 71.0] (25th and 75th percentile) in our cohort. In our cohort, 59.9 percent are male, and 40.1 percent are female.

The only identifiable disparity arises from the difference in care resources based on the hospital station type. Specifically, patients monitored in the private station might experience more attentive care due to a higher nurse-to-patient and doctor-to-patient ratio. This higher staffing level can lead to earlier identification of complications, more frequent monitoring, and quicker responses to patient needs, which could impact outcomes.

Objectives

4 Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both) This study aims to collect data to implement an alarm system in a hybrid ward, where patients are monitored using minimally intrusive wearable vital signs equipment to address the lack of monitoring. We created a multimodal dataset comprising free-text clinical notes, medications, ICU parameters, laboratory values, and common comorbidities. Develop and evaluate a prediction model that identifies complications after visceral surgery, which can then be used to inform clinical practice and improve the quality of care.

D.6.4 Methods

Data

5a Describe the sources of data separately for the development and evaluation datasets (eg, randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data We conducted a routine care study that uses data from clinical practice. The primary source is structured EHR data; however, we also utilize text and waveform data. Specific data sources include lab tests, medications, wearable vital signs on the regular ward (PPG and temperature sensors), ICU scores and vital signs, clinical notes, demographics, general patient information, and surgery outcome variables.

5b Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up The start of recruitment

was 2022.03.09. The recruitment ended on 2025.02.28. No follow-up is planned, as the desired endpoint typically occurs during the patient's hospital admission.

Participants

6a Specify key elements of the study setting (eg, primary care, secondary care, general population) including the number and location of centres We include hospitalised patients in a single tertiary care university hospital.

6b Describe the eligibility criteria for study participants Patients were included in the study with informed consent. Their age was 18 years or older. The groups included in the survey are: Visceral Surgical Major Resections in the Organ Systems of the Liver, Pancreas, Upper and Lower Gastrointestinal Tract. The German procedure classification (Operationen- und Prozedurenschlüssel - OPS) is the official classification for the encoding of operations, procedures, and general medical measures (https://www.bfarm.de/EN/Code-systems/Classifications/OPS-ICHI/OPS/_node.html [Accessed 2025-07-10]). The following OPS groups are included:

Intestine

- Small Intestine Resections: 5-454.0 to 5-545.y
- Reversal of a Double-Barreled Enterostomy: 5-465.0 to 5-465.y
- Restoration of Intestinal Continuity in Terminal Enterostomies: 5-466.0 to 5-466.y
- Partial Resection of the Colon: 5-455.0 to 5-455.y
- Total Colectomy, Variations: 5-456.0 to 5-456.y
- Rectum Resection with Sphincter Preservation: 5-484.0 to 5-484.y
- Rectum Resection without Sphincter Preservation: 5-485.0 to 5-485.y

Upper Gastrointestinal Tract:

- Partial Esophagectomy without Restoration of Continuity: 5-423.0 to 5-423.y
- Partial Esophagectomy with Restoration of Continuity: 5-424.0 to 5-424.y
- (Total) Esophagectomy without Restoration of Continuity: 5-425.0 to 5-425.y
- (Total) Esophagectomy with Restoration of Continuity: 5-426.0 to 5-426.y
- Atypical Partial Gastrectomy: 5-434.0 to 5-434.y
- Partial Gastrectomy (2/3 Resection): 5-435.0 to 5-435.y
- Subtotal Gastrectomy (4/5 Resection): 5-436.0 to 5-436.y
- (Total) Gastrectomy: With Esophagojejunostomy: 5-437.0 to 5-437.y
- (Total) Gastrectomy with Subtotal Esophagectomy: 5-438.0 to 5-438.y

Pancreas Surgery:

- Partial Resection of the Pancreas: 5-524.0 to 5-524.y
- (Total) Pancreatectomy: 5-525.0 to 5-525.y

Liver Surgery:

- Anatomical (Typical) Liver Resection: 5-502.0 to 5-502.y
- Atypical Resections: 5-501.0 to 5-501.y

The following groups of patients are excluded:

- Patients not capable of giving consent
- Patients who have undergone an organ transplant during the same stay
- Performance of an additional intraoperative hyperthermic chemotherapy (HiPEC)

6c Give details of any treatments received, and how they were handled during model development or evaluation, if relevant Treatments and endpoints

The study's primary endpoint was the occurrence of postoperative intra-abdominal fluid collections that necessitated further clinical intervention. These interventions included one of the following:

- **Surgical intervention**: Reoperation or surgical revision to manage the fluid collection.
- **Endoscopic intervention**: Placement of stents, endoscopic vacuum therapy (EndoVAC), endoscopically placed drains, or papillotomy procedures.
- **Radiological intervention**: Image-guided percutaneous drainage, percutaneous transhepatic cholangiographic drainage (PTCD), or embolization procedures performed under radiological guidance.

Following the interventions, intra-abdominal fluid collections were further categorized into specific subtypes based on intraoperative or interventional findings, as well as laboratory and microbiological results. The classifications included:

- **Bile leaks or insufficiency of the biliodigestive anastomosis**, as defined by Koch et al.
- **Pancreatic fistulas, type B or C**, according to the International Study Group on Pancreatic Surgery (Bassi et al.).
- **Surgical site infections (SSI-3)** of the abdominal cavity, based on the criteria of the National Healthcare Safety Network (Centers for Disease Control and Prevention).

- **Postoperative fluid collections** without signs of infection or organ leak, as described by Katayama et al.
- **Hematomas**, identified by imaging or intraoperative findings, also as described by Katayama et al.

Model development

The model developed in this study is specifically designed to detect the complications mentioned in the previous section. Subsequent data was cut from the training data, as this would compromise our cross-validation development setup with data of patients who had already experienced the endpoint. The previous surgery that all patients in this study underwent is recorded and contained as static data. This data is available to the model as a feature to detect the complications. A subset of patients develops an SSI-III, Pancreatic Fistula, or Bile Leakage (our primary endpoint) while in the ICU; therefore, they were not monitored with wearables on the normal ward, but we have continuous vital signs from the ICU.

Data Preparation

7 Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups All code used for preprocessing is available at https://github.com/rvandewater/CASHEWS

Processing and Quality Assurance of In-Hospital Transfer Data: Patient movement data during hospitalization were extracted using automated scripts from the electronic health record (EHR), with a focus on the perioperative period. Key time points used as anchors for temporal alignment included hospital admission, surgery start time (incision), transfer to the normal surgical ward, and hospital discharge. To improve the precision of these variables and enhance their temporal interpretability, a dedicated pipeline was developed to label and structure all in-hospital transfer events. This included differentiation between planned ICU stays (e.g., immediate postoperative observation) and unplanned ICU readmissions, which served as important clinical indicators for deterioration. The structured transfer data was derived from both EHR logs and the institution's Health Data Lake (HDL), allowing for redundancy and quality checks across data sources. Special attention was given to patients who underwent multiple surgeries during the same admission, where identifying the relevant index surgery (i.e., the major abdominal operation of interest) was crucial. By cross-referencing ward movements, surgical timestamps, and ICU transfers, we ensured accurate alignment of predictor variables with the clinical course. All predictors were based on structured data fields and did not require manual interpretation.

We simplified the patient care process to four stages: patient admission, surgery, ICU stay, and general ward stay (Figure 4.5). We used a diverse range of data sources for modelling patients in our study. Several digital documentation systems were used to collect data. Admission information was recorded using a RedCAP[59] database clinical trial application. The primary clinical information system was I.S.H.Med60, which was primarily used to collect patient stay information and laboratory results. The ICU vital signs were sourced from the COPRA system[61], which was used to collect vital signs,

scores, and other ICU data. The wearable data was collected using the Corsano Band 1 (CardioWatch 287-1B) and Band 2 (CardioWatch 287-2B), wrist-worn devices that collect various physiological signals, of which the most important is PPG. Moreover, we measured the core temperature with the GreenTEG CORE device. The device data is collected using a Bluetooth device gateway mesh that transmits the data to a central server via the DeviceHub software[62]. Many features were renamed and aggregated after data exploration had determined that they represent the same measurement. Values were translated from German to English and standardized to a common unit of measurement. The timestamps are harmonized from the source time zone to the UTC+0 (GMT) time zone. Moreover, all features are strictly filtered to their respective time periods within each segment to ensure that no data leakage occurs; all features extracted from any events that occur in the current segment occur either during the segment or before it.

We use early modality fusion[63] to combine low-level features after performing initial preprocessing, feature extraction, and (where applicable) embedding. The segmented combined features were input to the modelling pipeline.

Numerical ICU monitoring and lab values: features were treated uniformly. They are aggregated per subject into 30-minute segments after experimentally determining the right segment length. For each segment, we extract the following statistical features: mean, minimum, maximum, standard deviation, count, skewness, entropy, median, and sum.

Wearable data feature extraction: The preprocessing pipeline first ensures that all PPG traces are standardized to 32 Hz and normalized per patient (by subtracting the mean and dividing by the standard deviation). Then, it splits the traces into contiguous stay-aligned segments and fixed-length windows (default: 60 s, 1920 samples) with optional overlap; windows with a standard deviation of zero are skipped. For embedding, MiniRocket applies deterministic, extremely sparse 9-length convolutions with only three nonzero taps (84 unique 3-tap combinations) at multiple log-spaced dilations to capture patterns across temporal scales suitable for 32 Hz sampling. For each kernel/dilation pair, a set of scalar bias thresholds is assigned using a low-discrepancy quantile sequence computed from empirical convolution outputs, resulting in diverse binary detectors. At transformation time, each dilated convolutional response is filtered with the proportion of positive values (PPV) operator (1 if the response exceeds the bias, otherwise 0) and then pooled by computing the temporal mean (proportion positive), yielding compact, shift-robust statistics per detector. Implementations include a Numba-accelerated CPU path that computes dilations, biases, and PPV pooling, and an optional output. Outputs are saved per segment as Parquet rows with feature columns named by index and segment length. This design emphasizes reproducibility, computational efficiency, and scale for large PPG cohorts sampled at 32 Hz.

The wearable data is processed in multiple ways to ensure that the high-resolution data can be used for the real-time alarm model. The raw PPG data is preprocessed with our pipeline:

1. Remove duplicates from the data and sort by datetime.
2. Segment data by identifying longer gaps (60s).
3. Resample signal to clean sampling rate (25/32Hz).

• Fill in gaps by cubic spline interpolation.

4. Apply 4th Order Chebyshev Type II Filter 0.5-5Hz

5. Detect motion artifacts (outside STD) from accelerometer data, set to NaN, and interpolate.

Further preprocessing is then applied to be able to feed the data into the segmentation pipeline efficiently:

1. Segmentation to 5-minute snippets

2. Calculation of HRV features using NeuroKit2[64]

a. Time Domain

b. Frequency Domain

c. Non-linear Domain

3. Reduction of 76 features down to 32 by aligning with features used also by e.g., Kubios[65,66]

4. Statistical aggregation in 1-hour and 6-hour (etc.) periods.

Medications and clinical note features: Clinical notes were recorded by nurses and physicians as part of routine care and are completely in German. We optimized the clinical texts by manually determining and replacing common abbreviations, anonymizing personal identifiers, and removing non-informative footers. Medications and fluid balance information from the ICU and general ward were manually harmonized by clinical assistance: measurements and values considered the same or very similar were treated as the same feature. The clinical notes and free-text medications are fed into a flexible, pre-trained, general multi-lingual embedding model (jina-embeddings-v3[27]) from the HuggingFace[67] pre-trained model repository to obtain contextualized embeddings. The clinical notes dimensions were set at 64 dimensions, and the medications at 128 dimensions, after observing improved performance from these settings. These values were selected based on the model's performance in preliminary experiments. The embeddings were then aggregated by taking the element-wise mean into the same segments as the numerical features. Table 4.15 provides an overview of the collected and recorded signals.

Data fusion and segmentation: The data segmentation is a parallelized pipeline that segments the data into time intervals for monitoring across the patient's trajectory (ICU and general ward). It aggregates information for a patient within an adjustable time period, while also truncating the data after the adverse event to prevent data leakage and produce a dataset suitable for training early warning systems (EWS). The feature extraction pipeline is designed to extract features from the segmented data. It encompasses a range of techniques, including statistical analysis, time-series decomposition, and domain-specific feature engineering, to capture the relevant information for each modality. Multi-horizon historical aggregation is performed to capture the temporal dynamics of the data, aggregating the data backwards at 30-minute, 1-hour, 6-hour, 12-hour, and 24-hour intervals.

We are able to flexibly: add new modalities; adjust segmentation periods and add/process new patient timelines, extract a range of adjustable feature sets based on statistical features and retrospective, summarized time-series data; modify prediction

Table D.23: Patient journey described in terms of interval spent in each part. Values shown are the median and [25/75 percentile]. D: days, H: hours. Adapted from Daniela Zuluaga Lotero's Master thesis.

	All Patients (n = 1194)	**No Complication** (n = 813)	**Major Complication** (n = 381)
Preoperative period (D)	1.00 [0.92–1.12]	0.99 [0.89–1.11]	1.02 [0.95–1.15]
Surgery duration (H)	4.46 [2.92–5.92]	4.05 [2.67–5.55]	5.28 [3.90–6.78]
Postoperative period (D)	7.82 [4.97–13.91]	5.88 [4.22–8.28]	17.95 [12.22–27.87]
ICU stay (D)	1.72 [0.88–2.91]	1.00 [0.87–2.13]	2.03 [0.91–3.91]
Ward stay (D)	6.04 [4.08–11.00]	4.99 [3.96–6.94]	14.98 [9.97–24.04]
Length of stay (D)	9.18 [6.17–16.26]	7.17 [5.34–10.16]	20.26 [14.15–30.21]

horizons and exclude patients without certain modalities; aggregate data from periods before the monitoring starts (preoperative, operative, postoperative, ICU). Table 5 shows the resulting number of features per modality. To evaluate the temporal bin for our continuous analysis, data were aggregated into discrete time segments ranging from 7.5 minutes to 6 hours

Outcome

8a Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups The primary endpoint of this study (target major complication) was the occurrence of clinically relevant intra-abdominal complications. In this context, *clinically relevant* referred to complications that required an intervention—surgical, radiologic, or endoscopic. This definition corresponds to complications graded higher than Clavien–Dindo II, the most widely used classification system for postoperative complications among clinicians. Other major complications were defined as complications graded higher than Clavien–Dindo II, excluding surgical site infections grade III (SSI-III), bile leakage, and postoperative pancreatic fistula (POPF). A typical example would be postoperative respiratory failure requiring readmission to the intensive care unit and noninvasive ventilation (Clavien–Dindo IVa). Minor complications were defined as all complications with a Clavien-Dindo Grade less than III (i.e., pneumonia requiring antibiotic therapy). Below is the detailed definition of the primary endpoint.

8b If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors Postoperative complications were identified prospectively by a visceral surgeon and a general physician through daily review of clinical data. To ensure accuracy, a second visceral surgeon independently performed a manual chart review after the availability of final laboratory and microbiological results following the interventions (typically within 2–5 working days). This multi-step process was implemented to guarantee rigorous and unbiased outcome ascertainment. These outcome assessments were conducted by physicians not involved in the treatment of study patients, ensuring objectivity. Wearable devices used to monitor vital signs had no visible display, making the data inaccessible to both patients and treating physicians. This ensured that clinical decisions, such as ordering

CT scans, were based solely on the treating physicians' clinical judgement, entirely independent of the study's monitoring data, preserving unbiased outcome identification. In a subset of cases, a CT or MRI scan was performed prior to any intervention being indicated. The most common triggers for imaging, in descending order, were: highly abnormal laboratory values, Suspicious Fluid in the drainage, the patient's physical condition, and incidental findings on imaging performed for other indications.

8c Report any actions to blind assessment of the outcome to be predicted Outcome assessments were conducted by physicians not involved in the treatment of study patients, ensuring objectivity. Wearable devices used to monitor vital signs had no visible display; the data was inaccessible to both patients and treating physicians. This ensured that clinical decisions, such as ordering CT scans, were based solely on the treating physicians' clinical judgement, entirely independent of the study's monitoring data, preserving unbiased outcome identification.

Predictors

9a Describe the choice of initial predictors (eg, literature, previous models, all available predictors) and any pre-selection of predictors before model building Given the prospective study design, all predictors were determined before patient inclusion. Laboratory values with low completeness and low clinical significance (based on physicians' experience and clinical practice in the hospital) were excluded.

COPRA Modalities (ICU-Related Data) COPRA is identified as an Intensive Care Unit (ICU) data system focused on critical care monitoring and management.

- **COPRA Observations** (copra_observations):
  This dataset contains detailed recordings of patient vital signs and clinical observations in the ICU. It includes metrics such as heart rate, blood pressure, respiratory rate, and other real-time data essential for monitoring critically ill patients.
- **COPRA Scores** (copra_scores):
  This dataset comprises calculated clinical severity and outcome prediction scores for ICU patients. It includes standardized metrics like APACHE or SOFA scores, which are used to assess patient condition and guide clinical decisions.
- **COPRA Fluid Balance** (copra_fluid_balance):
  This dataset records the input and output of fluids for ICU patients. It tracks data critical for managing hydration, kidney function, and overall homeostasis in critical care settings.

ISHMED Modalities (Lab Values/General Information System) ISHMED is a hospital information system dedicated to managing lab values and general patient data.

- **ISHMED Lab Numeric** (ishmed_lab_numeric):
  This dataset includes numerical laboratory results for patients, such as blood test values including glucose levels, hemoglobin, and electrolytes. These diagnostic indicators are collected during hospital stays or outpatient visits.

Textual Modalities These datasets focus on categorized clinical information in text freeform relevant to patient care.

- **Medications** (cat_medications):
  This dataset details medications administered to patients, including drug names, dosages, and administration schedules. It supports analysis of treatment plans and potential drug interactions.
- **Clinical Notes** (cat_clinical_notes):
  This dataset consists of categorized data derived from clinical notes by healthcare providers. It includes patient history, symptoms, diagnoses, and treatment rationales, offering valuable context for machine learning applications.

Wearable Modalities (Activity and Physiological Data) These datasets are sourced from wearable devices monitoring patient activity and physiological signals. Vital signs were continuously measured using wearable devices placed on patients on the first day of transfer to the normal ward, whether directly after surgery or following ICU monitoring. The devices had no visible display, so data was inaccessible to both patients and treating physicians, and they were replaced every 3-4 days based on battery levels.

- **Wearable Activity** (wearable_activity):
  This dataset captures the aggregated precomputed features made available by the manufacturer. It includes derived metrics like heart rate and oxygen saturation.
- **Wearable PPG Feature** (wearable_ppgfeature):
  This dataset contains our custom features that we extracted from photoplethysmography (PPG) signals, measured by wearables, to assess blood volume changes.
- **Wearable PPG Embedding** (wearable_ppgembedding):
  This dataset includes processed PPG data transformed into embeddings or lower-dimensional representations. These are designed for machine learning tasks to identify complex physiological patterns.
- **Wearable Core** (wearable_core):
  This dataset encompasses the CORE device's features, which measure body temperature. It serves as foundational data for further health analysis.

Static Modalities (Retrospective and General Data) These datasets provide static, non-time-series records for baseline or historical patient information.

- **Static Retro ICU** (static_retro_icu):
  This dataset includes retrospective data from ICU stays, covering patient demographics, admission details, and discharge summaries. It provides a historical view of critical care episodes.
- **Static Retro Pre-Intra** (static_retro_pre_intra):
  This dataset contains retrospective data from preoperative and intraoperative periods. It includes surgical details and anesthesia records relevant to patient care before and during procedures.

- **Static Comorbidities and Demographics** (static):
  This dataset encompasses general static patient information, including age, gender, the biological system on which the patient is operated, and medical history. These fixed attributes provide essential baseline data for health studies.

9b Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors) See the description above for a detailed overview of modalities, which are sets of predictors.

Comorbidities were manually extracted from a questionnaire completed during the pre-surgical risk assessment in anesthesiology. A checklist of comorbidities was integrated into REDCap to facilitate the automatic calculation of the Charlson Comorbidity Index (CCI), ensuring consistency in scoring.

An automatic Nutritional Risk Screening (NRS) score calculation was introduced in REDCap, utilizing the patient's height, weight, and any reported recent weight changes extracted from the nutritional assessment conducted either prior to surgery or during hospitalization.

Vital signs were continuously measured using wearable devices placed on patients on the first day of transfer to the normal ward, whether directly after surgery or following ICU monitoring. The devices had no visible display, so data was inaccessible to both patients and treating physicians. They were replaced every 3-4 days based on battery levels.

9c If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors Predictor measurement was based solely on objective clinical data automatically retrieved from the electronic health record. No variables required a subjective assessment.

Sample Size

10 Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation No formal sample size calculation was performed in advance, as the study was designed to include all eligible surgical admissions during the study period. The resulting cohort size was considered sufficient for model training and evaluation, and the number of major complications observed supported an acceptable ratio of events per candidate predictor, reducing the risk of overfitting.

Missing Data

11 Describe how missing data were handled. Provide reasons for omitting any data Numeric variables undergo zero-fill imputation to address missing values, while categorical variables are processed through a two-step approach involving mode imputation to fill gaps, followed by label encoding to convert categories into numerical representations. The system creates missing value indicators for specified variables while

respecting user-defined exclusions, then implements forward-fill imputation to propagate the last observed values forward in time. Any remaining missing values are addressed through zero-fill imputation. We omit data in cases where they are very sparse (missing in 75 percent of the dynamic windows).

We note that our dataset contains considerable amounts of data points per subject. The dataset is, however, sparse over time as we collect different types of data in each stage of the patient's stay. It contains a mix of high-frequency time series, such as ICU and wearable data, as well as events that encapsulate a large amount of information, such as lab results.

One of our primary hypotheses is that wearable data will enhance the additional value, thereby improving prediction quality. Therefore, we excluded patients who had no wearable data. Moreover, we have a limited amount of notes due technical limitations.

Analytical Methods

12a Describe how the data were used (eg, for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements We expand an existing framework built for developing and benchmarking ICU models (see *Reproducibility and Adaptation to Different Clinical Settings*). We also developed a dynamic modelling training framework. We can train the framework with different types of Machine Learning models (XGBoost, GRU, LSTM, Transformer, Logistic Regression). This system produces prediction probabilities for a window of a patient's stay (in our analysis, this has been set at 30 minutes after experiments). Performance is evaluated using 5-fold cross-validation repeated 5 times. This means that we repeat the 5-fold cross-validation experiment five times to ensure a reliable assessment of the performance.

12b Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation)

12c Specify the type of model, rationale, all model-building steps, including any hyperparameter tuning, and method for internal validation In accordance with previous work, we performed runtime preprocessing steps to minimize data leakage between the training and test datasets.

The pipeline operates on three data segments: static patient characteristics, dynamic time-series measurements, and outcome labels. It processes training, validation, and test splits consistently using fitted transformations from the training data.

For static features, missing value indicators are generated across all static variables, followed by scaling and normalization procedures. Numeric variables undergo zero-fill imputation to address missing values, while categorical variables are processed through a two-step approach involving mode imputation to fill gaps, followed by label encoding to convert categories into numerical representations. The system creates missing value indicators for specified variables while respecting user-defined exclusions, then implements forward-fill imputation to propagate the last observed values forward in

time. Any remaining missing values are addressed through zero-fill imputation, and the pipeline concludes by generating comprehensive historical aggregate features, including minimum, maximum, count, and mean values computed over time windows to capture temporal patterns in the data. The system supports recipe caching for transfer learning scenarios, allowing preprocessing pipelines to be saved and reloaded. All transformations are fitted on training data only and consistently applied to validation and test sets. This preprocessing approach addresses common challenges in data, including irregular sampling, missing values, and the need to combine static patient characteristics with time-varying physiological measurements while maintaining temporal relationships and preventing data leakage.

ML architectures: Several architectures were used: Deep learning (Transformers[23], GRU[24], TCN[25]), Gradient Boosted Trees (GBT: XGBoost[20], LightGBM[21], CatBoost[22]), Random Forest, and Logistic Regression. Adding more model architectures is easily possible due to the framework's flexibility.

We selected the final classifier after thorough benchmarking, which aligns with earlier work indicating that boosted tree ensembles are a strong state-of-the-art classifier and often outperform Deep Learning models, such as Transformers[51,68], especially when domain-guided feature extraction is employed.

This system produces probabilities for a temporal window of a patient's stay. Each experiment was performed as a 5-times repeated 5-fold cross-validation to ensure the performance is balanced over the whole collected dataset. The test set was defined a 20 percent of subjects. The validation set was defined as 20% of the training subset. The patient timelines were split as a whole to ensure that data leakage between time windows could not occur (any patient window in the train set was not in the test set). Hyperparameter optimization was performed using predefined, publicly accessible configuration files. The traditional machine learning models were tested over 100 iterations of 5-fold cross-validation in a manner that the test set contained a unique subset of subject IDs each time. The Deep Learning models were instead trained with 50 iterations of tuning. Model calibration was performed, but needs to be redone when training the final deployment model for the population.

12d Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (eg, hospitals, countries). See TRIPOD-Cluster for additional considerations We performed a single-center study, so this is not applicable.

12e Specify all measures and plots used (and their rationale) to evaluate model performance (eg, discrimination, calibration, clinical utility) and, if relevant, to compare multiple models AUROC (Area Under Receiver Operating Characteristic Curve): useful for overall accuracy as it evaluates both sensitivity (TP complications/ TP+FN all real complications) and specificity (TN no complications/ TN+FP all predicted non-complications), providing an overall measure of the model's accuracy in identifying true complications and avoiding FP.

AUPRC (Area Under Precision-Recall Curve): useful for handling imbalanced data (complications are relatively rare) as it evaluates both precision (TP complications/

TP+FP all predicted complications) and recall (TP complications/ TP+FN all real complications), indicating the model's ability to identify rare complications without generating too many alarms (FP).

SHAP (SHapley Additive exPlanations) values (Feature Importance): to identify which features (lab values, vital signs) are most important in the model's predictions, providing insights for clinical relevance of predictors for detecting complications.

A comparison of modalities is presented in the study to determine which modalities are most important to include when developing a prediction model for this endpoint.

12f Describe any model updating (eg, recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings Our model was trained specifically for the setup in our clinic and can be adapted for use in other hospitals or clinical settings. Due to the dynamic nature of the model and the windowed prediction setup, the prevalence is low, at around 0.006, resulting in low probabilities for the model. Recalibration, however, is not needed in our experimental setting as we employ an alarming system setup that uses the predictions as input and can be tuned to different prevalences.

12g For model evaluation, describe how the model predictions were calculated (eg, formula, code, object, application programming interface) Code to preprocess the data on any computer or high-performance cluster can be found at:

`https://github.com/rvandewater/CASHEWS`

Code to train the models can be found at `https://github.com/rvandewater/YAIB`

Libraries used to preprocess data and generate predictions can be found within the respective code bases.

Class Imbalance

13 If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions The underlying model types that were used, Gradient Boosted Trees, are robust to class imbalance, and their performance decreased when oversampling was employed.

Fairness

14 Describe any approaches that were used to address model fairness and their rationale Age-based subgroup analysis (possibly oversampling/weighting female data). An abblation for the different surgery systems was performed

Model Output

15 Specify the output of the prediction model (eg, probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified The outputs of the online model were used for an alarm system. This means

the entire stay duration every hour, though this is adjustable to a custom amount of time segments, we predict the likelihood of the complication. The end of the prediction period is either the discharge or the main endpoint. This was a binary prediction for the Early online (hourly) prediction of SSI-3, Pancreatic Fistula, Bile-leakage complications. We used an adjustable 6-hour time horizon (and experimented with shorter and longer prediction horizons).Output: Probability score (0 = likely no complication, 1 = likely complication), hourly predicting the likelihood of SSI-3 complication in the next horizon period.

Training versus Evaluation

16 Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors No differences between the development and evaluation datasets; the data was randomly split over multiple experiments, and we ensured no information leakage.

Ethical Approval

17 Name the institutional research board or ethics committee that approved the study and describe the participant informed consent or the ethics committee waiver of informed consent The ethics committee of the Charité University Hospital approved study code EA4/198/21 from 20.10.2021.

D.6.5 Open Science

Funding

18a Give the source of funding and the role of the funders for the present study Funded by the "Gemeinsamer Bundesausschuss (G-BA) Innovationsausschuss" in the framework of "CASSANDRA – Clinical ASSist AND aleRt Algorithms" (project number 01VSF20015).

Conflicts of Interest

18b Declare any conflicts of interest and financial disclosures for all authors The study and the authors were partially employed/funded by "Gemeinsamer Bundesausschuss (G-BA) Innovationsausschuss".

Protocol

18c Indicate where the study protocol can be accessed or state that a protocol was not prepared The protocol can be retrieved upon request of the main clinician involved in the clinical trial (axel.winter@charite.de)

Registration

18d Provide registration information for the study, including register name and registration number, or state that the study was not registered

The study was registered in the German Clinical Trial Register (DRKS - Deutsches Register Klinischer Studien) under the study ID DRKS00034973. The link to this can be found here: https://drks.de/search/de/trial/DRKS00034973 [Accessed 2025-07-10].

Data sharing

18e Provide details of the availability of the study data The dataset used in this study is restricted to researchers involved in the ongoing study, as these are the conditions under which participants signed informed consent.

Code Sharing

18f Provide details of the availability of the analytical code Extensive modular and reusable packages and code for preprocessing the datasets and preparing it for ML can be found at:

https://github.com/rvandewater/CASHEWS.

An easy-to-use and flexible prediction framework to train the models can be found at:

https://github.com/rvandewater/YAIB/tree/cass.

Patient and Public Involvement

19 Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement There was no public involvement in the trial. However, patients were informed that they would contribute to scientific research by participating in it.

D.6.6 Result

Participants

20a Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful A total of 1,285 patients were prospectively enrolled in the study after providing written informed consent before the surgery, either in the outpatient clinic or on the surgical ward. These patients were consecutively identified upon admission for a planned major surgery, which fell within the predefined list of eligible surgical procedures.

Exclusions were applied as follows:

- Major surgery not performed (n=69): Patients who were scheduled for a major operation but ultimately did not undergo the procedure.
- Withdrawal from study (n=18): Patients who withdrew from the study after enrollment, either declining to continue participation or refusing to wear the required monitoring devices.
- Outcome adjudication (n=4): Patients who were excluded during the outcome adjudication process due to ambiguity in defining the presence or timing of postoperative complications. Their clinical trajectory was marked by different complications, and the censoring time point was not precisely defined.
- Lacking signal quality or amount (n=58): after retrieving, cleaning and analyzing the signal quality, if there was no PPG data at all, patients were excluded. For example, patients did not wear the device properly, which prevented it from getting a reading.

After exclusions, the final cohort included 1,136 patients for model development and evaluation. Patients were classified into four mutually exclusive groups based on the presence, severity, and timing of postoperative complications identified during their hospital stay. Patients are censored after reaching their first major endpoint and after a clinical prediction timeline of 20 days. Additionally, patients are censored after experiencing a major complication different from our endpoint.

The *target complications* are the most common highly granular complications after major abdominal surgery of the upper and lower gastrointestinal tract, major hepatobiliary and pancreatic surgery, are deep organ space infections (SSI-III), bile leakages, and pancreatic fistulas. These endpoints were combined because they all represent intra-abdominal pathologies associated with fluid collections that required intervention. These endpoints were subsumed under SSI-III for clinical reasons, reflecting everyday clinical practice.

Patients were classified into four mutually exclusive groups based on the presence, severity, and timing of postoperative complications identified during their hospital stay. Complications were first identified prospectively and later confirmed through structured chart review using strict clinical definitions based on the Clavien-Dindo classification classification:

- Group 1 – *No Complications:* Patients with no documented postoperative complications of any kind throughout their hospital stay.
- Group 2 – *Only Minor Complications (CD I&II):* Patients who had only minor complications during their stay—those not requiring significant interventions (e.g., conservative treatment only)—and no major complications at any time.
- Group 3 – *Target Major Complication First (CD III & IV)*: Patients who experienced at least one major complication, with the first major event matching the predefined target complication (e.g., intra-abdominal infection, anastomotic leak). This group reflects early clinical deterioration directly related to the study outcome of interest.

- Group 4 – *Other Major Complication First (CD III & IV):* Patients who also experienced a major complication, but their first major event was a different complication type than the predefined target. This group helps identify alternate deterioration patterns and competing risk events.

In cases where patients experienced both minor and major complications, the classification was based on the first major complication, as it marks the most clinically relevant point of deterioration. This logic ensures stratification by both severity and temporal sequence, aligning with the intended evaluation of early warning performance.

20b Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups The statistics of key periods and indicators of all patients can be found in Table D.23.

20c For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome) We used 5 times 5-fold repeated cross-validation such that the train/test split changes 5 times and the train/validation split 5 times as well resulting in 25 individual experiments.

Important predictors are shown in Figure 4.15.

Model Development

21 Specify the number of participants and outcome events in each analysis (eg, for model development, hyperparameter tuning, model evaluation) 1,194 patients are left after basic quality checks. Then, after selecting for wearable signal, we are left with 1,136 patients who were used for model development, hyperparameter tuning, and model evaluation. As described above, we use stratified repeated 5-fold cross-validation to quantify uncertainty.

Model Specification

22 Provide details of the full prediction model (eg, formula, code, object, application programming interface) to allow predictions in new individuals and to enable third-party evaluation and implementation, including any restrictions to access or reuse (eg, freely available, proprietary) A flexible prediction framework for training models, external validation, and evaluation can be found at:

https://github.com/rvandewater/YAIB/tree/cass.

The preprocessing code is available here:

https://github.com/rvandewater/CASHEWS.

Model Performance

23a Report model performance estimates with confidence intervals, including for any key subgroups (eg, sociodemographic). Consider plots to aid presentation Refer to the main paper for the differences in prediction between the ICU and general wards. Subgroup analysis was attempted, but it is challenging due to the limited sample size when performing 5-fold cross-validation.

23b If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD-Cluster for additional details‡ The data for the study were collected at a single clinic, making a traditional cluster analysis unnecessary. Ablations for specific sub-groups are reported in the paper. See the answer to the above question.

Model Updating

24 Report the results from any model updating, including the updated model and subsequent performance The model that is produced with our framework is low-cost to train and can be trained on commodity hardware. Retraining for the non-deep learning approach can be easily accomplished by adding data to the dataset and redoing hyperparameter tuning. Model updates should be done when more insights become available, as well as when clinical protocols are adjusted.

D.6.7 Discussion

Interpretation

25 Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies We demonstrate a high median area under the receiver operator characteristic (AUROC) of 0.90 (0.89-0.91) for the detection of surgery-related infections. Complications could be predicted 9 hours in advance with only a minor reduction of the AUROC: 0.89 (0.88-0.89).

Limitations

26 Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalisability This study has several limitations that may affect the generalizability of our findings. Despite being the largest reported cohort in surgical research using continuous vital sign monitoring, larger cohorts could enable more adaptable models. Moreover, as a single-center study conducted within our institution's Enhanced Recovery After Surgery (ERAS) program, the results reflect the implementation of strict, standardized postoperative procedures and an institutional culture with exceptionally low failure-to-rescue rates. While these factors contribute to optimal patient outcomes, they may limit the applicability of our findings to centers with different postoperative care standards or complication management approaches. Additionally, the standardized nature of our protocols, though beneficial for internal validity, may not reflect the variability in clinical practice encountered across different healthcare settings.

Usability of the Model in the Context of Current Care

27a Describe how poor quality or unavailable input data (eg, predictor values) should be assessed and handled when implementing the prediction model

The modeling/ preprocessing pipeline is robust to missing data and will automatically impute data, ensuring that the impact of poor-quality or unavailable input data is reduced. The entire pipeline is flexible such that missing datasets will still result in an ML-ready dataset. Evaluation of the logs produced can help the user or ML engineer determine if the data for a particular patient lacks sufficient information.

27b Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users A data pipeline needs to be set up to feed data into the model for inference. The current preprocessing pipeline offers an extensible interface for this purpose. A more flexible ward movement should transfer high-risk patients to a more heavily monitored ward. Diagnostic tools, such as CT scans, can be taken in a more targeted manner. The downstream model interface has not yet been conclusively determined.

27c Discuss any next steps for future research, with a specific view to applicability and generalisability of the model Despite the promising technical performance of multimodal continuous models, their utility can only be conclusively established through the evaluation of patient-centered outcomes: length of stay, comprehensive complication index, readmission rates, and mortality statistics. Future research is key to providing objective evidence of the system's meaningful improvement in healthcare delivery and justifying its implementation in clinical practice. Comparing the data quality or frequency from ICU monitoring versus wearable monitoring (e.g., missing data rates, data granularity) to assess if ICU data provides advantages over wearables in capturing early signs of SSI-III. Finally, propensity matching to match patients on baseline characteristics (static data) to minimise confounding factors and see whether monitoring setting alone affects prediction difficulty.

Appendix E

Supplements Chapter 5

E.1 Data Description

The Mount Sinai Data Warehouse (MSDW) dataset is stored in an Observational Medical Outcomes Partnership (OMOP) database on Mount Sinai Data Warehouse (MSDW). The data is taken from the the SAP HANA instance on AIR·MS.

Table E.1: Description of the MSHS dataset. We visualize the OMOP tables, record types, distinct patients, and record counts. Source: `https://labs.icahn.mssm.edu/msdw/data-sources/` [accessed 2025-07-01].

OMOP Table	Record Type	Distinct Patients	Record Count
person	Patient Demographics	12,613,207	12,613,207
death	Patient Date of Death	58,189	58,189
visit_occurrence	Mobile Unit Encounter	108,709	186,228
visit_occurrence	Inpatient Hospitalization from ED Visit	350,339	687,350
visit_occurrence	Hospital Outpatient Visit	1,077,260	3,191,746
visit_occurrence	Urgent Care Visit	7,214	7,910
visit_occurrence	ED Visit	1,392,155	3,518,199
visit_occurrence	Inpatient Hospitalization	698,296	1,045,651
visit_occurrence	Outpatient Visit	4,887,380	100,301,788
visit_occurrence	Telehealth Visit	844,573	4,048,646
visit_occurrence	Chart Documentation Event	6,418,280	124,955,281
condition_occurrence	Hospital Problem	995,145	4,311,335
condition_occurrence	Encounter Diagnosis	4,524,838	135,895,977
condition_occurrence	Problem List	2,545,781	14,529,514
condition_occurrence	Billing Diagnosis	3,026,875	86,859,811
measurement	Vital Signs	4,121,903	749,765,563
measurement	Flowsheet Measurement	2,024,535	278,407,737
measurement	Lab Component Result	4,475,629	1,191,858,706
drug_exposure	Outpatient Medication Order with Dispense	29,766	299,810
drug_exposure	Medication Order with Administration	1,777,567	148,207,486
drug_exposure	Immunization Administration	1,224,530	8,974,990
drug_exposure	Immunization from Procedure Order	866,077	4,017,133
drug_exposure	Outpatient Medication Order	2,964,849	53,607,610
drug_exposure	Medication Order with Dispense	1,076,397	17,403,808
drug_exposure	Immunization from Medication Order	241,680	314,829
drug_exposure	Medication Order	1,165,193	14,821,984
procedure_occurrence	Procedure Inferred from Flowsheet	25,261	298,534
procedure_occurrence	General Procedure	5,636,156	368,567,851
procedure_occurrence	Surgical Procedure	407,953	1,612,374
observation	Patient Demographics	12,613,206	67,395,891
observation	Family History	1,693,268	76,237,615
observation	Allergy	3,742,976	4,563,458
observation	Flowsheet Observation	2,458,935	127,019,929
observation	Social History	3,631,735	162,925,262
observation	Past Medical History	2,418,320	9,212,435

observation	Surgical History Procedure	1,742,624	3,918,654
note	Lab Component Result Note	2,697,131	18,688,874
note	Clinical Note	5,312,477	224,136,573
provider	Provider		1,372,008
provider_attribute_xtn	Physician specialty		606,877
provider_attribute_xtn	Personnel specialty		180,426
location	Patient Address		12,613,206
location	Care Site Address		16,126
location	Provider Address		1,372,008
care_site	Service Area		51
care_site	Care Area		844
care_site	Location		2,222
care_site	Department		6,526
care_site	Bed		83,723
care_site	Place of Service		7,274
care_site	Room		7,936
concept	EPIC DIAGNOSIS		1,594,594
concept	EPIC ALLERGEN		40,964
concept	EPIC VISIT TYPE		12,150
concept	EPIC SURGICAL PROCEDURE		42,040
concept	EPIC PROCEDURE		196,308
concept	EPIC IMMUNIZATION		543
concept	EPIC FLOWSHEET MEASURE		99,422
concept	EPIC MEDICATION		145,148
concept	EPIC FLOWSHEET TEMPLATE		4,102
concept	EPIC LAB COMPONENT		42,307
vocabulary	MSDW Vocabulary		7
vocabulary	Epic Vocabulary		118
vocabulary	Vocabulary		104
vocabulary	PowerPath Vocabulary		8

E.2 Task Definitions

Listing E.1: AUMCdb MEDS configuration predicates defined in the syntax of ACES for 8 endpoints. Defined for the extended fork of *MEDS-DEV* at `https://github.com/rvandewater/MEDS-DEV`

```
# Made by Robin van de Water (rvandewater) largely based on https://github.com/AmsterdamUMC/AmsterdamUMCdb/blob/2901fdda2ff8f7ae0f16191d0b3f6d0ac60f70b6/amsterdamumcdb/sql/common/legacy/
predicates:
  icu_admission:
    code: { regex: "^ICU_ADMISSION//.*" }
  icu_discharge:
    code: { regex: "^ICU_DISCHARGE//.*" }
  hospital_admission:
    code: { regex: "^ICU_ADMISSION//.*" }
  hospital_discharge:
    code: { regex: "^ICU_DISCHARGE//.*" }

  # Elevated creatinine task predicates
  creatinine_1:
    code: MEASURE//Kreatinine (bloed)//umol/l
  creatinine_2:
    code: MEASURE//Kreatinine//umol/l
  # Todo: convert to mg/dL
  abnormally_high_creatinine_1_umol_l:
    code: MEASURE//Kreatinine (bloed)//umol/l
    value_min: 115 # 1.3 mg/dL
    value_min_inclusive: False
    value_max: null
  abnormally_high_creatinine_2_umol_l:
    code: MEASURE//Kreatinine//umol/l
    value_min: 115 # 1.3 mg/dL
    value_min_inclusive: False
```

```
    value_max: null
  creatinine:
    expr: or(creatinine_1, creatinine_2)
  abnormally_high_creatinine_umol_l:
    expr: or(abnormally_high_creatinine_1_umol_l,
        abnormally_high_creatinine_2_umol_l)

  # Hyponatremia task predicates
  sodium:
    code: MEASURE//Natrium (bloed)//mmol/l # mmol/L = mEq/L for sodium

  abnormally_low_sodium:
    code: MEASURE//Natrium (bloed)//mmol/l
    value_min: null
    value_max: 135 # mEq/L
    value_max_inclusive: False

  # Metabolic Acidosis task predicates (not supported yet, only negative labels)
  # bicarbonate_1:
  #   code: MEASURE//Act.HCO3 (bloed)//mmol/l
  # bicarbonate_2:
  #   code: MEASURE//HCO3//mmol/l
  # abnormally_low_bicarbonate_1:
  #   code: MEASURE//Act.HCO3 (bloed)//mmol/l
  #   value_min: null
  #   value_max: 22 # mEq/L
  #   value_max_inclusive: False
  # abnormally_low_bicarbonate_2:
  #   code: MEASURE//HCO3//mmol/l
  #   value_min: null
  #   value_max: 22 # mEq/L
  #   value_max_inclusive: False
  # bicarbonate:
  #   expr: or(bicarbonate_1, bicarbonate_2)
  # abnormally_low_bicarbonate:
  #   expr: or(abnormally_low_bicarbonate_1, abnormally_low_bicarbonate_2)

  # Hemoglobin (mmol/L)=Hemoglobin (g/dL)x0.6206 = 13.0 g/dL x 0.6206 = 8.07 mmol/L
  hemoglobin_1:
    code: MEASURE//Hemoglobine//mmol/l
  hemoglobin_2:
    code: MEASURE//Hb (bloed)//mmol/l
  hemoglobin_3:
    code: "MEASURE//Hb(v.Bgs) (bloed)//mmol/l"
  abnormally_low_hemoglobin_1:
    code: MEASURE//Hemoglobine//mmol/l
    value_min: null
    value_max: 8 # g/dL
    value_max_inclusive: False
  abnormally_low_hemoglobin_2:
    code: "MEASURE//Hb (bloed)//mmol/l"
    value_min: null
    value_max: 8 # g/dL
    value_max_inclusive: False
  abnormally_low_hemoglobin_3:
    code: "MEASURE//Hb(v.Bgs) (bloed)//mmol/l"
    value_min: null
    value_max: 8 # g/dL
    value_max_inclusive: False
  hemoglobin:
    expr: or(hemoglobin_1, hemoglobin_2, hemoglobin_3)
  abnormally_low_hemoglobin:
    expr: or(abnormally_low_hemoglobin_1, abnormally_low_hemoglobin_2,
        abnormally_low_hemoglobin_3)

  # Leukocytosis task predicates
  wbc_1:
```

```
    code: MEASURE//Leucocyten//10^9/l
  wbc_2:
    code: MEASURE//Leuco's (bloed)//10^9/l
  abnormally_high_wbc_1:
    code: MEASURE//Leucocyten//10^9/l
    value_min: 11 # K/uL
    value_min_inclusive: False
    value_max: null
  abnormally_high_wbc_2:
    code: MEASURE//Leuco's (bloed)//10^9/l
    value_min: 11 # K/uL
    value_min_inclusive: False
    value_max: null
  wbc:
    expr: or(wbc_1, wbc_2)
  abnormally_high_wbc:
    expr: or(abnormally_high_wbc_1, abnormally_high_wbc_2)

  # Thrombocytopenia task predicates
  platelets_1:
    code: MEASURE//Thrombo's (bloed)//10^9/l
  platelets_2:
    code: MEASURE//Thrombocyten//10^9/l
  abnormally_low_platelets_1:
    code: MEASURE//Thrombo's (bloed)//10^9/l
    value_min: null
    value_max: 150 # K/uL
    value_max_inclusive: False
  abnormally_low_platelets_2:
    code: MEASURE//Thrombocyten//10^9/l
    value_min: null
    value_max: 150 # K/uL
    value_max_inclusive: False
  platelets:
    expr: or(platelets_1, platelets_2)
  abnormally_low_platelets:
    expr: or(abnormally_low_platelets_1, abnormally_low_platelets_2)

  # Hypotension task predicates
  map_1:
    code: MEASURE//ABP gemiddeld//mmHg
  map_2:
    code: MEASURE//Niet invasieve bloeddruk gemiddeld//mmHg

  abnormally_low_map_1:
    code: MEASURE//ABP gemiddeld//mmHg
    value_min: null
    value_max: 65 # mmHg
    value_max_inclusive: False
  abnormally_low_map_2:
    code: MEASURE//Niet invasieve bloeddruk gemiddeld//mmHg
    value_min: null
    value_max: 65 # mmHg
    value_max_inclusive: False
  map:
    expr: or(map_1, map_2)
  abnormally_low_map:
    expr: or(abnormally_low_map_1, abnormally_low_map_2)

  # Hyperkalemia task predicates
  potassium:
    code: MEASURE//Kalium (bloed)//mmol/l
  abnormally_high_potassium:
    code: MEASURE//Kalium (bloed)//mmol/l
    value_min: 5.0 # mmol/L
    value_min_inclusive: True
    value_max: null
```

```yaml

  # Hypoglycemia task predicates
  glucose:
    code: MEASURE//Glucose (bloed)//mmol/l
  abnormally_low_glucose:
    code: MEASURE//Glucose (bloed)//mmol/l
    value_min: null
    value_max: 70.2 # mg/dL
    value_max_inclusive: True
```

Listing E.2: Hyperkalemia prediction task configuration in the ACES syntax applicable to 8 different datasets. Defined for *MEDS-DEV* at `https://github.com/rvandewater/MEDS-DEV`

```yaml
metadata:
  description: >-
    Identify patients who develop hyperkalemia (high potassium levels) after 24
        hours of hospital admission using the first 24 hours of the admission,
        excluding those with prior hyperkalemia or children under 18.
  contacts:
    - name: "Robin van de Water"
      github_username: "rvandewater"
  supported_datasets:
    - MIMIC-IV
    - AUMCdb
    - INSPIRE
    - NWICU
    - HIRID
    - EHRSHOT
    - eICU
    - MSHS

predicates:
  hospital_admission: ???
  hospital_discharge: ???
  birth:
    code: { regex: "MEDS_BIRTH.*" }
  death:
    code: { regex: "MEDS_DEATH.*" }
  discharge_or_death:
    expr: or(death, hospital_discharge)

  #  Plain predicates
  potassium:
    code: ???
  # The following threshold is for potassium in mmol/L but can be adjusted in the
      dataset predicates
  abnormally_high_potassium:
    code: ???
    value_min: 5.0 # mmol/L
    value_min_inclusive: False
    value_max: null
  # This is if the dataset contains diagnosis codes (ICD for example) for
      hyperkalemia
  diagnosed_hyperkalemia:
    code: ???
  hyperkalemia_combined:
    expr: or(diagnosed_hyperkalemia, abnormally_high_potassium)

# Trigger event for the task
trigger: hospital_admission

windows:
  # Input window: first 24 hours after hospital admission
  input:
    start: null
    end: trigger + 24h
    start_inclusive: True
```

```
    end_inclusive: True
    index_timestamp: end
    # do not allow to have shown elevated values previously
    has:
      hyperkalemia_combined: (None, 0)
  # Do not allow hospital discharge to happen within the input + target window
  no_discharge:
    start: trigger
    end: start + 24h
    start_inclusive: False
    end_inclusive: True
    has:
      hospital_admission: (None, 0)
      discharge_or_death: (None, 0)
  # Do not allow to include children (age <= 18) patients
  no_children:
    start: end - 6570d # 18 years = 365 * 18 days
    end: trigger
    start_inclusive: True
    end_inclusive: True
    has:
      birth: (None, 0)
  target:
    start: input.end
    end: start + 24h
    start_inclusive: False
    end_inclusive: True
    has:
      # Has at least one measurement of potassium in the target window
      potassium: (1, None)
    label: hyperkalemia_combined
```

E.3 Hyperparameters

Table E.2: Hyperparameters for MEDS-Tab (Tiny) configuration.

Hyperparameter	Value
Tabularization Configuration	
Aggregation Functions	code/count, value/sum
Window Sizes	7d, 30d
Max Included Codes	100
Min Code Inclusion Count	null
Feature Engineering	
Static Features	Enabled
Time-Series Features	Enabled
Model Training	
Model Type	XGBoost
Hyperparameter Optimization Trials	10
Parallelization	
Number of Workers	1
Hydra Launcher	joblib
Sweeper Jobs	1
Data Processing	
Do Overwrite	False
Cache Tabularized Data	True
Cache Label Data	True

Table E.3: Hyperparameters for GenHPF model configuration.

Hyperparameter	Value
Model Architecture	
Model Name	genhpf_predictor
Structure	hierarchical
Embedding Method	text
Vocabulary Size	28,996
Tokenization	
Pad Token ID	0
Separator Token ID	102
Dummy Token ID	101
Ignore Index	-100
Optimization	
Learning Rate	1×10^{-4}
Max Epochs	200
Random Seed	42
Loss Function	
Criterion	Binary Cross Entropy
Task Type	meds_binary_classification
Number of Labels	1
Data Loading	
Data Format	MEDS
Batch Size	16
Number of Workers	6
Train Subset	train
Validation Subset	tuning
Test Subset	held_out
Checkpointing & Early Stopping	
Best Checkpoint Metric	AUROC
Maximize Metric	True
Save Interval	1 epoch
Keep Last Epochs	5
Early Stopping Patience	10
Distributed Training	
Distributed World Size	1
Find Unused Parameters	False
Logging	
Log Format	tqdm
Log Interval	10

Table E.4: Hyperparameters for CEHR-BERT pretraining.

Hyperparameter	Value
Architecture	
Hidden Size	768
Number of Hidden Layers	8
Max Position Embeddings	1,024
Vocabulary Size	100,000
Tokenization	
Min Frequency	100
Min Number of Tokens	10
MEDS Configuration	
Attention Function Type	cehr_bert
Inpatient Attention Function Type	mix
MEDS to CEHR-BERT Conversion	MedsToCehrbertOMOP
Include Auxiliary Token	True
Include Demographic Prompt	False
Include Value Prediction	False
Optimization	
Learning Rate	5×10^{-5}
Number of Epochs	10
Warmup Steps	500
Weight Decay	0.01
Batching	
Per Device Train Batch Size	4
Per Device Eval Batch Size	16
Gradient Accumulation Steps	2
Effective Batch Size	8
Data Loading	
Preprocessing Workers	10
Preprocessing Batch Size	1,024
Dataloader Prefetch Factor	2
Validation Split Percentage	0.1
Evaluation & Saving	
Evaluation Strategy	epoch
Save Strategy	epoch
Eval Accumulation Steps	10
Load Best Model at End	True
Metric for Best Model	eval_loss
Greater is Better	False

Table E.5: Hyperparameters for CEHR-BERT finetuning.

Hyperparameter	Value
Model Configuration	
Finetune Model Type	pooling
Do Train	True
Do Predict	True
LoRA Configuration	
Use LoRA	False
LoRA Rank	64
LoRA Alpha	16
Target Modules	query, value
LoRA Dropout	0.1
Architecture	
Vocabulary Size	100,000
Min Frequency	100
Min Number of Tokens	10
MEDS Configuration	
Attention Function Type	cehr_bert
Inpatient Attention Function Type	mix
MEDS to CEHR-BERT Conversion	MedsToCehrbertOMOP
Include Auxiliary Token	True
Include Demographic Prompt	False
Include Value Prediction	False
Optimization	
Learning Rate	5×10^{-5}
Number of Epochs	10
Warmup Steps	500
Weight Decay	0.01
Batching	
Per Device Train Batch Size	16
Per Device Eval Batch Size	16
Gradient Accumulation Steps	2
Effective Batch Size	32
Data Loading	
Preprocessing Workers	10
Preprocessing Batch Size	1,024
Dataloader Prefetch Factor	2
Validation Split Percentage	0.1
Evaluation & Saving	
Evaluation Strategy	epoch
Save Strategy	epoch
Eval Accumulation Steps	10
Load Best Model at End	True
Metric for Best Model	eval_loss
Greater is Better	False
Logging	
Logging Steps	10

Table E.6: Hyperparameters for CEHR-XGPT model pretraining.

Hyperparameter	Value
Architecture	
Hidden Size	768
Number of Layers	14
Max Position Embeddings	8,192
Attention Function Type	day
Inpatient Attention Function Type	day
Training	
Learning Rate	1×10^{-4}
Weight Decay	0.01
Warmup Ratio	0.01
Number of Epochs	50
Early Stopping Threshold	0.001
Random Seed	42
Batching & Data Loading	
Max Tokens per Batch	16,384
Sample Packing	Enabled
Dataloader Workers	8
Prefetch Factor	8
Evaluation & Saving	
Evaluation Strategy	epoch
Save Strategy	epoch
Load Best Model at End	True

Table E.7: Hyperparameters for MOTOR (FEMR) model configuration.

Hyperparameter	Value
Transformer Architecture	
Hidden Size	768
Number of Layers	11
Number of Attention Heads	12
Intermediate Size	3,072
Attention Width	496
Hidden Activation	SwiGLU
Vocabulary Size	10,126
Use Bias	False
Is Hierarchical	True
Use Normed Ages	True
Task Configuration	
Task Type	motor
Final Layer Size	512
Torch Dtype	float32
Training & Batching	
Per Device Train Batch Size	1
Per Device Eval Batch Size	1
Gradient Accumulation Steps	16
Effective Batch Size	16
Dataloader Workers	8
Prefetch Factor	2
Pin Memory	True
Early Stopping	
Early Stopping Patience	1
Early Stopping Threshold	0.001
Metric for Best Model	eval_loss
Greater is Better	False
Evaluation & Saving	
Evaluation Strategy	epoch
Save Strategy	epoch
Load Best Model at End	True

E.4 Extended Results

E.4.1 MEDS-DEV



Figure E.1: Performance metrics for the subjects weighted benchmarking *MEDS-DEV* compatible models. We perform the benchmarks on a large range of datasets (left) and tasks (right). Left column: AUROC; Right column AUPRC.

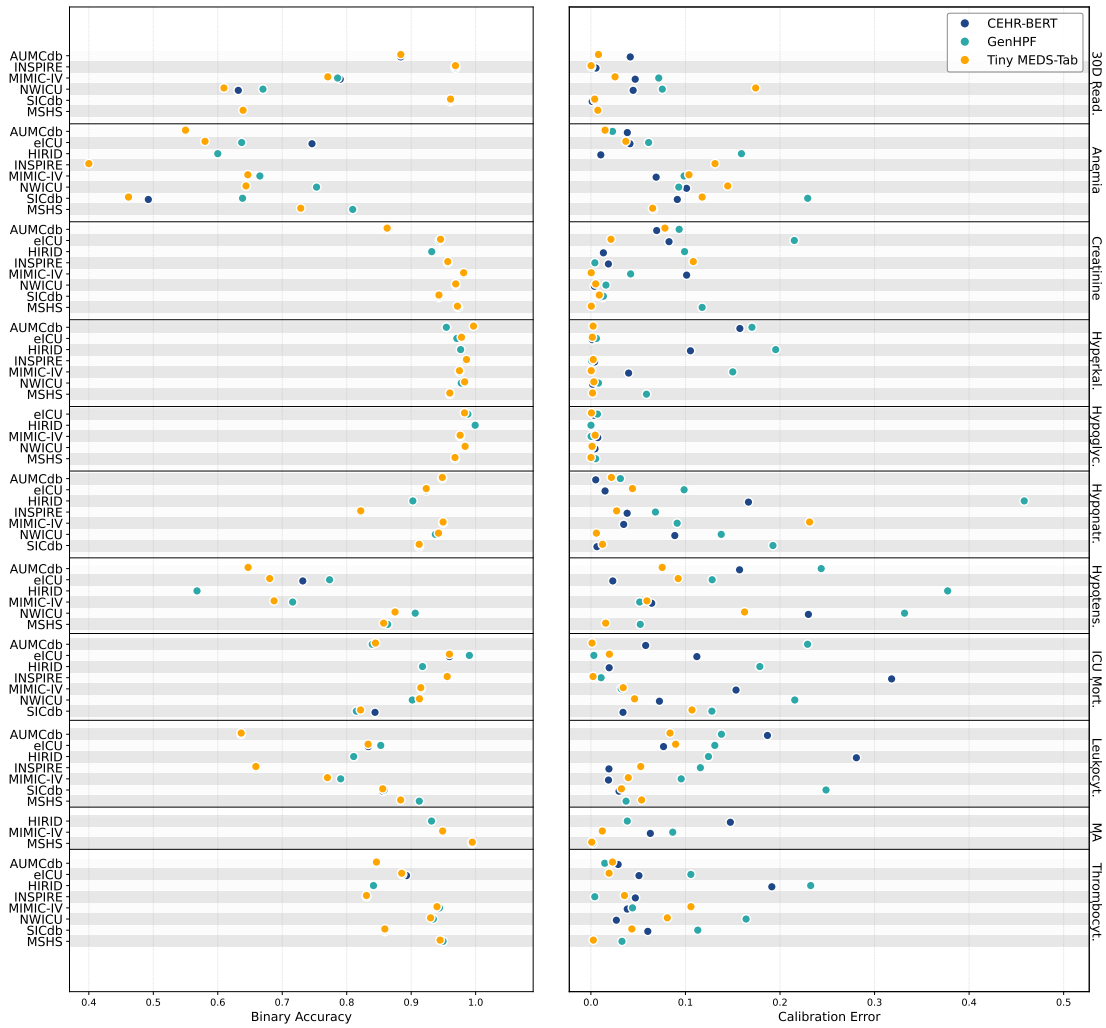


Figure E.2: Performance metrics for the subjects weighted benchmarking MEDS-DEV compatible models. We perform the benchmarks on a large range of datasets (left) and tasks (right). Left column: binary accuracy; Right column calibration error.

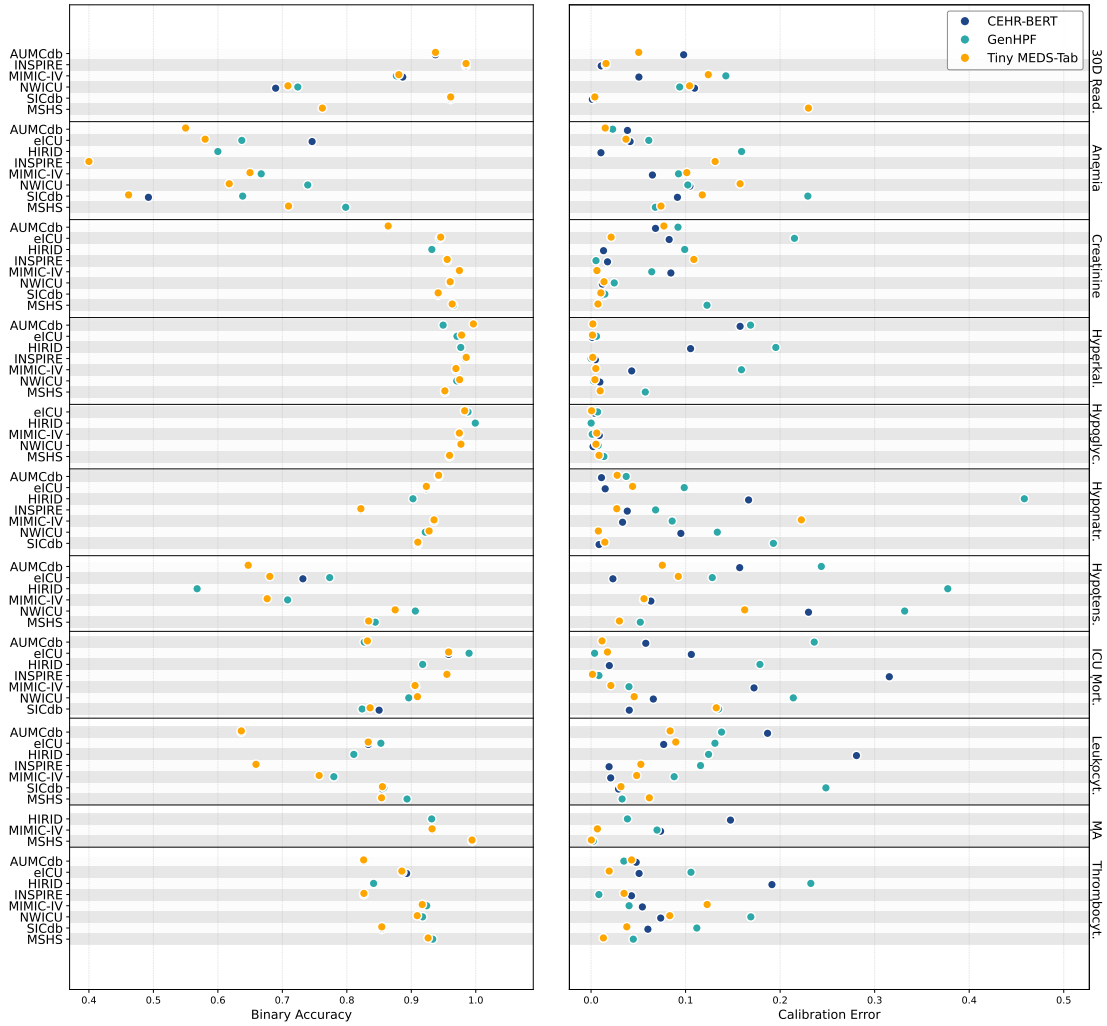


Figure E.3: Performance metrics for the subjects weighted benchmarking MEDS-DEV compatible models. We perform the benchmarks on a large range of datasets (left) and tasks (right). Left column: binary accuracy; Right column calibration error.

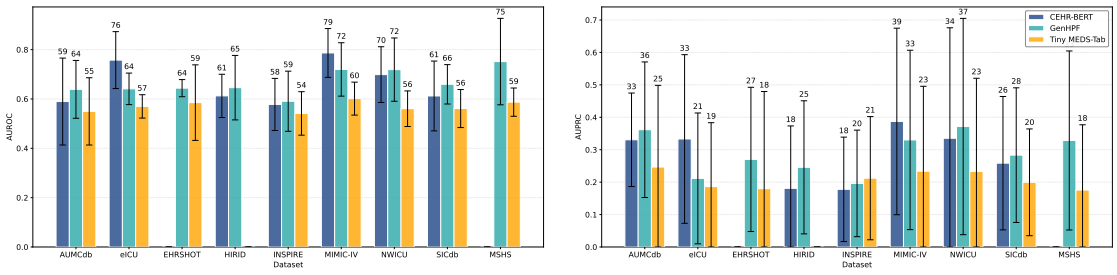


Figure E.4: Aggregated by dataset performance metrics for the samples weighted benchmarking *MEDS-DEV* compatible models. The error bars represent the standard deviation calculated over the individual samples. Mean values are shown on top of the error bars and rounded to two digits and multiplied by 100 for readability. Left: AUROC; Right AUPRC.

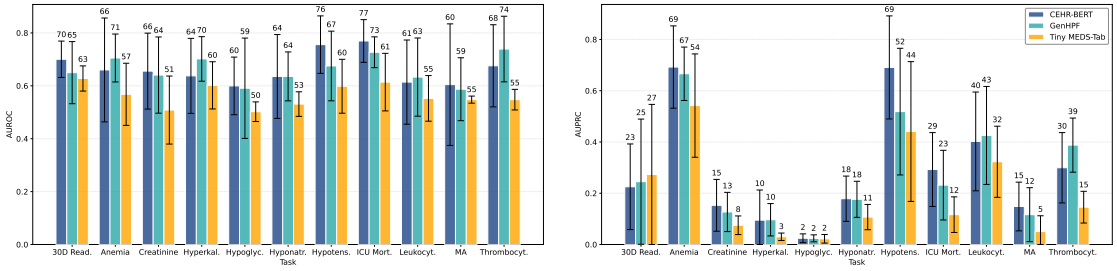


Figure E.5: Aggregated by task performance metrics for the subjects weighted benchmarking *MEDS-DEV* compatible models. The error bars represent the standard deviation calculated over the individual subjects. Mean values are shown on top of the error bars and rounded to two digits and multiplied by 100 for readability. Left: AUROC; Right AUPRC.

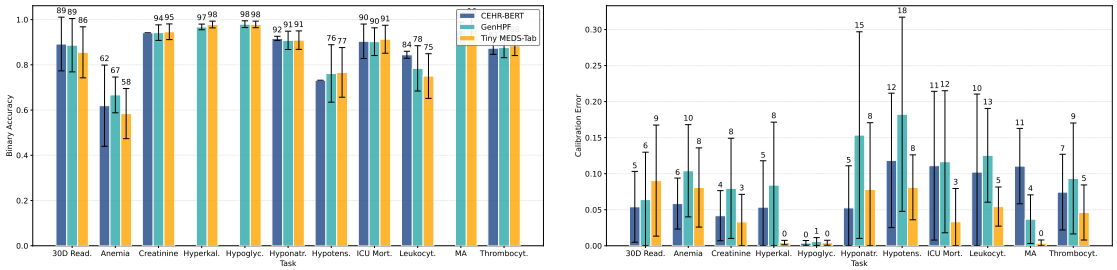


Figure E.6: Binary accuracy and calibration error aggregated by task for the subjects weighted benchmarking *MEDS-DEV* compatible models. The error bars represent the standard deviation calculated over the individual subjects. Mean values are shown on top of the error bars and rounded to two digits and multiplied by 100 for readability. Left: binary accuracy; Right: calibration error.

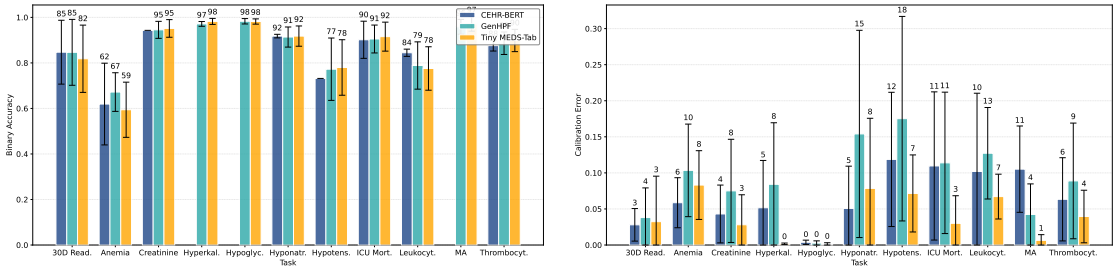
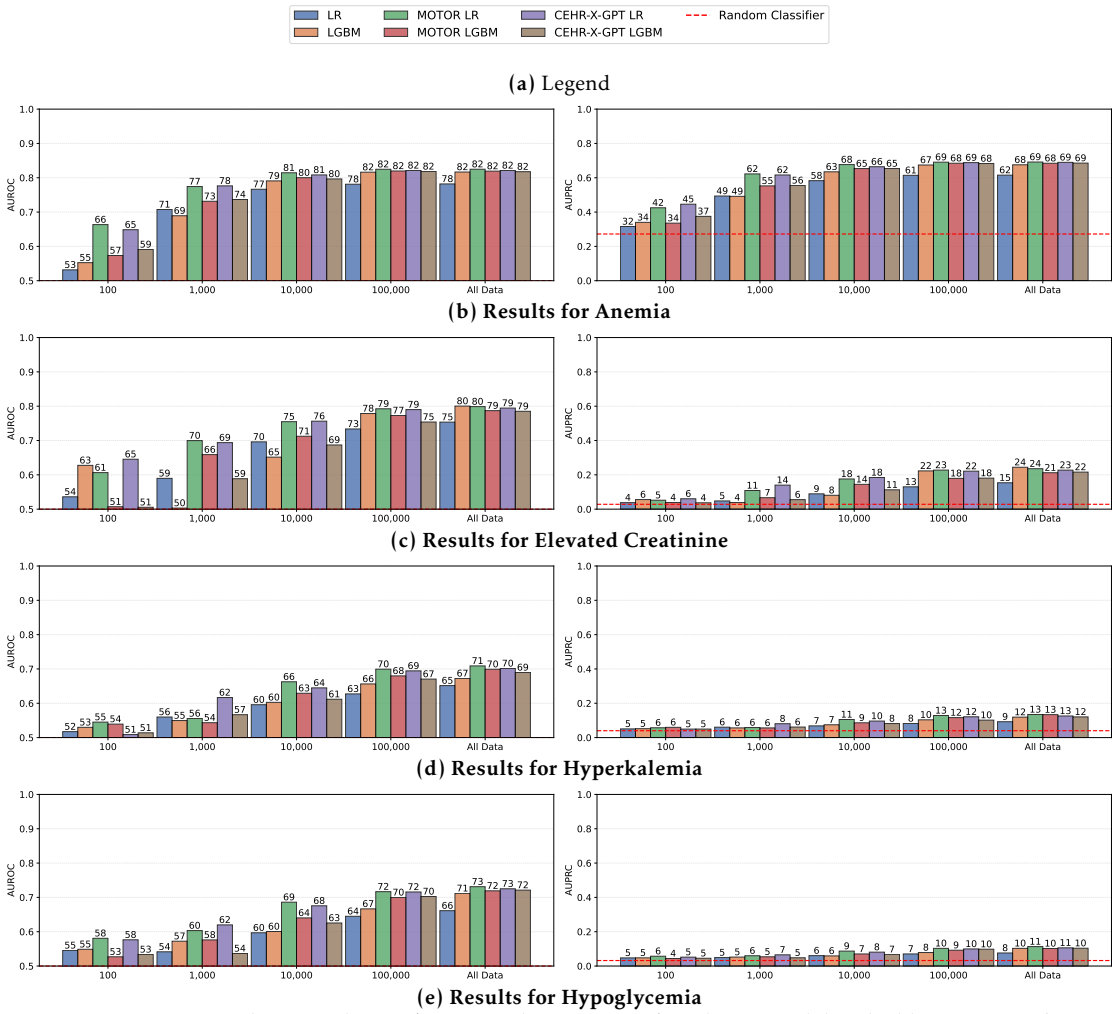


Figure E.7: Binary accuracy and calibration error aggregated by task for the samples weighted benchmarking *MEDS-DEV* compatible models. The error bars represent the standard deviation calculated over the individual subjects. Mean values are shown on top of the error bars and rounded to two digits and multiplied by 100 for readability. Left: binary accuracy; Right: calibration error.

E.4.2 Foundation Model Linear Probing Results



(a) Legend

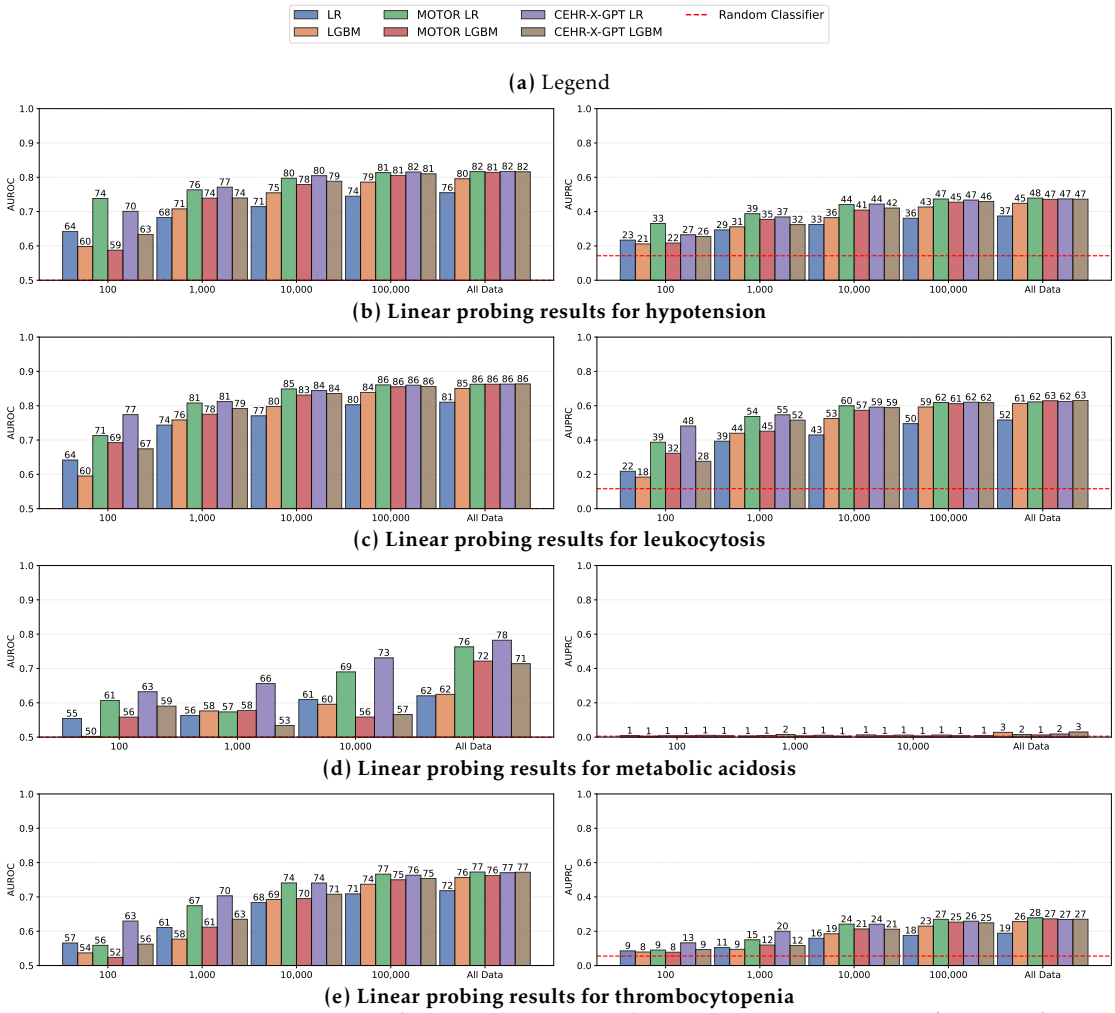
(b) Results for Anemia

(c) Results for Elevated Creatinine

(d) Results for Hyperkalemia

(e) Results for Hypoglycemia

Figure E.8: Linear probing results on the MSHS dataset using foundation model embeddings. We evaluate the performance of linear classifiers trained on frozen foundation model embeddings for various clinical tasks defined in the MSHS dataset. The results are presented in terms of AUROC and AUPRC metrics for each task.



(a) Legend

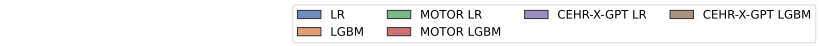
(b) Linear probing results for hypotension

(c) Linear probing results for leukocytosis

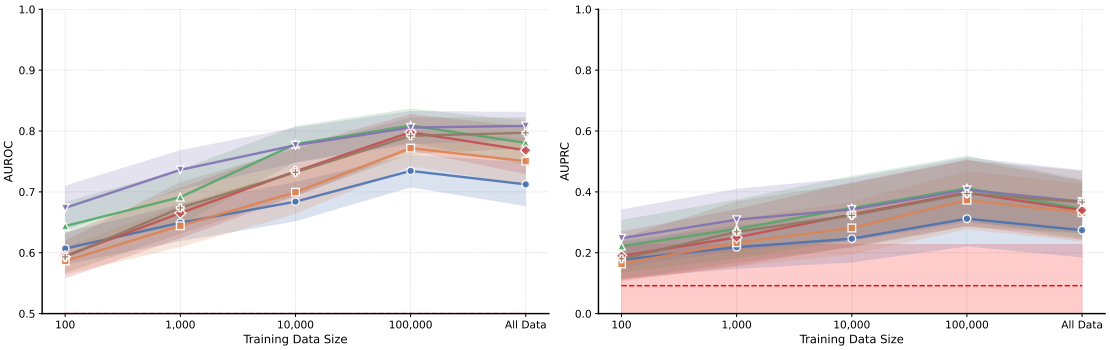
(d) Linear probing results for metabolic acidosis

(e) Linear probing results for thrombocytopenia

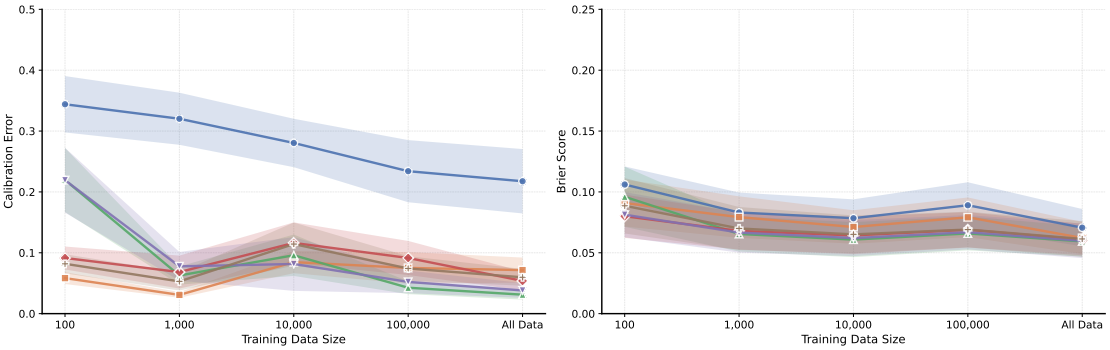
Figure E.9: Linear probing results on the MSHS dataset using foundation model embeddings (continued). We evaluate the performance of linear classifiers trained on frozen foundation model embeddings for various clinical tasks defined in the MSHS dataset. The results are presented in terms of AUROC and AUPRC metrics for each task.



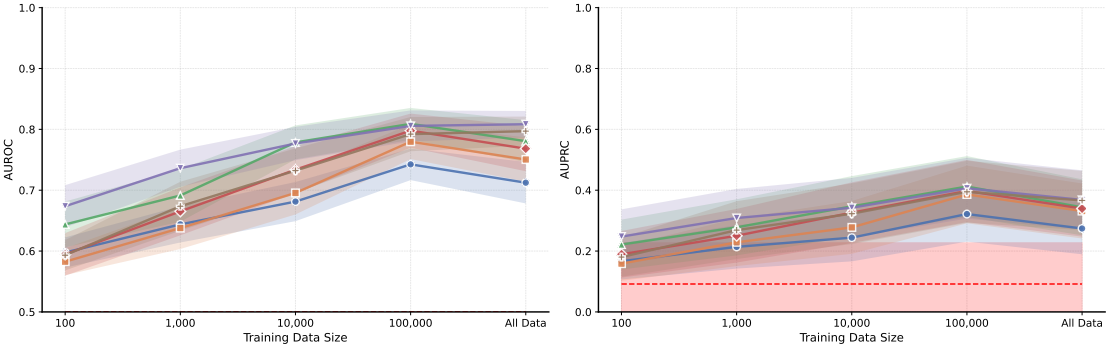
(a) Legend for aggregated linear probing results



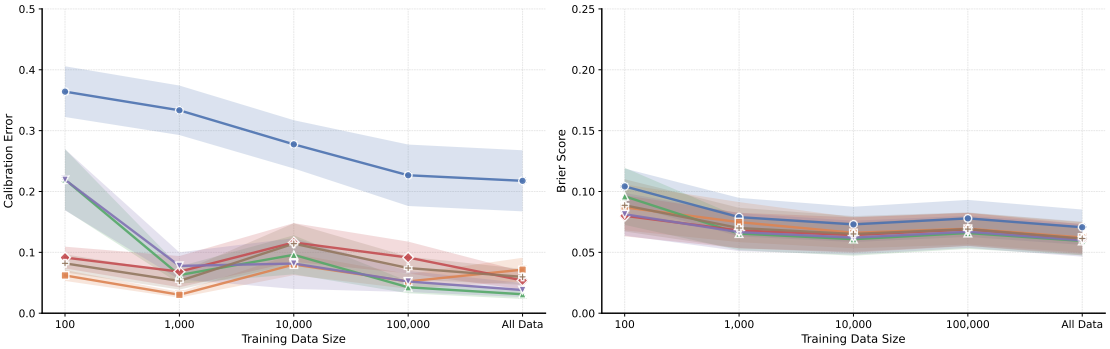
(b) Linear probing aggregated results for AUROC and AUPRC with tasks equally weighted (macro average).



(c) Linear probing aggregated results for calibration error and Brier score with tasks equally weighted (macro average).

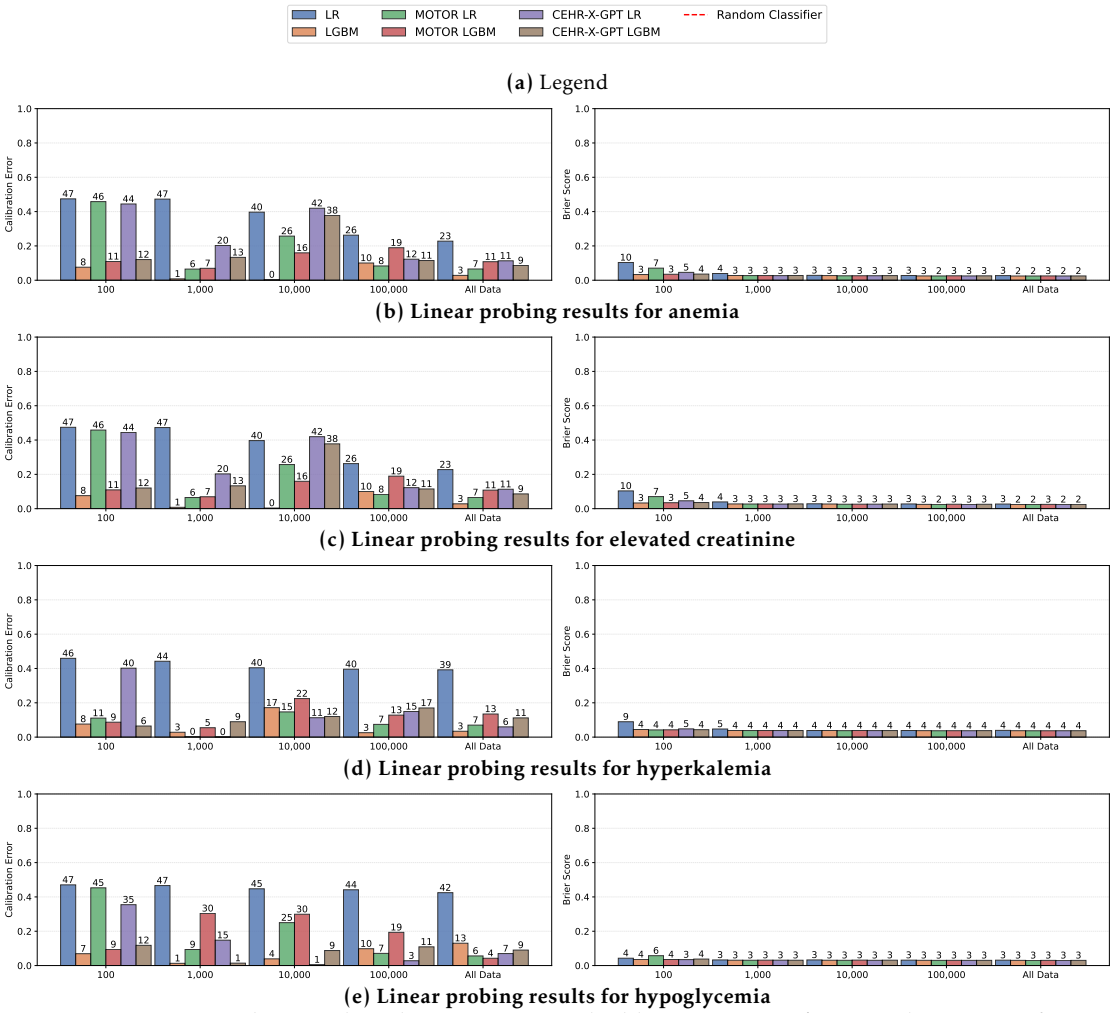


(d) Linear probing aggregated results for AUROC and AUPRC with tasks weighted according to cohort size (weighted average).



(e) Linear probing aggregated results for calibration error and Brier score with tasks weighted according to cohort size (weighted average).

Figure E.10: Aggregated results for linear probing with foundation models. We present the aggregated performance metrics for linear classifiers trained on frozen foundation model embeddings across various clinical tasks defined in the MSHS dataset. The results are shown for both macro and weighted averages of AUROC, AUPRC, calibration error, and Brier score metrics.



(a) Legend

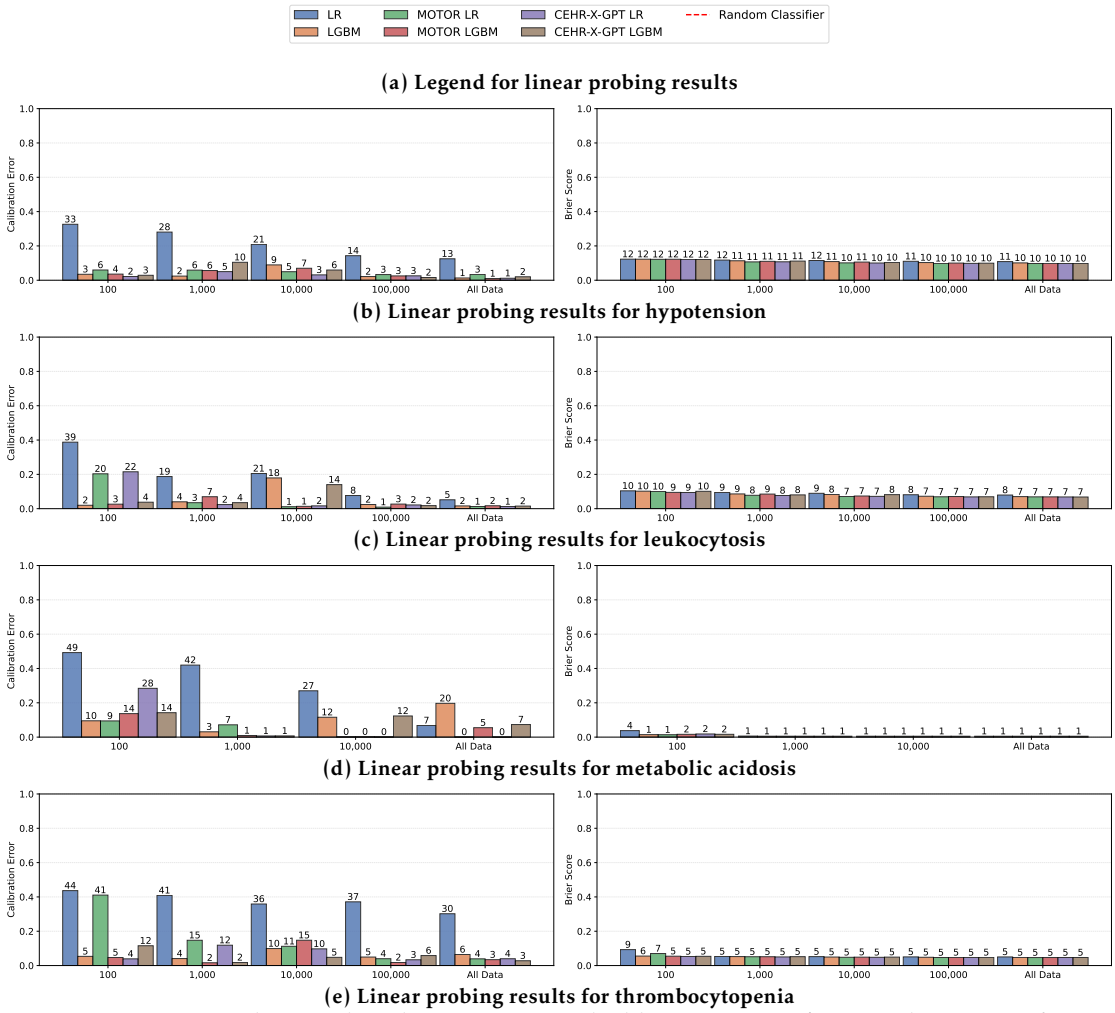
(b) Linear probing results for anemia

(c) Linear probing results for elevated creatinine

(d) Linear probing results for hyperkalemia

(e) Linear probing results for hypoglycemia

Figure E.11: Linear probing results in binary accuracy and calibration error on the MSHS dataset using foundation model embeddings. We evaluate the performance of linear classifiers trained on frozen foundation model embeddings for various clinical tasks defined in the MSHS dataset. Left column: binary accuracy; Right column calibration error.



(a) Legend for linear probing results

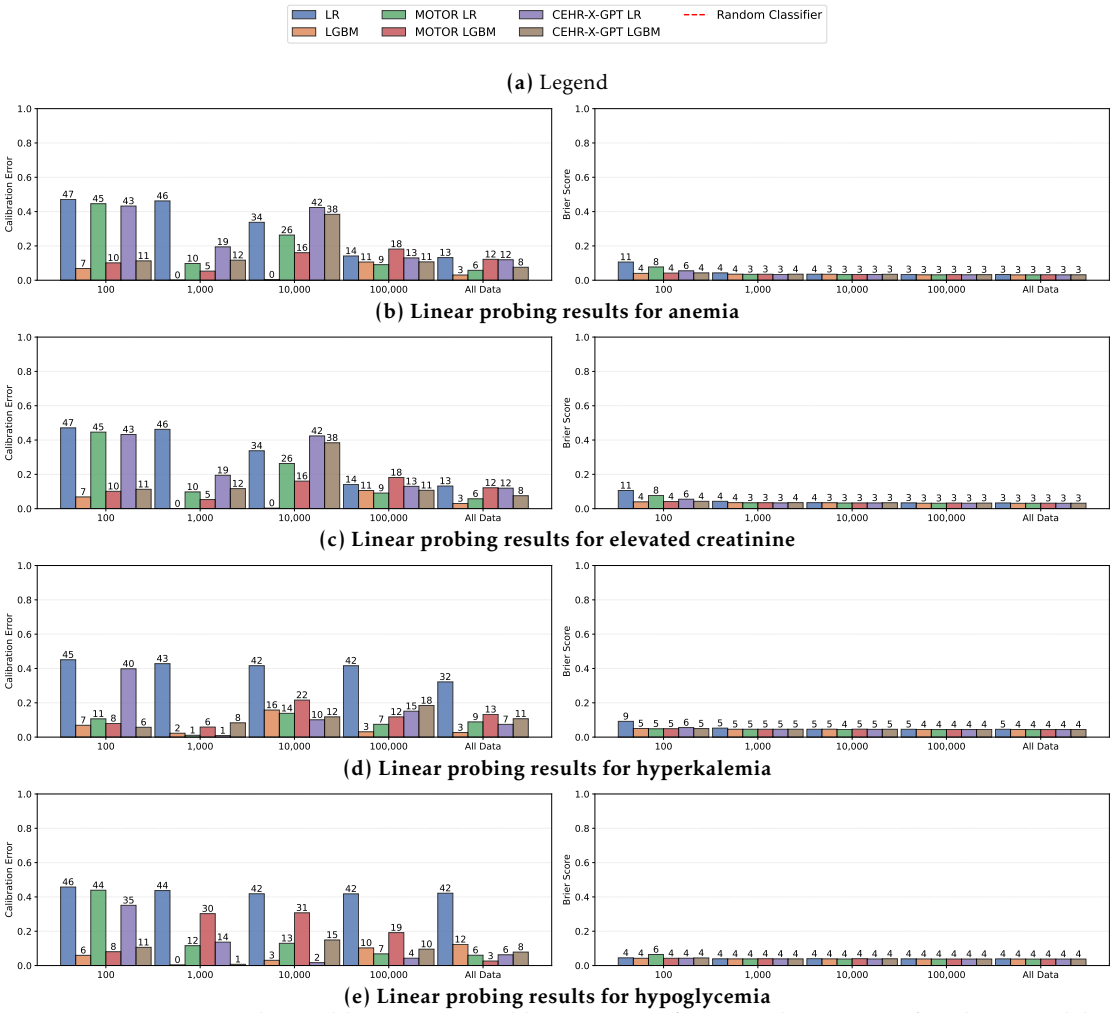
(b) Linear probing results for hypotension

(c) Linear probing results for leukocytosis

(d) Linear probing results for metabolic acidosis

(e) Linear probing results for thrombocytopenia

Figure E.12: Linear probing results in binary accuracy and calibration error on the MSHS dataset using foundation model embeddings (continued). We evaluate the performance of linear classifiers trained on frozen foundation model embeddings for various clinical tasks defined in the MSHS dataset. Left column: binary accuracy; Right column calibration error.



(a) Legend

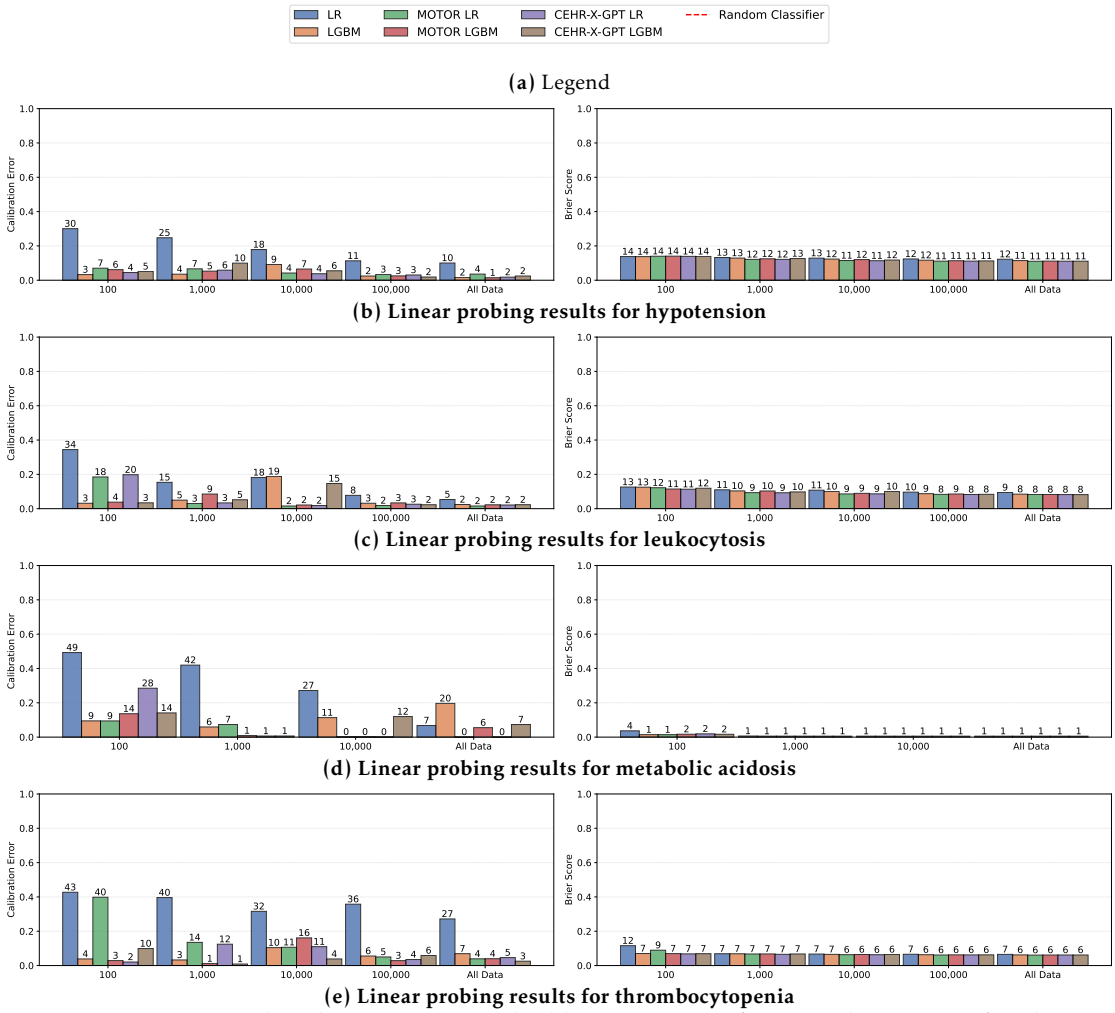
(b) Linear probing results for anemia

(c) Linear probing results for elevated creatinine

(d) Linear probing results for hyperkalemia

(e) Linear probing results for hypoglycemia

Figure E.13: Linear probing calibration error and brier score on the MSHS dataset using foundation model embeddings. We evaluate the performance of linear classifiers trained on frozen foundation model embeddings for various clinical tasks defined in the MSHS dataset. The results are presented in terms of AUROC and AUPRC metrics for each task.



(a) Legend

(b) Linear probing results for hypotension

(c) Linear probing results for leukocytosis

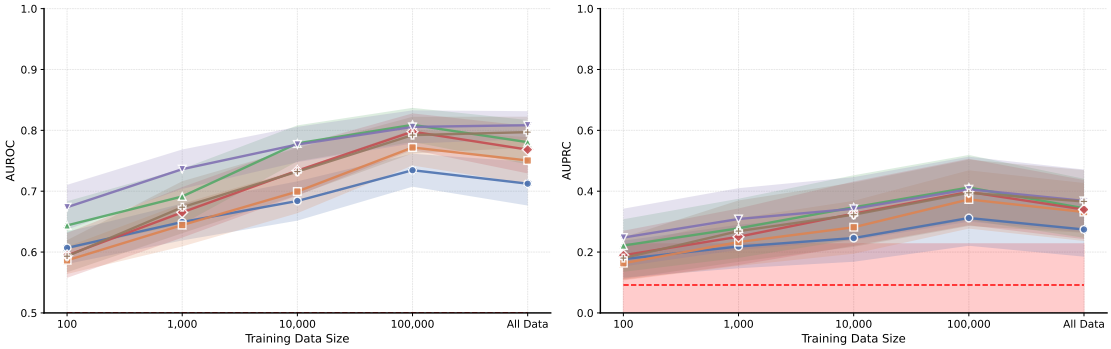
(d) Linear probing results for metabolic acidosis

(e) Linear probing results for thrombocytopenia

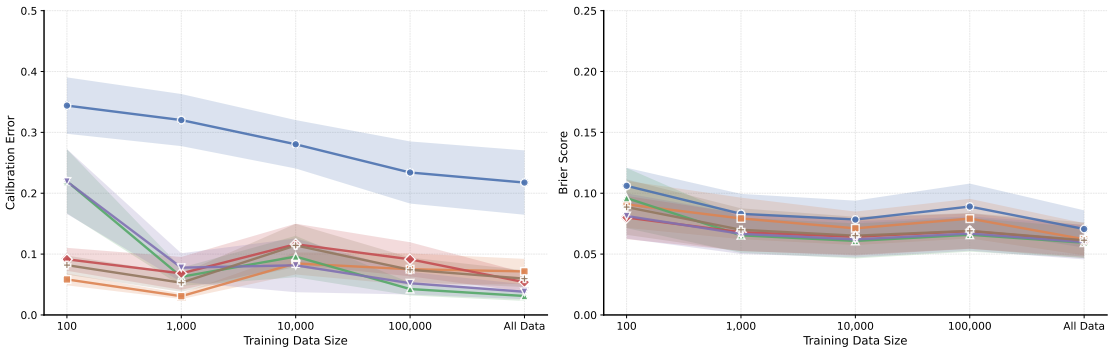
Figure E.14: Linear probing binary accuracy and calibration error on the MSHS dataset using foundation model embeddings (continued). We evaluate the performance of linear classifiers trained on frozen foundation model embeddings for various clinical tasks defined in the MSHS dataset. Left column: binary accuracy; Right column calibration error.



(a) Legend



(b) Linear probing aggregated results for AUROC and AUPRC with tasks equally weighted (micro average).



(c) Linear probing aggregated results for calibration error and Brier score with tasks equally weighted (micro average).

Figure E.15: Aggregated results for linear probing by task with foundation models.

E.5 Dimensionality Reduction of Foundation Model Embeddings

We perform dimensionality reduction on the foundation model embeddings using three techniques: t-distributed Stochastic Neighbor Embedding (t-SNE) [vdMH08], Uniform Manifold Approximation and Projection (UMAP) [MHM20], and Principal Component Analysis (PCA). We use the default settings for every library set in their respective package version: scikit-learn 1.5.2 and umap-learn 0.5.3. The results are visualized in the following figures.

E.5.1 Embedding Visualization Methodology

For each prediction task $t \in \mathcal{T}$ and model $m \in \{\texttt{motor}, \texttt{cehr}\}$, we computed 2D embeddings $\mathbf{Z}^{(t,m,d)} \in \mathbb{R}^{N_{t,m} \times 2}$ using dimensionality reduction methods $d \in \{\texttt{UMAP}, \texttt{t-SNE}, \texttt{PCA}\}$ and extracted binary labels $\mathbf{y}^{(t)} \in \{0,1\}^{N_{t,m}}$ and sequence lengths $\mathbf{s}^{(t)} \in \mathbb{R}_{+}^{N_{t,m}}$ from parquet files [`{task}_{model}_reduced.parquet`]. Each embedding was visualized twice using the most colorblind-friendly `cividis` colormap [K25] to ensure consistent visual comparison across plots.

Contour and density labeled plots with marginal distribution: Colors encoded $y_i \in \{0,1\}$ using colormap quantiles at $\{0.25, 0.75\}$. All density overlays used kernel density estimation (KDE) with Gaussian kernel ($h = 0.1$) over a 80×80 evaluation grid spanning data limits ± 3. A 2D pixel heatmap visualized total point density (128×128 bins, masked below 2% max density). Heatmaps used 2D histograms normalized to $[0,1]$ with zero-masking for sparse regions. Marginal KDEs were computed separately for each dimension and label.

Sequence-length scatter plots: Points $(z_{i,1}, z_{i,2})$ were plotted with settings $s = 8$ (scatterpoint size), $\alpha = 0.5$. We randomly subsampled points to a maximum of 10,000 to ensure a reasonable file size and to prevent issues from rendering the vectorized graphs with PDF readers. Colors were encoded as heatmap with a log-scale with per-figure normalization.

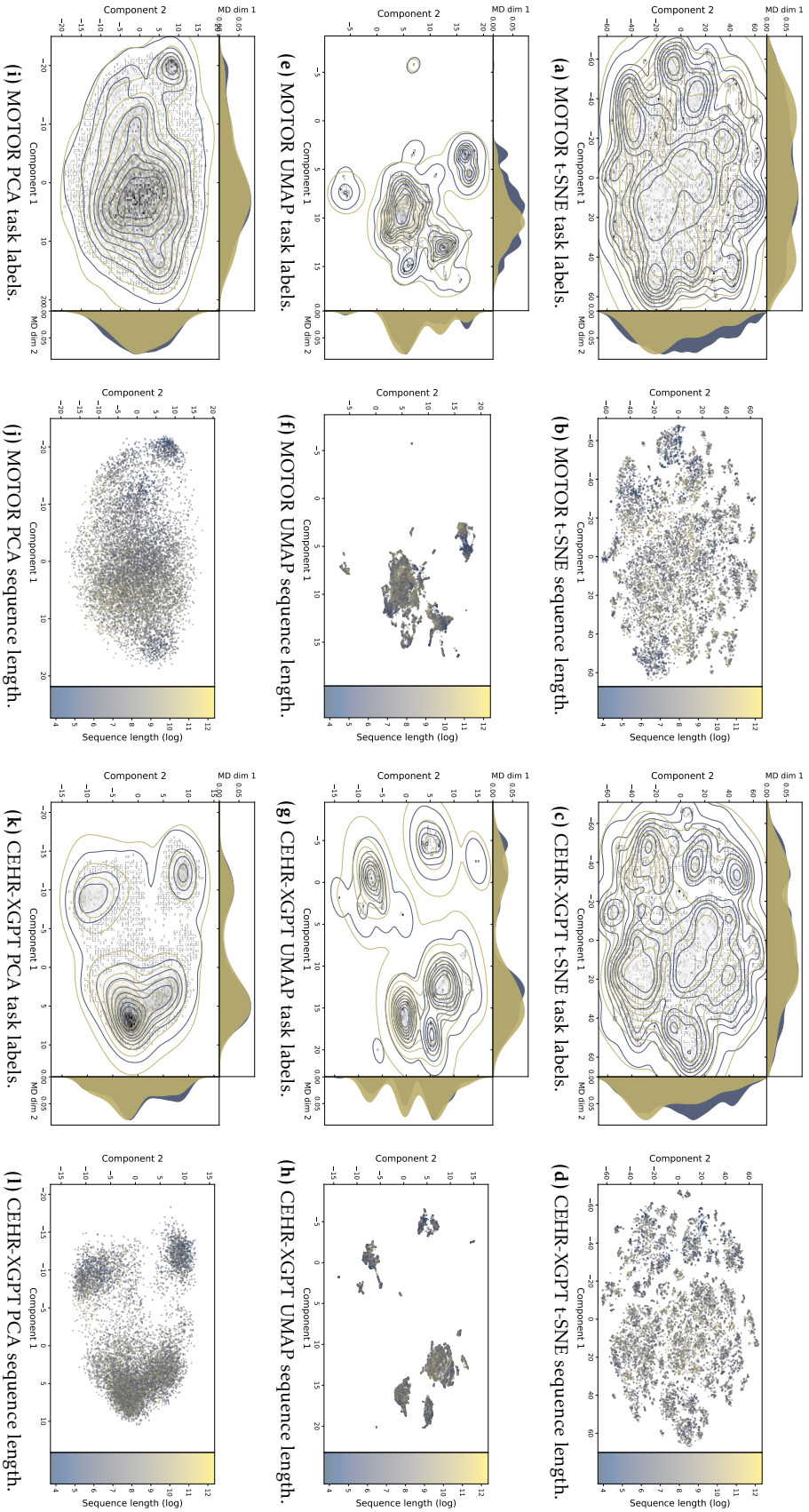
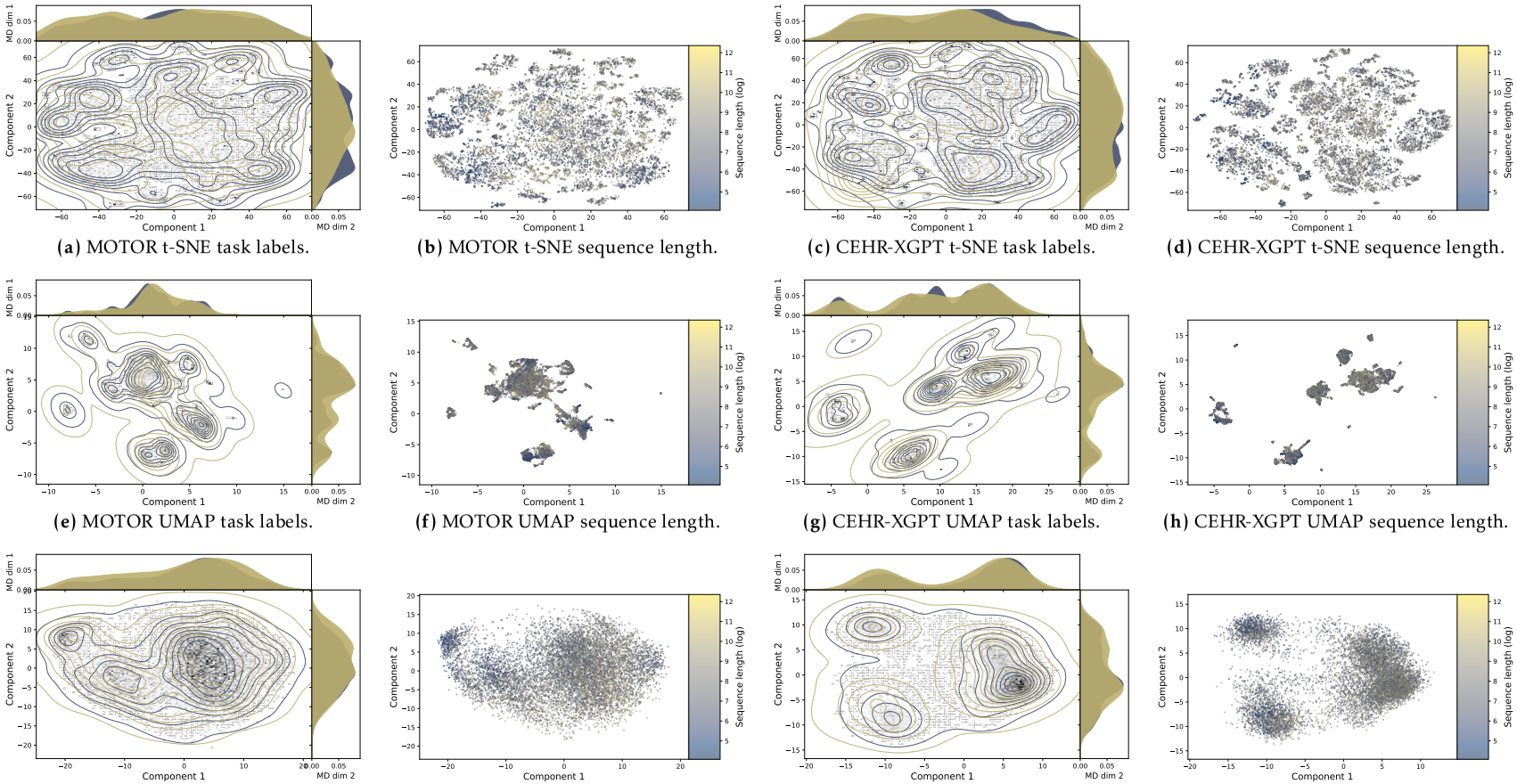


Figure E.16: Dimensionality reduction for hypoglycemia. Top row to bottom: t-SNE, UMAP, PCA. Left column: MOTOR embeddings, Right column: CEHR-XGPT embeddings. First and third columns: colored by task labels (positive/negative), second and fourth columns: colored by sequence length.

(a) MOTOR t-SNE task labels.
(b) MOTOR t-SNE sequence length.
(c) CEHR-XGPT t-SNE task labels.
(d) CEHR-XGPT t-SNE sequence length.
(e) MOTOR UMAP task labels.
(f) MOTOR UMAP sequence length.
(g) CEHR-XGPT UMAP task labels.
(h) CEHR-XGPT UMAP sequence length.
(i) MOTOR PCA task labels.
(j) MOTOR PCA sequence length.
(k) CEHR-XGPT PCA task labels.
(l) CEHR-XGPT PCA sequence length.

Figure E.17: Dimensionality reduction for elevated creatinine. Top row to bottom: t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), Principal Component Analysis (PCA). Left column: MOTOR embeddings, Right column: CEHR-XGPT embeddings. First and third columns: colored by task labels (positive/negative), second and fourth columns: colored by sequence length.

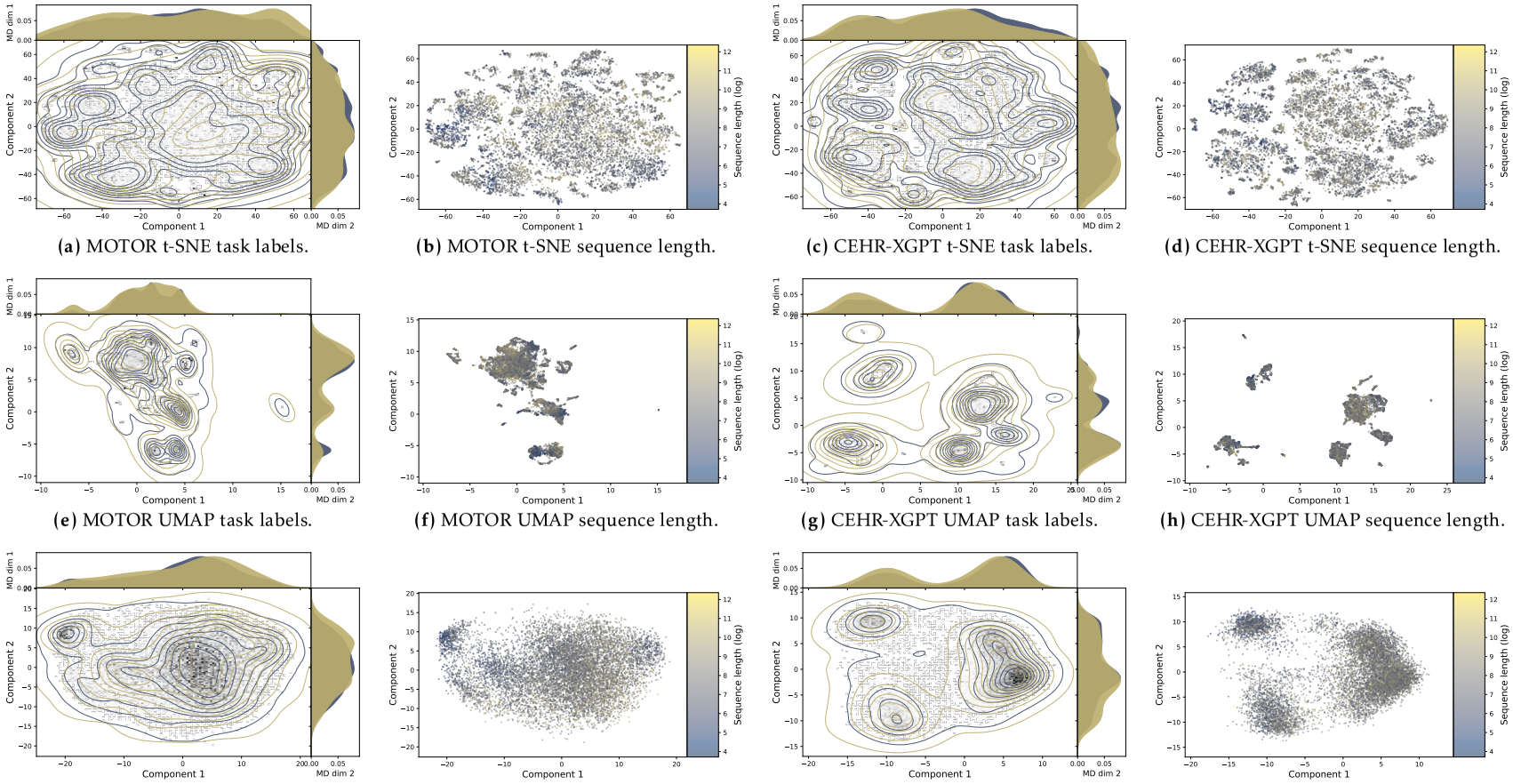


(a) MOTOR t-SNE task labels. **(b)** MOTOR t-SNE sequence length. **(c)** CEHR-XGPT t-SNE task labels. **(d)** CEHR-XGPT t-SNE sequence length.

(e) MOTOR UMAP task labels. **(f)** MOTOR UMAP sequence length. **(g)** CEHR-XGPT UMAP task labels. **(h)** CEHR-XGPT UMAP sequence length.

(i) MOTOR PCA task labels. **(j)** MOTOR PCA sequence length. **(k)** CEHR-XGPT PCA task labels. **(l)** CEHR-XGPT PCA sequence length.

Figure E.18: Dimensionality reduction for hyperkalemia. Top row to bottom: t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), Principal Component Analysis (PCA). Left column: MOTOR embeddings, Right column: CEHR-XGPT embeddings. First and third columns: colored by task labels (positive/negative), second and fourth columns: colored by sequence length.



(a) MOTOR t-SNE task labels.
(b) MOTOR t-SNE sequence length.
(c) CEHR-XGPT t-SNE task labels.
(d) CEHR-XGPT t-SNE sequence length.
(e) MOTOR UMAP task labels.
(f) MOTOR UMAP sequence length.
(g) CEHR-XGPT UMAP task labels.
(h) CEHR-XGPT UMAP sequence length.
(i) MOTOR PCA task labels.
(j) MOTOR PCA sequence length.
(k) CEHR-XGPT PCA task labels.
(l) CEHR-XGPT PCA sequence length.

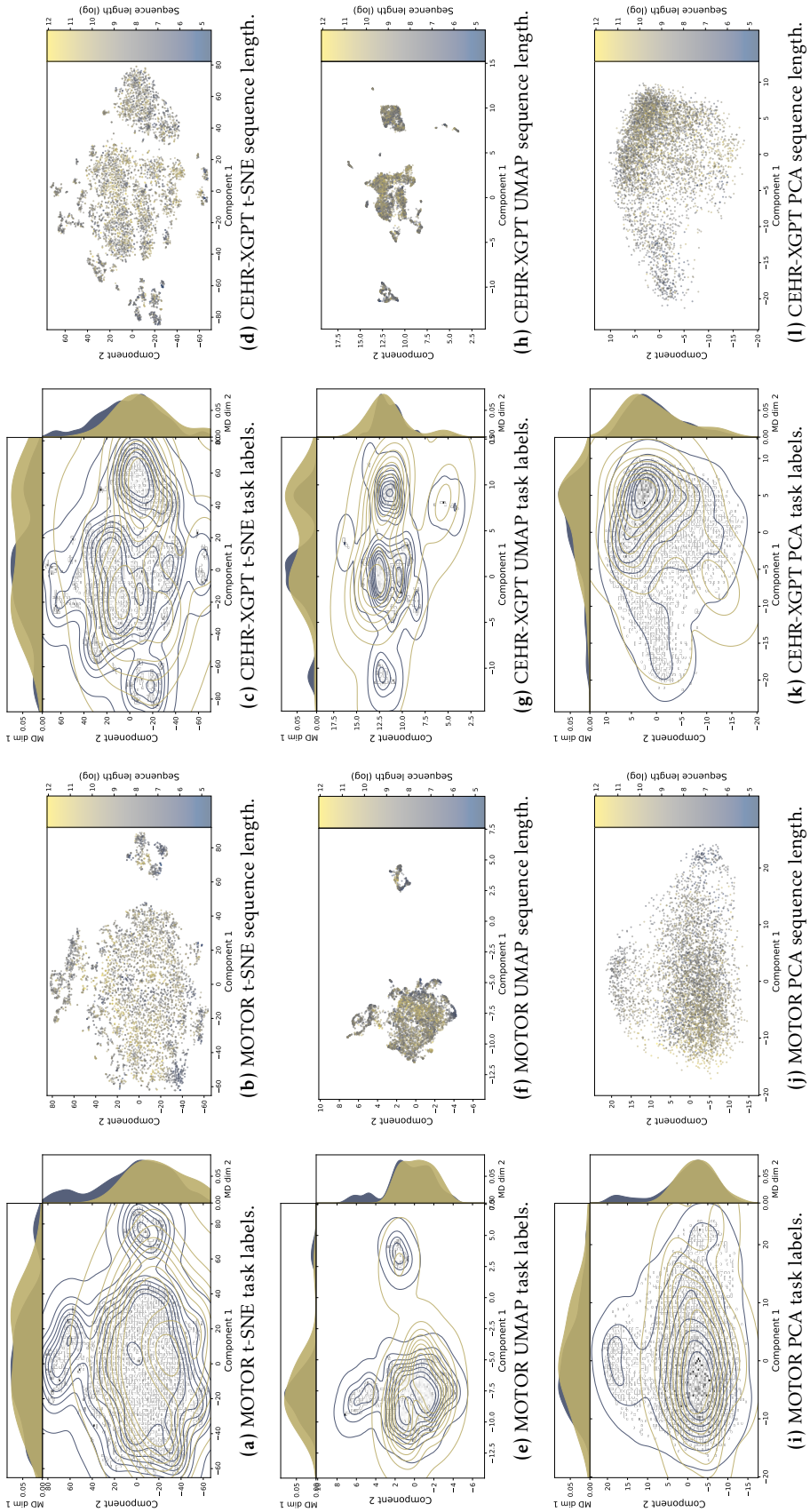


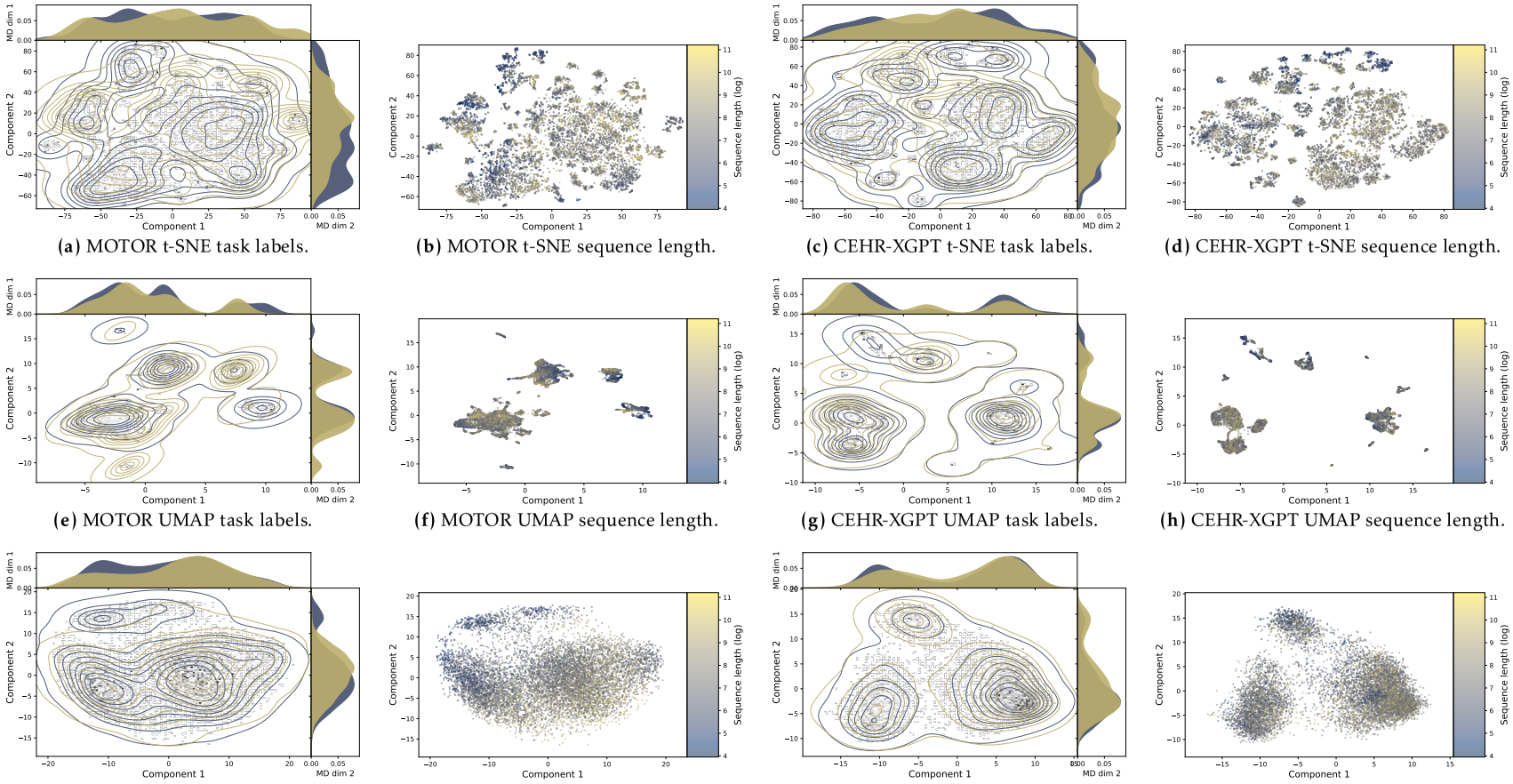
Figure E.19: Dimensionality reduction for metabolic acidosis. Top row to bottom: t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), Principal Component Analysis (PCA). Left column: MOTOR embeddings, Right column: CEHR-XGPT embeddings. First and third columns: colored by task labels (positive/negative), second and fourth columns: colored by sequence length.

(a) MOTOR t-SNE task labels. **(b)** MOTOR t-SNE sequence length. **(c)** CEHR-XGPT t-SNE task labels. **(d)** CEHR-XGPT t-SNE sequence length.

(e) MOTOR UMAP task labels. **(f)** MOTOR UMAP sequence length. **(g)** CEHR-XGPT UMAP task labels. **(h)** CEHR-XGPT UMAP sequence length.

(i) MOTOR PCA task labels. **(j)** MOTOR PCA sequence length. **(k)** CEHR-XGPT PCA task labels. **(l)** CEHR-XGPT PCA sequence length.

Figure E.20: Dimensionality reduction for anemia. Top row to bottom: t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), Principal Component Analysis (PCA). Left column: MOTOR embeddings, Right column: CEHR-XGPT embeddings. First and third columns: colored by task labels (positive/negative), second and fourth columns: colored by sequence length.



(a) MOTOR t-SNE task labels.
(b) MOTOR t-SNE sequence length.
(c) CEHR-XGPT t-SNE task labels.
(d) CEHR-XGPT t-SNE sequence length.
(e) MOTOR UMAP task labels.
(f) MOTOR UMAP sequence length.
(g) CEHR-XGPT UMAP task labels.
(h) CEHR-XGPT UMAP sequence length.
(i) MOTOR PCA task labels.
(j) MOTOR PCA sequence length.
(k) CEHR-XGPT PCA task labels.
(l) CEHR-XGPT PCA sequence length.

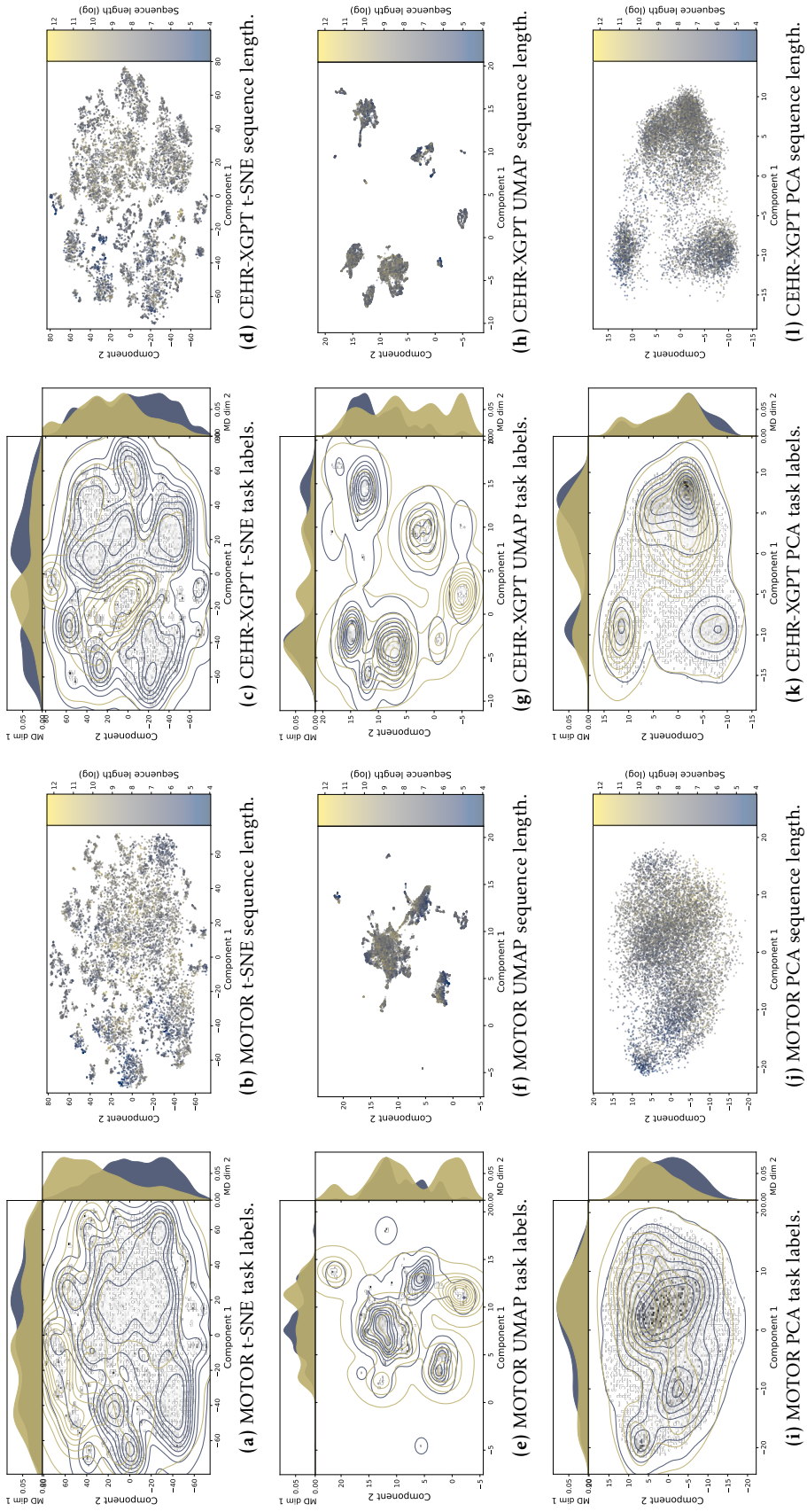


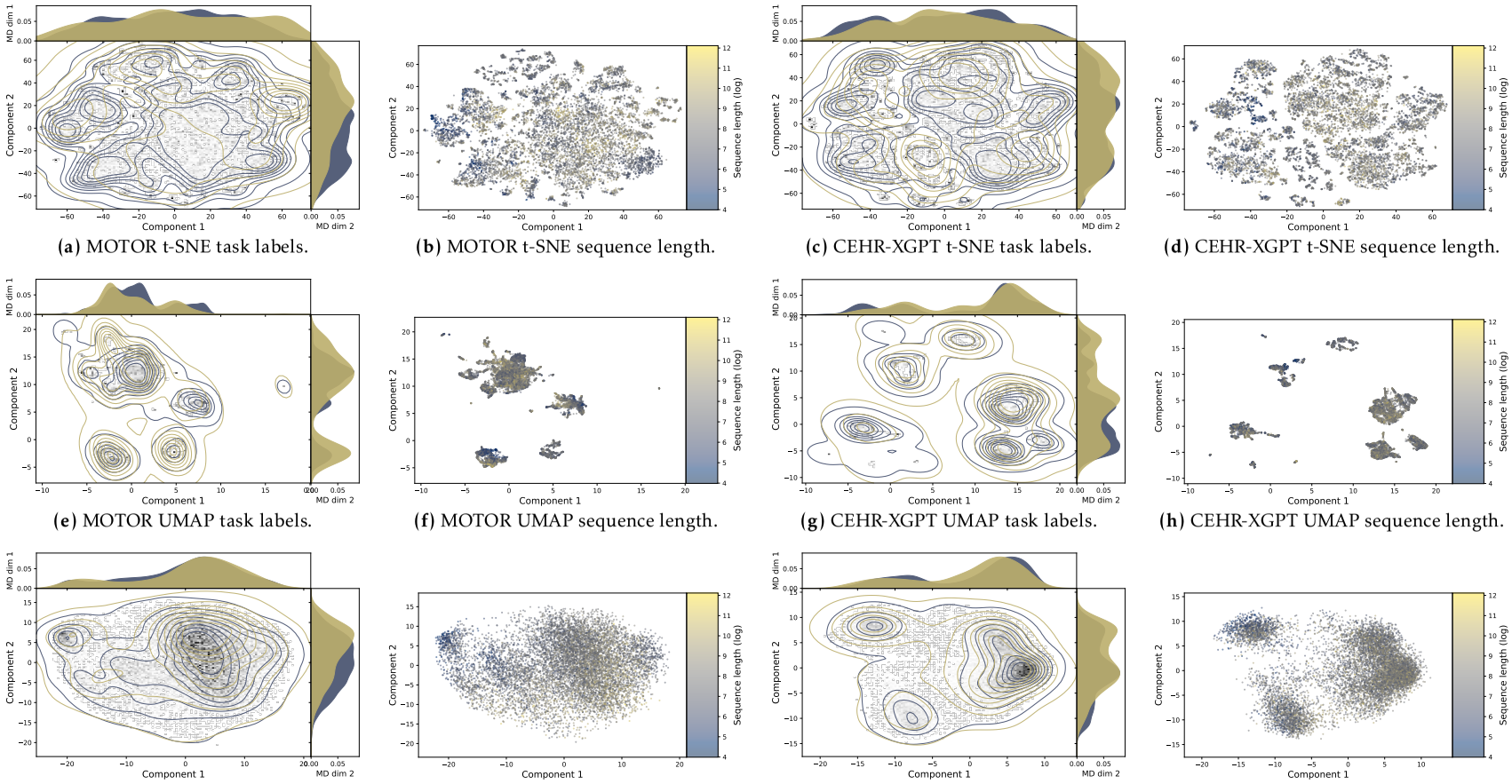
Figure E.21: Dimensionality reduction for leukocytosis. Top row to bottom: t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), Principal Component Analysis (PCA). Left column: MOTOR embeddings, Right column: CEHR-XGPT embeddings. First and third columns: colored by task labels (positive/negative), second and fourth columns: colored by sequence length.

(a) MOTOR t-SNE task labels. **(b)** MOTOR t-SNE sequence length. **(c)** CEHR-XGPT t-SNE task labels. **(d)** CEHR-XGPT t-SNE sequence length.

(e) MOTOR UMAP task labels. **(f)** MOTOR UMAP sequence length. **(g)** CEHR-XGPT UMAP task labels. **(h)** CEHR-XGPT UMAP sequence length.

(i) MOTOR PCA task labels. **(j)** MOTOR PCA sequence length. **(k)** CEHR-XGPT PCA task labels. **(l)** CEHR-XGPT PCA sequence length.

Figure E.22: Dimensionality reduction for thrombocytopenia. Top row to bottom: t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), Principal Component Analysis (PCA). Left column: MOTOR embeddings, Right column: CEHR-XGPT embeddings. First and third columns: colored by task labels (positive/negative), second and fourth columns: colored by sequence length.



(a) MOTOR t-SNE task labels.
(b) MOTOR t-SNE sequence length.
(c) CEHR-XGPT t-SNE task labels.
(d) CEHR-XGPT t-SNE sequence length.
(e) MOTOR UMAP task labels.
(f) MOTOR UMAP sequence length.
(g) CEHR-XGPT UMAP task labels.
(h) CEHR-XGPT UMAP sequence length.
(i) MOTOR PCA task labels.
(j) MOTOR PCA sequence length.
(k) CEHR-XGPT PCA task labels.
(l) CEHR-XGPT PCA sequence length.

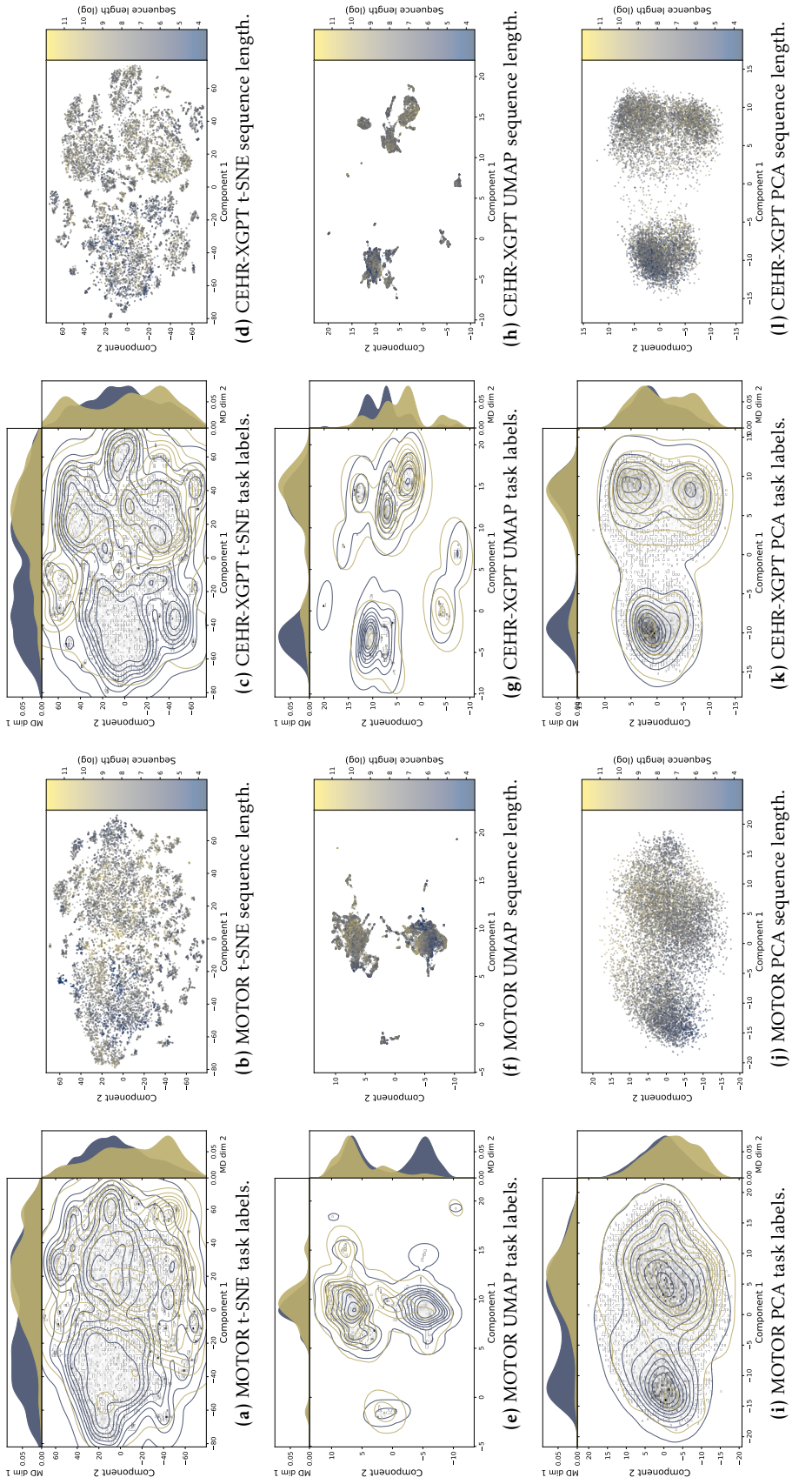


Figure E.23: Dimensionality reduction for hypotension. Top row to bottom: t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), Principal Component Analysis (PCA). Left column: MOTOR embeddings, Right column: CEHR-XGPT embeddings. First and third columns: colored by task labels (positive/negative), second and fourth columns: colored by sequence length.

(a) MOTOR t-SNE task labels. **(b)** MOTOR t-SNE sequence length. **(c)** CEHR-XGPT t-SNE task labels. **(d)** CEHR-XGPT t-SNE sequence length.

(e) MOTOR UMAP task labels. **(f)** MOTOR UMAP sequence length. **(g)** CEHR-XGPT UMAP task labels. **(h)** CEHR-XGPT UMAP sequence length.

(i) MOTOR PCA task labels. **(j)** MOTOR PCA sequence length. **(k)** CEHR-XGPT PCA task labels. **(l)** CEHR-XGPT PCA sequence length.

Appendix F

German Layperson Summary

Gesundheitssysteme sammeln täglich riesige Datenmengen. Doch die Umwandlung der erhaltenen Informationen in zuverlässige maschinelle Lernmodelle (ML), die tatsächlich Patienten helfen, bleibt schwierig. Mit einer alternden Bevölkerung und steigenden Gesundheitskosten wächst der Bedarf an praktischen, skalierbaren Systemen, die vielfältige medizinische Daten sinnvoll nutzen. Obwohl erste Arbeiten Potenzial zeigen, gibt es grundlegende Probleme: Ergebnisse lassen sich oft nicht reproduzieren und die Systeme kommen in der Praxis nicht an. Diese Dissertation geht diese verbundenen Probleme an indem sie sowohl konkrete Anwendungen in der Medizin erforscht als auch die notwendige Infrastruktur schafft, um zuverlässigere und einsetzbare künstliche Intelligente (KI)-Systeme zu ermöglichen.

Ein zentrales Problem ist die mangelnde Reproduzierbarkeit der ML-forschung. Obwohl öffentliche Intensivstationsdatensätze verfügbar sind, fehlt ein Standard dafür, wie Daten vorbereitet, Prediktionsaufgaben definiert und Ergebnisse überprüft werden. Jeder Forscher verwendet andere Methoden, weshalb es unmöglich ist, ML-Modelle fair zu vergleichen oder auf frühere Arbeiten aufzubauen. Diese Dissertation präsentiert: *YAIB* (Yet Another ICU Benchmark), ein flexibles Framework für standardisierte medizinische ML-Experimente, und *ReciPies*, ein einfaches System zur Dokumentation und Wiederverwendung von Datenaufbereitungsschritten. Überraschenderweise zeigt sich durch diese Arbeit, dass Entscheidungen bei der Datenauswahl und -vorbereitung oft wichtiger sind als die Wahl des Modells selbst.

Der zweite Teil zeigt die praktische Kraft direkter klinischer Zusammenarbeit. Wir haben zusammen mit Chirurgen Daten von über 1.200 Patienten gesammelt, einschließlich kontinuierlicher Messwerte von tragbaren Sensoren. Damit entwickelten wir ein Frühwarnsystem, das an Pflegestationen subtile Komplikationen wie Wundinfektionen oder Pankreasfissteln zeitig erkennt. Die hochauflösenden Sensordaten verbesserten die Erkennungsgenauigkeit deutlich, was zeigt, wie wichtig präzise, kontinuierliche Messungen sind.

Die dritte Komponente adressiert Skalierbarkeit. Um KI-Systeme über Krankenhäuser und Datensätze hinweg zu verbreiten, braucht man standardisierte Werkzeuge. Diese Arbeit trägt zum *Medical Event Data Standard* (MEDS) bei — einem einfachen, offenen Datenformat für medizinische Aufzeichnungen, das die Zusammenarbeit und Reproduzierbarkeit erleichtert. Mit diesem Standard entstanden umfangreiche Tests

mit elf klinischen Aufgaben über mehrere Datensätze und 12 Millionen Patienten, die zeigen, wie viel sich durch standardisierte, institutionsübergreifende Forschung erreichen lässt.

Ein durchgehendes Muster ergibt sich aus all dieser Arbeit: Erfolgreiche KI in der Medizin braucht mehr als nur clevere Algorithmen. Sie benötigt sorgfältige Datenarbeit, klare Methoden und robuste Werkzeuge. Zwar ist jede Anwendung unterschiedlich und braucht maßgeschneiderte Lösungen, doch die Bausteine dieser Dissertation bilden ein Fundament, auf dem Forscher und Kliniker weltweit aufbauen können. Das Ziel ist klar: maschinelles Lernen aus dem Forschungslabor in Krankenhäusern zu bringen, damit es vertrauenswürdig und gerecht das Leben verbessert.

Declaration

I hereby confirm that

- this dissertation is the result of my own work, it was prepared without unauthorized help and using only the given literature,
- this dissertation has not been previously submitted, in part or whole, to any other university,
- I am aware of the doctorate regulations of the Digital Engineering Faculty of the University of Potsdam from November 27, 2019.

Ich erkläre hiermit, dass

- ich die vorliegende Dissertationsschrift selbständig und ohne unerlaubte Hilfe angefertigt sowie nur die angegebene Literatur verwendet habe,
- die Dissertation keiner anderen Hochschule in gleicher oder ähnlicher Form vorgelegt wurde,
- mir die Promotionsordnung der Digital Engineering Fakultät der Universität Potsdam vom 27. November 2019 bekannt ist.

Robin Philippus van de Water – January 21, 2026