

Fall 2021

Secondary use of Structured Electronic Health Records Data: From Observational Studies to Deep Learning-based Predictive Modeling

Laila Bekhet

University of Texas Health Science Center at Houston, Laila.Rasmy.GindyBekhetBekhet@uth.tmc.edu

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthshis_dissertations



Part of the [Biomedical Informatics Commons](#)

Recommended Citation

Bekhet, Laila, "Secondary use of Structured Electronic Health Records Data: From Observational Studies to Deep Learning-based Predictive Modeling" (2021). *Dissertations (Open Access)*. 53.

https://digitalcommons.library.tmc.edu/uthshis_dissertations/53

This is brought to you for free and open access by the McWilliams School of Biomedical Informatics at DigitalCommons@TMC. It has been accepted for inclusion in Dissertations (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digcommons@library.tmc.edu.

Secondary use of Structured Electronic Health Records Data: From
Observational Studies to Deep Learning-based Predictive
Modeling

By

Laila Rasmy Gindy Bekhet, M.S.

APPROVED:

Degui Zhi, PhD, Chair

Hua Xu, PhD

Angela Ross, PhD

Date approved: _____

Secondary use of Structured Electronic Health Records Data: From Observational
Studies to Deep Learning-based Predictive Modeling

A
Dissertation

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
School of Biomedical Informatics
in Partial Fulfilment of the Requirements for the Degree of
Doctor of Philosophy

By

Laila Rasmy Gindy Bekhet, M.S.

University of Texas Health Science Center at Houston

2021

Dissertation Committee:

Degui Zhi, PhD¹, Advisor
Hua Xu, PhD¹
Angela Ross, PhD¹

¹The School of Biomedical Informatics

Copyright by
Laila Rasmy Gindy Bekhet
2021

Dedication

Dedicated to my parents, whose loss ignited my current research interest, my brother Remon Rasmy, my great supporter, and the whole family.

Acknowledgements

I acknowledge the exceptional support and guidance of my advisor Dr. Degui Zhi and my committee members Dr. Hua Xu and Dr. Angela M. Ross. I also acknowledge the UTHealth Innovation for Cancer Prevention Research Training Program Pre-Doctoral Fellowship (CPRIT Grant RP160015) and the amazing support of Dr. Patricia Mullen, Dr. Sahiti Myneni, and all of my peer fellows. I also would like to acknowledge the EHR Working Group, led by Dr. Hulin Wu, for the group's amazing collaboration opportunities and Dr. Kathleen McGrow from Microsoft for the insights in regard to artificial intelligence-based models' implementability. I am grateful to the UTHealth School of Biomedical Informatics faculty, especially Dr. Cui Tao, Dr. Dean Sittig, and Dr. Robert Murphy, for their collaboration and valuable insights, and Dr. Wenjin Zheng, and Dr. Xiaoqian Jiang, for their support and facilitating my access to powerful computational resources. I also am grateful to our clinical collaborators, George Williams, MD; David Aguilar, MD; Masayuki Nigo, MD; and Bijun S. Kannadath, MBBS, MS, for the helpful discussions on definitions of cohorts and the evaluation of results. Last but not least, I acknowledge all of my co-authors, Dr. Vahed Maurofy, Dr. Yang Xiang, and Dr. Ziqian Xie, as well as all of Dr. Zhi's team members. All of my research was undertaken with the assistance of resources and services from the School of Biomedical Informatics Data Service team, Firat Tiriyaki, Yujia Zhou, and Judy Young.

Abstract

With the wide adoption of electronic health records (EHRs), researchers, as well as large healthcare organizations, governmental institutions, insurance, and pharmaceutical companies have been interested in leveraging this rich clinical data source to extract clinical evidence and develop predictive algorithms. Large vendors have been able to compile structured EHR data from sites all over the United States, de-identify these data, and make them available to data science researchers in a more usable format. For this dissertation, we leveraged one of the earliest and largest secondary EHR data sources and conducted three studies of increasing scope. In the first study, which was of limited scope, we conducted a retrospective observational study to compare the effect of three drugs on a specific population of approximately 3,000 patients. Using a novel statistical method, we found evidence that the selection of phenylephrine as the primary vasopressor to induce hypertension for the management of nontraumatic subarachnoid hemorrhage is associated with better outcomes as compared to selecting norepinephrine or dopamine. In the second study, we widened our scope, using a cohort of more than 100,000 patients to train generalizable models for the risk prediction of specific clinical events, such as heart failure in diabetes patients or pancreatic cancer. In this study, we found that recurrent neural network-based predictive models trained on expressive terminologies, which preserve a high level of granularity, are associated with better prediction performance as compared with other baseline methods, such as logistic

regression. Finally, we widened our scope again, to train Med-BERT, a foundation model, on more than 20 million patients' diagnosis data. Med-BERT was found to improve the prediction performance of downstream tasks that have a small sample size, which otherwise would limit the ability of the model to learn good representation.

In conclusion, we found that we can extract useful information and train helpful deep learning-based predictive models. Given the limitations of secondary EHR data and taking into consideration that the data were originally collected for administrative and not research purposes, however, the findings need clinical validation. Therefore, clinical trials are warranted to further validate any new evidence extracted from such data sources before updating clinical practice guidelines. The implementability of the developed predictive models, which are in an early development phase, also warrants further evaluation.

Vita

2002.....B.Sc., Pharmaceutical Sciences, Ain Shams University

2011.....MBA, Globalization, Maastricht School of Management

2016.....Graduate Teaching Assistant, School of Biomedical Informatics, University of Texas Health Science Center in Houston

2017.....M.Sc., Biomedical Informatics, University of Texas Health Science Center in Houston

2018.....Programmer Analyst, School of Biomedical Informatics, University of Texas Health Science Center in Houston

2019.....Graduate Research Assistant, School of Biomedical Informatics, University of Texas Health Science Center in Houston

2020 to present.....Pre-doctoral Research Fellow, Innovation for Cancer Prevention Research Training Program, University of Texas Health Science Center in Houston

Publications

- Rasmy L**, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*. 2021 May 20;4(1):86. doi: 10.1038/s41746-021-00455-y. PMID: 34017034; PMCID: PMC8137882.
- Rasmy L**, Tiriyaki F, Zhou Y, Xiang Y, Tao C, Xu H, Zhi D. Representation of EHR data for predictive modeling: a comparison between UMLS and other terminologies. *Journal of the American Medical Informatics Association*. 2020 Oct 1;27(10):1593-1599. doi: 10.1093/jamia/ocaa180. PMID: 32930711; PMCID: PMC7647355.
- Williams G (co-first), Maorify V (co-first), **Rasmy L (co-first)**, Brown D, Yu D, Zhu H, Talebi Y, Wang X, Thomas E, Zhu G, Yaseen A, Zhi D, Aguilar D, Wu H. Vasopressor treatment and mortality following nontraumatic subarachnoid hemorrhage: a nationwide electronic health record analysis. *Neurosurgical Focus*. 2020;48(5):E4. doi: 10.3171/2020.2.FOCUS191002. PMID: 32357322.
- Rasmy L**, Wu Y, Wang N, Geng X, Zheng WJ, Wang F, Wu H, Xu H, Zhi D. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR dataset. *Journal of Biomedical Informatics*. 2018 Aug;84:11-16. doi: 10.1016/j.jbi.2018.06.011. Epub 2018 Jun 15. PMID: 29908902; PMCID: PMC6076336.
- Nigo M, **Rasmy L**, May S, Rao A, Karimaghaei S, Mao B, Kannadath B, Hoz A, Arias C, Li L, Zhu D. Real world assessment of the efficacy of tocilizumab in patients with COVID19: results from a large de-identified multicenter electronic health record dataset in the United States. *International Journal of Infectious Diseases*. 2021; S1201-9712. doi: 10.1016/j.ijid.2021.09.067. PMID: 34597766; PMCID: PMC8479513.
- Xiang Y, Ji H, Zhou Y, Li F, Du J, **Rasmy L**, Wu ST, Zheng WJ, Xu H, Zhi D, Zhang Y, Tao C. Asthma exacerbation prediction and risk factor analysis based on a time-sensitive, attentive neural network: retrospective cohort study. *Journal of Medical Internet Research*. 2020 Jul 31;22(7):e16981. doi: 10.2196/16981. PMID: 32735224; PMCID: PMC7428917.
- Xiang Y, Xu J, Si Y, Li Z, **Rasmy L**, Zhou Y, Tiriyaki F, Li F, Zhang Y, Wu Y, Jiang X. Time-sensitive clinical concept embeddings learned from large electronic health records. *BMC Medical Informatics and Decision Making*. 2019 Apr 9;19(Suppl 2):58. doi: 10.1186/s12911-019-0766-3. PMID: 30961579; PMCID: PMC6454598.

Field of Study

Health Informatics

Table of Contents

Dedication	ii
Acknowledgements	iii
Abstract	iv
Vita	vi
Table of Contents	ix
List of Tables	xv
List of Figures	xvii
Chapter 1: Introduction	1
1.1. Secondary Use of Electronic Health Records Structured Data	1
1.1.1. Collaborative initiatives and common data models	2
1.1.2. Commercial secondary EHR data sources	4
1.2. Secondary Use of Structured EHR Data for Observational Studies	5
1.3. Secondary Use of Structured EHR Data for Predictive Modeling	6
1.3.1. Deep learning for clinical events prediction	7
1.3.2. Terminology Normalization of Clinical Data for EHR Based Studies	8
1.3.3. Foundation deep learning models for clinical data	9

1.4. Innovation	10
1.5. Significance.....	11
Chapter 2: Vasopressor treatment and mortality following nontraumatic subarachnoid hemorrhage: a nationwide electronic health record analysis	
2.1. Abstract	14
2.2. Introduction.....	15
2.3. Methods.....	17
2.3.1. Study design and patient population	17
2.3.2. Statistical and causal inference analysis methods.....	19
2.3.3. Sensitivity analysis.....	21
2.4. Results.....	22
2.4.1. Baseline demographics and vital signs	24
2.4.2. Comorbidities and outcomes.....	24
2.4.3. Glasgow Coma Score (GCS)	24
2.4.4. Vasopressor treatment effects based on causal inference analysis	26
2.4.5. Sensitivity analyses	27
2.5. Discussion	30
2.6. Conclusion	33
2.7. Acknowledgments.....	33

2.7.1. Sources of funding	34
2.7.2. Disclosures	34
2.8. Supplementary Material	35
2.9. References	50
Chapter 3: Representation of EHR data for predictive modeling: a comparison between	
UMLS and other terminologies	57
3.1. Abstract	58
3.2. Background	59
3.3. Objective	61
3.4. Methods.....	62
3.4.1. Prediction tasks and cohort description	62
3.4.2. Diagnosis terminology	64
3.4.3. Tasks and models	67
3.4.4. Statistical analysis for model comparison.....	68
3.5. Results.....	68
3.5.1. Effect of information loss due to terminology mapping	71
3.6. Discussion	72
3.7. Conclusion	76
3.8. Acknowledgments.....	77

3.8.1. Funding	77
3.8.2. Authors' contribution.....	78
3.8.3. Competing interests	78
3.9. Supplementary Material.....	78
Appendix A. Cohort Definitions.....	78
Appendix B. Terminology mappings.....	82
Appendix C. Recurrent Neural Network model architecture.....	85
Appendix D. Diabetes Heart Failure full Cohort Results	87
Appendix E. DHF LR Additional Results	88
Appendix F. Statistical Significance using Tukey-Kramer HSD	89
3.10. References.....	92
Chapter 4: Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction.....	98
4.1. Abstract	99
4.2. Introduction.....	100
4.3. Methods.....	107
4.3.1. Cohort Definition	107
4.3.2. The Data Modality of Structured EHR	112
4.3.3. Med-BERT Architecture.....	114

4.3.4. Pre-training Med-BERT.....	116
4.3.5. Applying Med-BERT for Downstream Prediction Tasks by Fine-tuning....	118
4.3.6. Evaluation of Med-BERT	119
4.3.7. Implementation Details	121
4.3.8. Code Availability	122
4.4. Results.....	123
4.4.1. Data Source	123
4.4.2. Performance Boost of Med-BERT on Fine-tuning Tasks.....	124
4.4.3. Visualization of Attention Patterns in Med-BERT	129
4.5. Discussion	133
4.6. Acknowledgments.....	137
4.6.1. Competing Interests	138
4.6.2. Authors Contributions.....	138
4.7. Supplementary material	139
Supplementary 4 Table 1: Additional benchmark results.....	141
Supplementary 4 Table 2: Additional performance results	144
4.8. References.....	155
Chapter 5: Conclusions, Discussions, and Recommendations	162
5.1. Use of secondary EHR data for evidence-based medicine	162

5.2. Sequential Deep Learning Modeling using EHR data.....	164
5.2.1. Recurrent neural network-based models.....	164
5.2.2. Transformers.....	165
5.2.3. Model explainability	166
5.3. Contribution to Science.....	167
5.4. In-progress Work and Future Directions	168
5.4.1. Pytorch_ehr framework to train and evaluate an implementable deep learning predictive model.....	170
5.4.2. Implementability evaluation factors	172
5.5. Final Thoughts	176
References.....	178

List of Tables

Table 2.1. Baseline demographics and clinical characteristics/outcomes of non-traumatic aneurysmal SAH patients by vasopressor treatment group	
Table 2.2. Percentage of patients according to GCS scores in each of the three treatment groups and mortality breakdown by treatment group and GCS categories	
Table 2.3. Association of treatment with mortality using a logistic regression model (unweighted), and propensity adjustment via IPW (weighted) for nontraumatic aneurysmal SAH subjects, adjusted for age, race, marital status, and sex	
Table 2.4. Logistic regression for phenylephrine and norepinephrine versus dopamine (the reference treatment) adjusted for variables selected using LASSO and stepwise methods	
Table 3.1. Description of cohorts.....	
Table 3.2. Prediction performance of different diagnosis terminologies for the DHF and PC tasks.....	
Table 3.3. Differences in AUROC between primary mapping to ICD-9/10 codes and reversed mapping to ICD-9'/10' codes.....	
Table 4.1: Comparison of Med-BERT with BEHRT and G-BERT from multiple perspectives.....	
Table 4.2: Comparison of characteristics of EHR data versus Natural language data	
Table 4.3. Descriptive analysis of the cohorts.	
Table 4.4. Average AUC values and standard deviations (in parentheses) for the different methods for the three evaluation tasks.....	
Table 5.1. Deep learning based models Implementability evaluation factors	
Supplementary 2 Table 1: National Drug Codes for Vasopressors used for the Study	
Supplementary 2 Table 2: Association of treatment with mortality for non-traumatic SAH subjects in the subgroup that had baseline blood pressure and heart rate.....	

Supplementary 2 Table 3: Association of treatment with mortality for non-traumatic SAH subjects in the subgroup that missing baseline blood pressure or heart rate.	
Supplementary 2 Table 4: Association of treatment with mortality for non-traumatic SAH subjects in the subgroup having Glasgow Coma Score recorded	
Supplementary 2 Table 5: Distribution comparison of demographic variables and Glasgow Coma Score after propensity matching between phenylephrine and dopamine treatment groups.....	
Supplementary 2 Table 6: Distribution comparison of demographic variables and Glasgow Coma Score after propensity matching between phenylephrine and norepinephrine treatment groups	
Supplementary 2 Table 7: Distribution comparison of demographic variables and Glasgow Coma Score after propensity matching between norepinephrine and dopamine treatment groups.....	
Supplementary 2 Table 8: Mortality by treatment class separated by the cumulative comorbidity of at least one of the three diseases MI, ARF, or Sepsis	
Supplementary 2 Table 9: Mortality based on vasopressor choice for patients with only one of the three vasopressor prescription.....	
Supplementary 2 Table 10: Shows the percentage of patients with one, two, or three different vasopressors within each of the three treatment groups.....	
Supplementary 3 Table 1: LR and RNN results of DHF full cohort	
Supplementary 3 Table 2: Results of different variations of LR models the DHF cohort	

List of Figures

Figure 1.1. PubMed indexed publications per year using MIMIC versus using secondary EHR data sources available through collaborative initiatives such as i2b2, OHDSI OMOP, PCORnet, and eMERGE	
Figure 1.2. PubMed indexed publications on EHR data based predictive models per year.	
Figure 3.1. Terminology conversion roadmap.....	
Figure 3.2. Significance of AUROC difference.....	
Figure 4.1. An example of structured EHR data of a hypothetical patient as it would be available from a current EHR system	
Figure 4.2. Med-BERT structure.	
Figure 4.3. Selection pipeline for the pre-training cohort.....	
Figure 4.4. Comparison of prediction AUC for the test sets by training on different sizes of data on various Cohorts between the methods with or without the pre-trained Med-BERT layer. Logistic regression (LR) results are included as a baseline.....	
Figure 4.5. Example of different connections of the same code, “type 2 diabetes mellitus,” in different visits.....	
Figure 4.6. Example of the dependency connections in the DHF-Cerner cohort.....	
Figure 4.7. Example of the dependency connections in the PaCa-Cerner cohort.....	
Figure 5.1. Pytorch_EHR Framework	
Figure 5.2. CovRNN Data Mechanics	
Supplementary 2 Figure 1. Graphical representation of covariate balance as measured by the standardized mean difference (SMD) within the unweighted and weighted data	
Supplementary 2 Figure 2. Mortality rate vs number of admitted SAH patients by hospital	
Supplementary 3 Figure 1. Flowchart for DHF cohort definition	
Supplementary 3 Figure 2. Flowchart for PC cohort definition	
Supplementary 3 Figure 3. RNN based model training.....	
Supplementary 4 Figure 1. Flowchart for the DHF cohort definition.	

Supplementary 4 Figure 2. Flowchart for the PaCa cohort definition.....	
Supplementary 4 Figure 3. Attention connections from the first three transformer layers (a top-down direction) of a sample patient sequence.....	

Chapter 1: Introduction

1.1. Secondary Use of Electronic Health Records Structured Data

Ancient Egyptian papyri and inscriptions indicate the use of medical records as early as 1,600 BC[1,2]; however, the use of paper medical records did not become an established practice until the 19th century, when it began in France and Germany[2]. By the early 90th, with advances in computer technology and advocacy by the Institute of Medicine, academic healthcare systems in the United States started to shift from paper-based to electronic health records(EHRs)[1,2], which was fostered by government initiatives and incentive programs, such as Meaningful Use (MU) by the Centers for Medicare & Medicaid Services (CMS) and the Office of the National Coordinator for Health Information Technology (ONC) in 2010[3].

The main purpose of EHR utilization is to provide authorized users with secure access to patients' real-time data to improve the efficiency and the quality of care, including the coordination of care while protecting patient privacy. Based on the increasing adoption rate of EHR systems in the US hospitals, which reached 84% in 2015[4], healthcare organizations and vendors could begin to compile rich clinical data and make them available for biomedical informatics researchers to mine them, extract information, and create knowledge. In this regard, “secondary” EHR data refer to the clinical data warehouses that extract and combine data from different sources and healthcare systems and make them useable by researchers.

1.1.1. Collaborative initiatives and common data models

There are several initiatives supported by governmental institutions and non-profit organizations to facilitate the wide collection of EHR data from different sites and harmonize them through common data models (CDM), in order to support translational research[5]. One of the earliest initiatives is the informatics for integrating biology at the bedside (i2b2) which started in 2004 as a part of the National Institutes of Health (NIH) Roadmap initiative[6]. I2b2 schema offers the flexibility to hold denormalized non-standard and local data[7]. I2b2 is utilizing the star schema format commonly used in retail data warehouses, which is characterized by the large narrow fact tables that include all individual observations along with ontology tables that provide definitions such as concept_dimension and tables defining the hierarchical arrangements of concepts such as modifier_path[8]. Over 200 institutions worldwide are utilizing i2b2 CDM and ancillary software[9] for data linkage and harmonization to form large federated data networks that can facilitate international clinical research such as the consortium for clinical characterization of COVID-19 by EHR (4CE) studies to understand COVID-19 clinical trajectories[10,11]. The most widely adopted CDM nowadays is the observational medical outcomes partnership (OMOP) CDM developed by the observational health data sciences and informatics (OHDSI) consortium[8]. Some of the largest and latest collaborative initiatives such as the All of Us research program[12,13] and the national COVID cohort collaborative (N3C)[14,15] are following the specifications of the OMOP data model. OMOP CDM is optimized for typical observational research purposes as it

provides clinical events tables such as condition occurrences, drug exposures, and observations, as well as derived elements tables such as condition era and drug era, and standardized health economics tables[16]. That's beside metadata and vocabulary tables which include terminology information and concepts relationships[17]. Other government-supported initiatives include the electronic medical records and genomics (eMERGE) consortium funded and organized by the NIH[18,19], the PCORnet supported by the Patient-Centered Outcome Research Institute (PCORI) [20,21], and the biologics effectiveness and safety system (BEST) supported by the FDA center for biologics evaluation and research (CBER)[22,23]. The majority of such collaborative initiatives are providing tools to facilitate data exchange between different CDM, for example, i2b2 offers a “multi-fact-table querying” feature that can query OMOP and PCORnet models[7,24].

Similar to the US government-supported initiatives, The UK department of health and social care sponsored the clinical practice research datalink (CPRD) initiative through the medicines and healthcare products regulatory agency (MHRA) and the national institute for health research (NIHR). CPRD collects de-identified patient data from primary care practices throughout the UK and links it to other healthcare data sources available through the National Health Services (NHS). As appears in Figure 1.1, CPRD is the most cited secondary EHR data source in PubMed indexed articles throughout the period from 2014 to 2019, while the freely available medical information mart for intensive care dataset (MIMIC) is the most cited data source in 2020 and 2021.

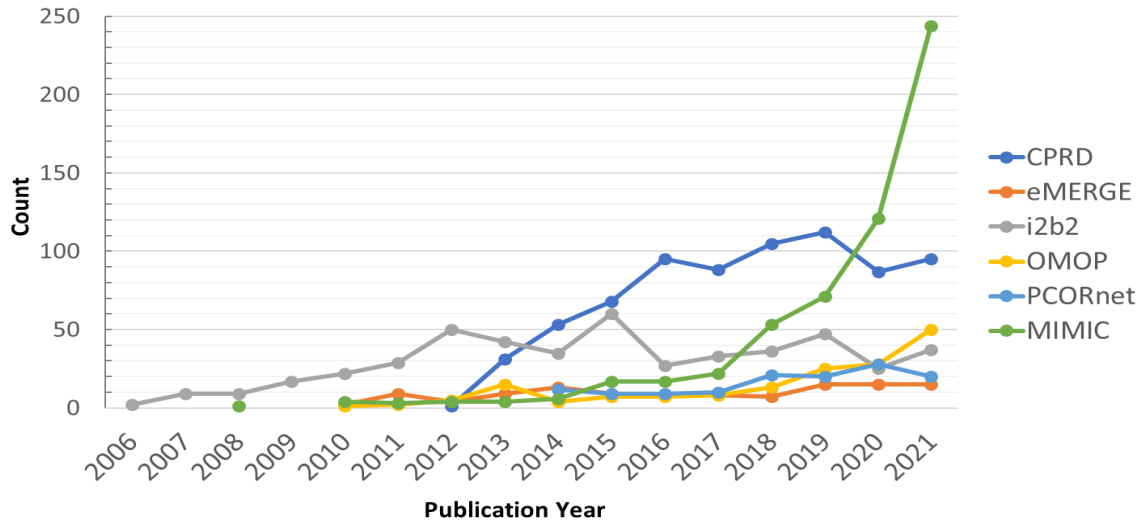


Figure 1.1. PubMed indexed publications per year using MIMIC versus using secondary EHR data sources available through collaborative initiatives such as i2b2, OHDSI OMOP, PCORnet, and eMERGE.

1.1.2. Commercial secondary EHR data sources

As large EHR vendors such as Cerner[25] and EPIC[26–28] and big insurance companies such as Blue Cross Blue Shield (BCBS)[29] envision the powerful impact of big clinical data to transform healthcare through data-informed decisions, they started to offer clinicians and researchers access to multi-institutional data resources to facilitate clinical research using secondary EHR and claims data. However, access to such powerful resources is conditional either through sharing data such as joining the Cerner learning health network (LHN) or through paying licensed access fees. There are several secondary EHR and claims data vendors[30] including the popular IBM MarketScan[31], Optum[32], and IQVIA[33].

Cerner HealthFacts®, an earlier version of the current Cerner real-world data (CRWD), is the main source for secondary EHR data that we used to extract the cohorts described in chapters 2,3, and 4. Cerner HealthFacts® is a database of de-identified structured EHR data from approximately 50 million patients from more than 600 hospitals across the United States. Available de-identified patient data include diagnosis, medications, procedures, laboratory results, and other recorded clinical events as well as patient demographics and encounter-level administrative data. For generalizability evaluation of Med-BERT described in chapter 4, we used the Truven® claims database, an earlier version of the IBM® MarketScan® Research Databases. The Truven® version which we used, contains individual-level, de-identified, healthcare claims information from employers, health plans, hospitals, and Medicare and Medicaid programs, for the period between 2011 and 2015. Although we acknowledge the differences between secondary EHR and claims data[34,35], the use of a claims dataset to evaluate the generalizability of a foundation model trained on secondary EHR data was appropriate as the foundation model was only trained on diagnosis codes which are consistent between EHR and claims data, that's beside the cohort definition was only based on age and diagnoses codes.

1.2. Secondary Use of Structured EHR Data for Observational Studies

With the increased availability of secondary EHR data sources as described in section 1.1, hundreds of observational studies were conducted to compare patterns and trends between different populations and different interventions[36,37]. However, extracting new evidence from secondary EHR data is challenging given the data quality limitations of inconsistency, incompleteness, and inaccuracy[38–41]. Therefore, we need to account

for the data quality and check for the underlying model assumptions to verify the reliability of our findings[42].

In Chapter 2, we provide an example of how we extracted new evidence from secondary EHR data[43]. The research question for this study was raised by neuro-intensivists, who note that, although current evidence-based guidelines suggest that maintaining a vasopressor-induced elevation in blood pressure, when managing aneurysmal subarachnoid hemorrhage (SAH), may reduce the incidence of delayed cerebral ischemia (DCI), there is no clear evidence, as described in Chapter 2 sections 2.2. and 2.5, in regard to which vasopressor is associated with better outcomes, i.e., reduced long-term adverse outcomes and lower mortality risk. Therefore, the main objective of this study was to determine the association between the initial vasopressor choice and in-hospital mortality of nontraumatic SAH patients, based on a comparison of the three most commonly used vasopressors, namely dopamine, norepinephrine, and phenylephrine.

1.3. Secondary Use of Structured EHR Data for Predictive Modeling

Similar to the increased numbers of retrospective observational studies using secondary EHR data sources, there are hundreds of published studies that describe predictive models trained on secondary EHR data sources[44–48]. However, the majority of the predictive models employ few predictors and do not fully address biases due to training data quality issues such as incompleteness[49]. The majority of the recently published predictive models are based on machine learning techniques, with approximately 20% based on deep learning, as seen in Figure 2.1. In Chapters 3 and 4 we utilized sequential

deep learning architectures to train predictive models using tens of thousands of categorical predictors. Our proposed models showed improved prediction accuracy and generalizability compared to baseline machine methods such as logistic regression.

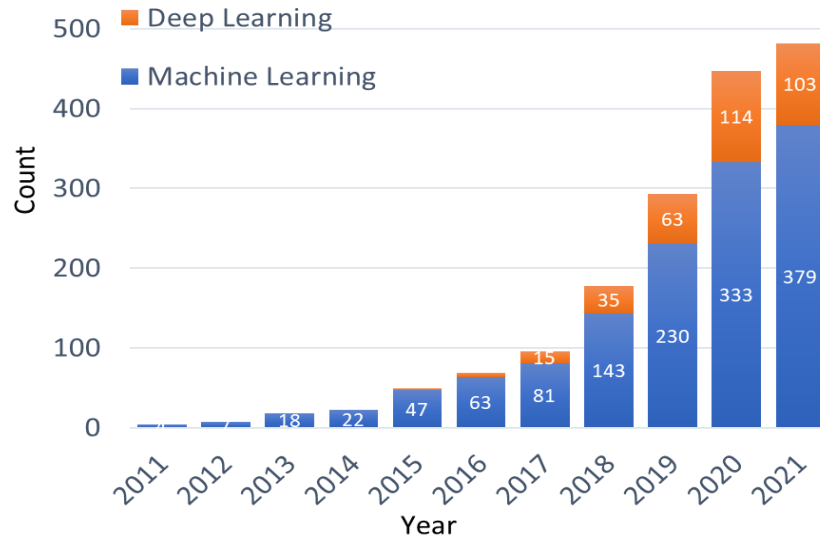


Figure 1.2. PubMed indexed publications on EHR data based predictive models per year

1.3.1. Deep learning for clinical events prediction

In the last decade, hundreds of articles were published on the development of deep learning(DL)-based models that used EHR data in either a structured[50–57] or unstructured format[58–61]. Most of these developed models provide promising prediction accuracy and demonstrate that DL algorithms outperform other standard statistical or machine learning algorithms[62–64]. Such models, however, are rarely implemented in practice, and most clinical decision support systems are either rule-based

or depend on simple algorithms such as logistic regression, which uses only a certain number of features.

The lack of adoption of DL-based models was justified by the “black box” nature of DL algorithms, i.e., the lack of interpretability required by physicians to understand the predicted risk scores and to assess their reliability[65–68]. Although many techniques had been developed to improve the interpretability/explainability of DL-based models, the implementation of such models is still rare[69–72]. Notably, there are other elements to be considered in regard to improving the implementability of DL-based predictive models.

In Chapters 3 and 4, we present a sample of our work, in which we demonstrate that DL-based algorithms have superior performance as compared with other traditional machine learning methods. In addition, we address the issue of the generalizability of the trained models from a different perspective, which is an important factor in the overall implementability of the proposed models.

1.3.2. Terminology Normalization of Clinical Data for EHR Based Studies

A common obstacle to the evaluation of predictive model generalizability is that clinical data at different sites are available in different formats. For example, some sites prefer to use SNOMED-CT to record patient diagnoses, while others prefer to use ICD-10-CM codes. This can limit the transferability of the trained models between hospitals as well as limit further external validation to evaluate their generalizability. Therefore, there is a need to normalize different clinical data types to a common terminology. One way to do so is to

map the data to the unified medical language system (UMLS)[73] or another common data model[8,74,75]. Other solutions include mapping each clinical data category to the most common standard terminology in use for that category, for example, mapping laboratory tests to LOINC codes.

In Chapter 3, we investigated the impact of different transformation methods on the performance of logistic regression and DL-based predictive models by comparing the prediction discriminative accuracy of the models when trained and evaluated on the same patients’ data but in different terminologies. In this study[76], we mapped the diagnosis codes recorded in Cerner HealthFacts®, mainly in ICD-9 and ICD-10-CM format, into UMLS concept unique identifiers (CUIs)[77], ICD-9[78], ICD-10[78], PheWAS[79], and CCS codes in both the single-level[80,81] and the refined version (CCSR)[82]. We evaluated the effect of terminology mapping on two different disease prediction tasks. We refer to the first task, which was to predict diabetic patient risk to develop heart failure, as the *DHF task*. We refer to the second task, which involved predicting the patient risk of being diagnosed with pancreatic cancer in the next visit, as the *PaCa task*. The results of this study helped us to better understand the need for and how to normalize categorical clinical data, including diagnoses, medications, and procedures, for predictive modeling.

1.3.3. Foundation deep learning models for clinical data

Foundation models, such as bidirectional encoder representations from Transformers (BERT), are models trained on broad data at scale such that they can be adapted to a wide range of downstream tasks[83]. Such models merge self-supervised learning and transfer-learning concepts. Transformers proved to be a valid alternative for sequential modeling.

Since the end of 2018, after Google released its first pre-trained BERT model, based on the transformer structure, transformer-based models have continued to evolve and serve as the basis to train large foundation models.

In Chapter 4, we describe our foundation model, Med-BERT[84], which is trained on structured EHR diagnosis data from more than 20 million patients, extracted from Cerner HealthFacts®, and how such pre-trained contextualized embedding can improve the prediction performance for different downstream tasks. We used the same DHF and PaCa prediction tasks as the evaluation downstream tasks. We also used the Truven MarketScan™ claims data to evaluate the generalizability of our trained foundation model. In addition, we evaluated prediction performance, using multiple metrics, including the area under the receiving operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), sensitivity, specificity, precision, and F1-score.

1.4. Innovation

The key innovation of our work involves the use of secondary EHR data for knowledge discovery and the development of a highly accurate DL-based model to predict patient risk for diseases such as heart failure and pancreatic cancer. We are the first to take advantage of access to a large heterogeneous EHR database and to extract a cohort to compare the effect of phenylephrine, dopamine, and norepinephrine on nontraumatic SAH patient outcomes, using novel statistical methods[85]. Similarly, we are the first to use such a large heterogeneous database to train accurate and generalizable DL predictive models[84,86].

Translational research has always been an area in which the major focus was on transferring findings from laboratory settings to actual clinical trials in hospital settings. Several concerns, however, hinder the feasibility of transferring our findings and trained models to the clinical settings for further evaluation. Therefore, in this dissertation, we addressed some of the concerns that may hinder further external validation of our trained models. We are the first to systematically compare the effect of terminology normalization to coding standards with different levels of granularity on the performance of predictive models. We are also the first to train a medical foundation model, using structured diagnosis data in ICD-9 and ICD-10 codes, which are globally accepted standards, for more than 20 million patients. Finally, we are the first to evaluate the generalizability of our trained foundation model. Although we did not tackle all of the concerns that hinder the further validation of the developed models, we took the first steps.

1.5. Significance

With the advancement of the science of artificial intelligence and the availability of powerful computational resources and large observational healthcare databases, we can extract knowledge and develop more accurate, generalizable, and personalized prediction models. Our findings on the vasopressor effect on non-traumatic SAH patients' outcomes favored phenylephrine, which is known to be the least potent among the studied vasopressors. These findings are promising but require validation through clinical trials, after which the clinical practice guidelines should be updated accordingly. We also demonstrated the value of utilizing a large heterogeneous healthcare database to develop

accurate and generalizable predictive models. Our evaluation of the impact of terminology normalization on the predictive models' performance indicated the value of using more expressive terminologies with a high level of granularity. The use of data in their raw format can achieve good performances, and future researchers can draw upon such raw data directly, especially if they have a large sample size. If, however, their methods require much lower input dimensions, such as certain statistical algorithms, we recommend the use of PheWAS over CCSR or CCS. When integrating with other data modalities, such as unstructured text, is desired, we recommended the normalization to UMLS. We disseminate the details of our methods, along with our findings and codebase, to facilitate the reproducibility of our work and allow further improvement and validation by future researchers. Notably, our Med-BERT work demonstrated the value of training large foundation models on large heterogeneous healthcare databases. Our publicly available pretraining codebase is in use by researchers all over the world to train their own foundation models and further evaluate these models on more practical downstream tasks.

Chapter 2: Vasopressor treatment and mortality following nontraumatic subarachnoid hemorrhage: a nationwide electronic health record analysis

This chapter is adapted from:

Williams G (co-first), Maroufy V (co-first), **Rasmy L (co-first)**, Brown D, Yu D, Zhu H, Talebi Y, Wang X, Thomas E, Zhu G, Yaseen A, Zhi D, Aguilar D, Wu H. Vasopressor treatment and mortality following nontraumatic subarachnoid hemorrhage: a nationwide electronic health record analysis. *Neurosurgical Focus*. 2020;48(5):E4. doi: 10.3171/2020.2.FOCUS191002. PMID: 32357322.

Title: Vasopressor Treatment and Mortality Following Non-Traumatic Subarachnoid Hemorrhage: A Nationwide EHR Analysis

Authors: George Williams, M.D.^{1#}; Vahed Maroufy, Ph.D.^{2#}; Laila Rasmy, M.Sc.^{3#}; Derek Brown²; Duo Yu, M.Sc.²; Hai Zhu²; Yashar Talebi²; Xueying Wang²; Emy Thomas²; Gen Zhu, M.Sc.²; Ashraf Yaseen, Ph.D.²; Hongyu Miao, Ph.D.²; Luis Leon Novelo, Ph.D.²; Degui Zhi, Ph.D.^{2,3}; Stacia DeSantis, Ph.D.²; Hongjian Zhu, Ph.D.²; Jose-Miguel Yamal, Ph.D.^{2*}; David Aguilar, M.D.^{1,2*}; Hulin Wu, Ph.D.^{2,3*}

- 1- McGovern Medical School, University of Texas Health Science Center, Houston, TX.
- 2- School of Public Health, University of Texas Health Science Center, Houston, TX.
- 3- School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX.

Co-First Authors

* Co-Corresponding Authors

Corresponding Author:

Hulin Wu, Ph.D
Department of Biostatistics and Data Science
The University of Texas School of Public Health

Key Words: Subarachnoid hemorrhage, vasopressors, Norepinephrine, Phenylephrine, Dopamine, mortality

2.1. Abstract

Background and Object: Subarachnoid hemorrhage (SAH) is a devastating cerebrovascular condition not only due to the effect of initial hemorrhage, but also due to the complication of delayed cerebral ischemia. While hypertension facilitated by vasopressors is often initiated to prevent delayed cerebral ischemia, which vasopressor is most effective in improving outcomes is not known. The objective of this study was to determine associations between initial vasopressor choice and mortality in patients with non-traumatic subarachnoid hemorrhage.

Methods: We conducted a retrospective cohort study using a large, national EMR Dataset from 2000-2014 to identify patients with a new diagnosis of non-traumatic subarachnoid hemorrhage (based on ICD-9 codes) that were treated with vasopressors (dopamine, phenylephrine or norepinephrine). We examined the relationship between the initial choice of vasopressor therapy and the primary outcome, which was defined as in-hospital death or discharge to hospice care.

Results: In total, 2,634 patients were identified with non-traumatic subarachnoid hemorrhage who were treated with a vasopressor. In this cohort, the average age was 56.5 years, and 63.9% were female. 36.5% of patients developed the primary outcome. The incidence of the primary outcome was higher in those initially treated with either norepinephrine (47.6%) or dopamine (50.6%) than with phenylephrine (24.5%). After adjusting for possible confounders using propensity score methods, the adjusted odds ratio (OR) of the primary outcome was higher with dopamine (OR 2.19, 95% CI: 1.70-2.81) and norepinephrine (OR 2.24, 95% CI: 1.80-2.80), compared with phenylephrine.

Sensitivity analyses using different variable selection procedures, causal inference models, and machine learning methods confirmed the main findings.

Conclusions: In patients with non-traumatic subarachnoid hemorrhage, phenylephrine was significantly associated with reduced mortality in SAH patients compared to dopamine or norepinephrine. Prospective randomized clinical studies are warranted to confirm this finding.

2.2. Introduction

Aneurysmal subarachnoid hemorrhage (SAH) is a potentially devastating cerebrovascular condition due not only to the effect of initial hemorrhage but also the complicated treatment regimen required to manage such patients[1]. The incidence of SAH in the population has been estimated to be around 14.5 per 100,000 person years with a mean age of 55 years[2,3]. As a result of the pathophysiology of this disease and its treatments, 25-44% of patients who present with SAH die, and half of the survivors are left with some degree of neurological deficit[4–6]. The average age of onset for SAH may be young, though this particular type of stroke syndrome is associated with traditional risk factors, such as hypertension, which predispose patients to active aneurysmal disease[7]. Substantial resources are dedicated to currently accepted management paradigms, which include staffing by neurosurgeons, neurologists, neuro-radiologists, neuro-intensivists, neuro-anesthesiologists, and specialized nursing and rehabilitation personnel. Given the relatively young age of patients with SAH and associated disability, the health and economic burden to the individual can be devastating.

One of the complications of SAH that medical management seeks to avoid or minimize is delayed cerebral ischemia (DCI), as it is the main source of morbidity following SAH[8]. Decreased cerebral perfusion (relative hypotension) and cerebral arterial vasospasm that may commonly occur after SAH are felt to be significant contributors to DCI. While 70% of SAH patients show signs of radiographic cerebral vasospasm, about 20-30% of SAH patients have clinical signs of cerebral vasospasm[9]. There are several proposed mechanisms contributing to cerebral vasospasm, including damage to the endothelium, smooth muscle contraction, changed vascular responsiveness, and inflammatory changes to the vascular wall[9]. Besides the administration of nimodipine prior to clinical evidence of cerebral vasospasm, the avoidance of hypovolemia and hypotension are accepted as mainstays of therapy when cerebral vasospasm is suspected. Induced hypertension to reduce the incidence of DCI was first described in the 1970s and its benefit is clinically accepted by many neuro-intensivists[10]. A goal systolic blood pressure of 160-180 mmHg is a widely accepted blood pressure target for SAH patient treatment following coiling of a cerebral aneurysm; this hemodynamic approach combined with avoidance of hypovolemia is commonly described as Hypertensive Hypervolemic Therapy (HHT)[11,12]. Recent studies showed that Vasopressor induced hypertension is recommended over only fluid regimen[13,14]. Dopamine, norepinephrine, and phenylephrine are the most commonly accepted drugs to achieve the desired increase in blood pressure[15]. Vasopressin is occasionally used, but is not generally accepted as a mainstay of therapy[16]. While current guidelines[12] suggest that maintaining an elevation in blood pressure when managing aneurysmal SAH may reduce the incidence

of DCI, little clinical data exists to demonstrate which vasopressor is most efficacious to achieve induced hypertension and ultimately reduce mortality and other long-term adverse outcomes. In the absence of clinical trial data, the optimal vasopressor choice is based on multiple other factors, including the patient's hemodynamic status, comorbid conditions, and institutional preferences.

Cerner Health Facts® EMR database comprises de-identified EHR data from over 600 participating Cerner client hospitals and clinics in the United States including patient demographics, encounters, diagnoses, procedures, lab results, medications, vital signs, and other clinical observations[17,18]. The large population in this database, approximately 50 million patients, allows for performing in-depth studies on rare disorders with sufficient power to detect clinically meaningful effect sizes in which clinical trials may be difficult to perform. In this study, we queried the Cerner database in order to determine the association of vasopressor choice with in-patient mortality. We hypothesized that the choice of vasopressor is associated with mortality for SAH patients after adjusting for possible confounding factors.

2.3. Methods

2.3.1. Study design and patient population

Because of the sensitive nature of the data collected for this study, requests to access the dataset from qualified researchers trained in human subject confidentiality protocols may be sent to Hulin Wu, PhD at [Hulin.Wu@uth.tmc.edu]. Following institutional IRB approval (IRB HSC-MS-18-0124), we used the Cerner Health Facts® EMR database,

which consists of EMRs from more than 700 hospitals and clinics that use the Cerner Corporation's electronic health record system. Cerner Health Facts® EMR database (version 2015) includes around 50 million unique patients (49,826,219) out of which 39,017 patients had at least one encounter with one SAH diagnosis identified by ICD-9 code (430,800.2x, 800.7x, 801.2x, 801.7x, 803.2x, 803.7x, 804.2x, 804.7x, 852.x) with 17,273 aneurysmal SAH. We excluded all patients younger than 17 years old. A total of 4,850 patients were administered at least one of the three vasopressors: dopamine, phenylephrine, or norepinephrine [National Drug Codes (NDC) for included vasopressors are shown in Supplementary Table 1]. If a patient had multiple SAH encounters within 24 hours of each other, we combined the encounters into a single encounter; otherwise, we only used the data from the first SAH encounter per patient in our analysis.

In this study, we examined the first vasopressor administered to eligible SAH patients. We excluded subjects who initiated vasopressor treatment with two vasopressors simultaneously (n=40). Among the remaining 4,810 SAH patients, 3,078 (64%) patients were prescribed only one of the three vasopressors during the encounter period, 1,437 (30%) patients received a second vasopressor after the initial vasopressor, and 295 (6%) patients were treated with all three vasopressors after the initial vasopressor. We grouped the patients into either dopamine, phenylephrine, or norepinephrine groups based on the initial vasopressor received, regardless of whether they subsequently received a different vasopressor. Among 4,810 eligible SAH patients, 2,176 (45.2%) patients had trauma associated SAH (ICD 9 codes: 800.2x, 800.7x, 801.2x, 801.7x, 803.2x, 803.7x, 804.2x, 804.7x, 852.x) and 2,634 patients had a non-traumatic SAH diagnosis. We analyzed SAH

patients and non-traumatic SAH patients separately, which resulted in similar conclusions; this report will focus on non-traumatic SAH patients.

2.3.2. Statistical and causal inference analysis methods

The primary outcome was in-hospital mortality, defined as in-hospital death or discharge to hospice care. We calculated the Charlson Comorbidity Index (CCI) to provide an indicator of clinical morbidity across the populations. We also compared descriptive statistics for baseline, demographic, comorbidity, and outcome differences among the three treatment groups. For continuous variables, we used ANOVA or Kruskal-Wallis test, depending on whether the ANOVA assumptions were met. For categorical variables, Fisher's exact test was used.

As high grade SAH patients have prolonged ICU stays with multiple nosocomial complications such as pneumonia, anemia, infection/sepsis, renal failure, myocardial infarction, and heart failure (cardiomyopathy), we compared proportions of these diagnoses among the treatment groups in a secondary analysis. Additionally, we included diabetes and liver disease as they are in general associated with worse outcomes. Furthermore, in order to account for patient severity, we included Glasgow Coma Score (GCS), which has been shown comparable to *Hunt and Hess* and *World Federation of Neurological Surgeons Scales*, and a significant factor for predicting severity at discharge[19].

We used propensity score models to account for potential confounding variables and facilitate causal inference between vasopressor treatments and mortality. In a binary treatment case, the propensity score is the probability of receiving the treatment

conditional on a given set of observed potential confounding factors[20]. This probability can be calculated using standard regression techniques (typically logistic regression), with the treatment being considered the outcome and the potential confounding factors as the predictors. Treated and control subjects with similar estimated values for their propensity scores should have, on average, similar sets of covariate factors[21]. We used inverse probability weighting (IPW) in the propensity score model, given its well-studied benefits[22,23]. We used propensity score matching and generalized gradient boosted models (GBM)[24], a machine learning approach based on trees that selects predictive variables while higher order variable interactions are automatically taken into account, to create the propensity score. Then we used this machine learning approach to estimate the average treatment effect (ATE) inverse probability weights to balance treatment groups. We estimated propensity score vectors representing the probability of receiving each of 3 treatments (i.e., the generalized propensity score, GPS) using GBM. The potential confounders we included in the propensity model are age, race, marital status, gender, GCS, 750 medications (generic names) administered prior to the vasopressor treatment, CCI, and total IV-fluid administered throughout the encounter. Also, to account for other complications, we adjusted for the 443 diagnoses which were present during admission across all patients. Prior to GBM, we conducted variable selection using L1-penalized generalized linear models (GLM Lasso)[25] so that the significant confounders associated with the choice of vasopressor treatment were identified. We then applied GBM to the selected variables, generated the inverse probability weights (where the weight is the inverse propensity of the treatment an individual actually received[26,27])

in weighted regression models[24] and estimated ATEs of the vasopressor treatments on mortality.

We assessed the balance achieved through inverse probability weighting by graphical representations (Supplementary Figure 1). For each pairwise treatment comparison (dopamine vs phenylephrine, dopamine vs norepinephrine, and phenylephrine vs norepinephrine), we calculated a standardized mean difference before (unweighted) and after implementation of IPW. Then, for each selected confounder we graphed the maximum standardized mean difference across the three vasopressor treatment comparisons. From these plots, we evaluated the balance of potential confounders between the comparison groups.

2.3.3. Sensitivity analysis

We conducted several sensitivity analyses to assess the robustness of these results. We used multiple logistic regression models to fit the outcome mortality and covariate vasopressors, as well as, variables selected using GLM Lasso and stepwise selection[28] based on Akaike Information Criterion (AIC)[29]. We further applied other causal inference methods, but since the results were consistent and for the sake of brevity, we did not include them here. In order to account for baseline blood pressure and heart rate as potential confounders, we performed a subgroup analysis limited to only those that had these data available (available in 37% of the subjects) and included blood pressure and heart rate in the final outcome model. To account for other complications, we also included the same 443 diagnosis present on admission as confounders in the model. Furthermore, to adjust the model by

initial GCS, we used propensity score matching based on a subgroup of patients (40% of non-traumatic cohort) with at least one record of GCS prior to the treatment.

2.4. Results

Of 2,634 non-traumatic SAH patients, 559 (21.2%), 1,342 (50.9%), and 733 (27.8%) were initially treated with dopamine, phenylephrine, and norepinephrine respectively.

Baseline demographics, characteristics, mortality, and comorbidities are summarized by the three treatment groups in Table 2.1.

Table 2.1: Baseline demographics and clinical characteristics/outcomes of non-traumatic aneurysmal SAH patients by vasopressor treatment group.

Characteristic	Total	Dopamine	Phenylephrine	Norepinephrine	pValue
<i>Baseline/demographics</i>					
No. of patients (%)	2634 (100)	559 (21.2)	1342 (50.9)	733(27.8)	<0.001
Mean age \pm SD, yrs	56.5 \pm 14.7	58.3 \pm 15.4	56.2 \pm 14.3	55.7 \pm 15.0	0.003
Female sex, %	63.9	65.1	63.6	63.4	0.79
Race, %					<0.001
White	66.0	66.9	68.0	61.8	
African American	22.7	22.2	19.7	28.5	
Other	11.3	10.9	12.4	9.7	
Mean SBP*	122.4	113.8	126.4	117.6	<0.001

Characteristic	Total	Dopamine	Phenylephrine	Norepinephrine	pValue
± SD	± 28.9	± 30.7	± 27.2	± 30.1	
Mean DBP*	65.4	64.3	67.0	63.2	0.004
± SD	± 16.3	± 17.5	± 15.2	± 17.6	
Mean HR*	84.3	84.4	80.4	92.0	<0.001
± SD	± 21.9	± 24.0	± 18.7	± 24.6	
<i>Outcomes/comorbidities</i>					
Mortality, %	36.5	50.6	24.5	47.6	<0.001
Mean LOS†	17.5	13.1	19.2	17.9	<0.001
± SD, days	± 19.8	± 15	± 15.9	± 27.7	
Pneumonia, %	22.3	16.1	22.1	29.2	<0.001
Anemia, %	20.0	17.2	18.8	24.6	0.001
Sepsis, %	10.1	7.3	7.5	17.1	<0.001
MI, %	6.5	7.5	4.4	9.5	<0.001
Acute renal failure, %	9.9	8.9	6.6	17.2	<0.001
Pulmonary edema, %	1.3	0.4	1.2	2.0	0.03
Heart failure, %	9.4	9.5	7.7	12.4	0.002
Diabetes, %	10.8	11.6	9.6	12.4	0.11
Liver disease, %	2.1	1.4	1.5	3.5	0.003
Mean CCI ± SD	1.8 ± 1.0	1.7 ± 0.9	1.7 ± 1.0	2 ± 1.2	<0.001

*LOS = hospital length of stay. All p values are calculated using an ANOVA or Fisher's exact test. *Baseline blood pressure and heart rate (the last measurement prior to the administration of the first vasopressor) were available and analyzed for the 37% available subjects. †For LOS one patient is excluded due to incomplete data; medians are 15, 8, 16, and 14, respectively (p < 0.001).*

2.4.1. Baseline demographics and vital signs

The population includes 64% females and the majority (66%) are white race. Patients administered dopamine were on average slightly older than those in the other two treatment groups (58.3, 56.2 and 55.7 years old for dopamine, phenylephrine, and norepinephrine groups, respectively, $p=0.003$). The phenylephrine group had the largest percentage of whites (68%) and the smallest percentage of African-Americans (19.7%) among the three groups ($p < 0.001$). The patients in the three treatment groups had different average baseline blood pressures and heart rates, where the average heart rates were higher in the norepinephrine group, and the average blood pressures were higher in the phenylephrine group, compared to that of other two treatment groups ($p=0.001$, 0.004).

2.4.2. Comorbidities and outcomes

Among 2,634 non-traumatic SAH patients, 59.8% were discharged alive, 36.5% died in hospital or discharged to hospice care, and 3.7% had an unknown mortality status. There was a markedly lower mortality rate in the phenylephrine group compared to dopamine and norepinephrine groups (24.5%, 50.6%, and 47.6%, respectively; $p < 0.001$, Table 2.1). The norepinephrine group had a higher burden of comorbidities including CCI, pneumonia, anemia, sepsis, myocardial infarction, acute renal failure, and heart failure.

2.4.3. Glasgow Coma Score (GCS)

Table 2.2 presents the patient percentages and mortality by treatment group and GCS categories for the 980 (40% of non-traumatic SAH) patients with at least one GCS record prior to administration of the vasopressors. Patients in the phenylephrine group had more

favorable initial GCS scores than those in dopamine and norepinephrine. For example, there were a smaller percentage of patients with GCS between 3-5 among the phenylephrine group (48.1%) compared to dopamine and norepinephrine (69.9% and 67.6%, respectively). However, when comparing the treatment groups among each subgroup of GCS category, phenylephrine consistently had lower mortality rates. This observation was confirmed in propensity-matched sensitivity analysis below.

Table 2.2: Percentage of patients according to GCS scores in each of the three treatment groups and mortality breakdown by treatment group and GCS categories

Treatment Group	GCS Score					Total
	3–5	6–8	9–11	12–14	15	
<i>Percentage of patients</i>						
Dopamine	69.9	13.3	1.8	11.5	3.5	100
Norepinephrine	67.6	14.3	4.1	10.2	3.8	100
Phenylephrine	48.1	19.7	10.8	17.5	4.0	100
<i>Mortality rate</i>						
Dopamine	73.4	40.0	50.0	15.4	25.0	60.2
Norepinephrine	69.5	26.7	15.4	6.3	50.0	54.0
Phenylephrine	40.6	11.1	4.7	1.0	4.2	22.5
Weighted averages	55.7	17.5	7.6	3.4	20.0	

All data given as percentages.

2.4.4. Vasopressor treatment effects based on causal inference analysis

Table 2.3 summarizes the results from the propensity score model, and Supplementary Figure 1 presents a graphical representation of the balance achieved through IPW. The adjusted (controlled for age, race, marital status, and gender) odds for mortality in the dopamine and norepinephrine groups were significantly higher than that for the phenylephrine group, in both unweighted and propensity weighted models. Specifically, propensity adjustment via IPW modestly attenuated the adjusted odds ratio, but mortality remain increased for both dopamine and norepinephrine compared with phenylephrine (unweighted OR =3.02, 95% CI=[2.42, 3.76]; weighted OR=2.19, 95% CI=[1.70, 2.81]) and (unweighted OR=2.63, 95% CI=[2.15, 3.22]; weighted OR=2.24, 95% CI=[1.80, 2.80]). Furthermore, there was no significant difference between the odds of mortality for dopamine and norepinephrine (OR=0.97, 95% CI=[0.75-1.27]).

Table 2.3: Association of treatment with mortality using a logistic regression model (unweighted), and propensity adjustment via IPW (weighted) for nontraumatic aneurysmal SAH subjects, adjusted for age, race, marital status, and sex

		Unweighted Model		Weighted Model	
Comparison	Deaths/Cohort (%)	OR (95% CI)	p Value	OR (95% CI)	pValue
<i>Dopamine vs phenylephrine</i>					
Dopamine	251/492 (51%)	3.02 (2.42–3.76)	<0.0001	2.19 (1.70–2.81)	<0.001
Phenylephrine	318/1253 (25%)	Ref	Ref	Ref	Ref

		Unweighted Model		Weighted Model	
Comparison	Deaths/Cohort (%)	OR (95% CI)	p Value	OR (95% CI)	pValue
<i>Norepinephrine vs phenylephrine</i>					
Norepinephrine	311/672 (46%)	2.63 (2.15–3.22)	<0.0001	2.24 (1.80–2.80)	<0.001
Phenylephrine	318/1253 (25%)	Ref	Ref	Ref	Ref
<i>Dopamine vs norepinephrine</i>					
Dopamine	251/492 (51%)	1.15 (0.91–1.45)	0.25	0.97 (0.75–1.27)	0.85
Norepinephrine	311/672 (46%)	Ref	Ref	Ref	Ref

2.4.5. Sensitivity analyses

The vital sign records are not available in the database prior to 2009, and it is the hospital's decision which tables to report to the Cerner database. Hence, important baseline measurements such as systolic blood pressure (SBP), diastolic blood pressure (DBP), and heart rate (HR) are only available for 37% of patients in our cohort. We conducted subgroup analyses using the patients who had available data on baseline DBP, SBP, HR, and potential confounders of the treatment outcome association (Supplementary Tables 2-3). Accounting for these measurements did not change the findings. Furthermore, a subgroup analysis based on 40% of non-traumatic patients with at least one GCS record and matched based on GCS, and demographic variables, confirmed the results, with OR=4.31, 95% CI=[2.53-7.32] for norepinephrine vs phenylephrine, OR=3.01, 95% CI=[1.07-8.48] for dopamine vs phenylephrine and

OR=0.98, 95% CI=[0.39-2.44] for dopamine vs norepinephrine (Supplementary Table 4). Supplementary tables 5-7 show that GCS and demographic factors were well balanced between each pair of treatments. We confirmed the associations between the initial vasopressor and mortality with several causal inference statistical models. For example, we conducted pairwise treatment analyses by creating separate propensity score models for each group of treatments (i.e., dopamine vs phenylephrine, norepinephrine vs phenylephrine, dopamine vs norepinephrine). We also examined whether the percentage of subjects that died differed by treatment group in the subgroups of subjects who had at least one of three comorbidities: myocardial infarction, acute renal failure, or sepsis (Supplementary Table 8).

Table 2.4 summarizes the treatment effect on mortality in a logistic regression with two underlying variable selection approaches: Stepwise and LASSO. Although in this study we are primarily interested in the non-traumatic SAH cohort (2634 patients), we repeated the analyses for all SAH patients (4,810 patients, including traumatic SAH patients) and for the subgroup of patients with records for DBP, SBP, and HR and for those without these vital signs. The results were consistent with our propensity score models. Sensitivity analysis showed the OR of mortality for the dopamine versus phenylephrine group ranged from [1.54-2.86], which contains our reported point estimate from the full sample. No significant difference in the odds of mortality was seen between dopamine

and norepinephrine range = [0.86-1.11]. The same results stand when we take into account other complications and diagnoses which were present at admission.

Table 2.4: Logistic regression for phenylephrine and norepinephrine versus dopamine (the reference treatment) adjusted for variables selected using LASSO and stepwise methods

Cohort	Dopamine vs Phenylephrine	Dopamine vs Norepinephrine
<i>All SAH Patients</i>		
LASSO		
All	2.00 (<0.001)	0.86 (0.16)
All w/ BP-HR	2.22 (<0.001)	0.88 (0.55)
All w/o BP-HR	1.78 (<0.001)	0.88 (0.34)
Stepwise		
All	1.98 (<0.001)	0.86 (0.14)
All w/ BP-HR	2.39 (<0.001)	0.92 (0.7)
All w/o BP-HR	1.87 (<0.001)	0.89 (0.37)
<i>Nontraumatic SAH</i>		
LASSO		
Nontraumatic	1.79 (<0.001)	0.86 (0.32)
Nontraumatic w/ BP-HR	2.47 (0.004)	1.01 (0.98)
Nontraumatic w/o BP-HR	1.54 (0.007)	0.87 (0.46)
Stepwise		
Nontraumatic	1.85 (<0.001)	0.86 (0.3)
Nontraumatic w/ BP-HR	2.84 (0.001)	1.11 (0.74)

Cohort	Dopamine vs Phenylephrine	Dopamine vs Norepinephrine
Nontraumatic w/o BP-HR	1.60 (0.003)	0.87 (0.44)

BP = blood pressure. For each model the OR (p value) is given. The analyses are also repeated for all SAH patients, for subgroups of patients with baseline blood pressure and HR, and those without baseline blood pressure and HR.

2.5. Discussion

Patients who received phenylephrine, dopamine, or norepinephrine as their first or only vasopressor had a mortality rate of 24.5%, 47.6%, and 50.6% respectively (Table 1). The treatment administered to the most patients was phenylephrine (48%). Importantly, the mortality benefit associated with phenylephrine was preserved even when the acute comorbidities of acute myocardial infarction, renal failure, and sepsis were included.

Our data indicates that phenylephrine administered for non-traumatic SAH is associated with reduced mortality over other vasopressors, which the investigators found surprising due to the sympathomimetic role of phenylephrine in causing vasoconstriction while not increasing cardiac output[30]. Some research up to this point suggests phenylephrine is not the optimal drug for aneurysmal SAH. In an evaluation by Roy et al, among 63 patients who developed DCI associated with SAH, phenylephrine was associated with worse outcomes[8]. While our population did have a substantial percentage of patients who had a change in vasopressor during the same encounter of care (33% of non-traumatic SAH patients), for the purposes of our analysis we followed the intention to treat approach by categorizing each patient to their initial vasopressor administered.

Interestingly, Joseph et al analyzed cerebral blood flow using xenon tomography and found that cerebral blood flow increased by 75% with phenylephrine, though this study did not directly compare phenylephrine to dopamine or norepinephrine[31]. Similarly, Muizelaar found increases in cerebral blood flow with phenylephrine administration using xenon tracers[32].

The study population appears to reflect previous studies assessing SAH management; in fact the fraction of patients receiving phenylephrine identically matched those reported in 2011 by Meyer et al when she performed a survey of practicing neuro-intensivists[33]. Meyer's survey had an excellent response rate of 45%. Phenylephrine has several attributes which set it apart from norepinephrine and dopamine, including the potential to be given peripherally for many hours safely with appropriate concentration and safety protocols, thereby avoiding potential delays related to hospital policies preventing peripheral administration of dopamine or norepinephrine[34]. As norepinephrine and dopamine are recognized to be relatively more potent than phenylephrine, their use may potentially lead to blood pressure increases which exceed the endpoint established by the neurocritical care team. Such a phenomenon may lead to complications stemming from excessive hypertension, such as posterior reversible encephalopathy syndrome (PRES). While PRES is noted to be rare, complications from this disease are known to occur and may potentially be unrecognized in routine clinical care[35]. Phenylephrine has not been associated with changes in ICP, while dopamine is associated with slightly higher ICPs than norepinephrine in a small clinical study of head trauma patients where patients were switched between norepinephrine and dopamine[36].

The trend to refer patients to high volume centers is well founded and rationalized by the presence of tertiary and experienced multidisciplinary personnel with a full array of interventional capabilities including neuroangiography; the threshold previously described is >60 cases per year[37]. We were not able to precisely validate such findings, though there was a significant trend favoring care in a more experienced center which supports the current health system paradigm (Supplementary Figure 2).

Limitations to our study include the retrospective nature of the Cerner Health Facts database. While our population is relatively large and representative given the incidence rate of SAH in USA[2,38,39] and includes granular data of drugs administered, functional status upon discharge (modified Ranking, etc.) was not available and such outcomes cannot be considered with this dataset. Additionally, while we used the new diagnosis of SAH and the concomitant administration of pressors to indicate severe disease, we were not able to confirm with imaging whether or not included patients had aneurysmal disease. When adding GCS to the propensity score model to account for patient severity, IPW did not return a desirable GCS balance among the vasopressor treatment groups; hence instead of IPW we used propensity scores to match the subjects and obtained satisfactory GCS and other covariate balance. Given that this analysis confirmed our primary analyses, we expect that the medication variables and demographics probably captured patient severity and factors associated with mortality. Finally, the important comorbidities that could be confounders of the treatment-mortality relationship, have no time of onset information in the database. Therefore, we cannot determine whether comorbidities existed before the administration of vasopressors, or

whether they occurred due to the vasopressors or associated complications, and could be considered as surrogate outcomes. Hence, we only adjusted for known pretreatment medications (surrogates for comorbidities) and demographics in the propensity model. Due to potential imprecision in available baseline diagnostic variables for use in the propensity models, there is the possibility for unmeasured confounding since the odd ratios become more and more attenuated as the adjustment performed becomes more stringent, but never to null. This could be a sign of unmeasured confounding, violating one of the two fundamental assumptions of propensity scoring. Being able to include more pretreatment clinical covariates in the model would further strengthen the analysis. Future work includes studying the difference conflicting potential for multi-drug therapy versus non-drug pressor therapy

2.6. Conclusion

Dopamine, phenylephrine, and norepinephrine are frequently administered in the setting of high grade SAH. Phenylephrine administration is associated with a substantial reduction in mortality among these three agents for patients admitted with a new diagnosis of aneurysmal subarachnoid hemorrhage requiring vasopressors. Dopamine usage was associated with the highest mortality. Prospective studies are warranted to further evaluate these findings.

2.7. Acknowledgments

The authors acknowledge the contributions from all members of the EHR Working Group at the CBD-HS.

2.7.1. Sources of funding

This project is supported by the Center for Big Data in Health Sciences (CBD-HS) at School of Public Health, University of Texas Health Science Center at Houston (UTHealth), and partially supported (support on data preparation) by the SBMI Data Service Office and Data Science and Informatics Core for Cancer Research (funded by CPRIT RP170668) at UTHealth. Some of the trainees are supported by the NIH training grant 2T32GM074902.

2.7.2. Disclosures

None

2.8. Supplementary Material

Supplementary 2 Table 1: National Drug Codes for Vasopressors used for the Study

National Drug Code (NDC)	Brand Name	Generic Name	Strength	Route
24134204	Neo-Syneprine	phenylephrine	10 mg/ml	injectable
74180001	Neo-Syneprine	phenylephrine	10 mg/ml	injectable
409180001	Neo-Syneprine	phenylephrine	10 mg/mL	injectable
364242646	phenylephrine Hydrochloride	phenylephrine	10 mg/ml	injectable
517029925	phenylephrine Hydrochloride	phenylephrine	10 mg/ml	injectable
517040525	phenylephrine Hydrochloride	phenylephrine	10 mg/ml	injectable
641048225	phenylephrine Hydrochloride	phenylephrine	10 mg/ml	injectable
641608825	phenylephrine Hydrochloride	phenylephrine		injectable
641614225	phenylephrine Hydrochloride	phenylephrine	10 mg/mL	injectable
703163104	phenylephrine Hydrochloride	phenylephrine	10 mg/ml	injectable
10019016301	phenylephrine Hydrochloride	phenylephrine	10 mg/ml	injectable
10019016312	phenylephrine Hydrochloride	phenylephrine	10 mg/ml	injectable
24200009150	phenylephrine Hydrochloride	phenylephrine		intravenous
24200109610	phenylephrine Hydrochloride	phenylephrine		intravenous
52533017112	phenylephrine Hydrochloride	phenylephrine	100 mcg/mL- NaCl 0.9%	intravenous
61553030765	phenylephrine Hydrochloride	phenylephrine		intravenous
61553030772	phenylephrine Hydrochloride	phenylephrine		
66647600942	phenylephrine Hydrochloride	phenylephrine		intravenous
66758001604	phenylephrine Hydrochloride	phenylephrine	10 mg/mL	injectable

National Drug Code (NDC)	Brand Name	Generic Name	Strength	Route
66758001701	phenylephrine Hydrochloride	phenylephrine	10 mg/ml	injectable
409337525	Levophed	norepinephrine	1 mg/mL	intravenous
24112302	Levophed Bitartrate	norepinephrine	1 mg/ml	intravenous
74144304	Levophed Bitartrate	norepinephrine	1 mg/ml	intravenous
247120004	Levophed Bitartrate	norepinephrine	1 mg/mL	intravenous
409144304	Levophed Bitartrate	norepinephrine	1 mg/mL	intravenous
409337504	Levophed Bitartrate	norepinephrine	1 mg/mL	intravenous
61553013461	norepinephrine	norepinephrine	4 mg/250 mL-NaCl 0.9%	injectable
74704101	norepinephrine Bitartrate	norepinephrine	1 mg/ml	intravenous
574085410	norepinephrine Bitartrate	norepinephrine	1 mg/ml	intravenous
703115303	norepinephrine Bitartrate	norepinephrine	1 mg/ml	intravenous
781893285	norepinephrine Bitartrate	norepinephrine		intravenous
24200011610	norepinephrine Bitartrate	norepinephrine		injectable
24200011810	norepinephrine Bitartrate	norepinephrine		injectable
24200111610	norepinephrine Bitartrate	norepinephrine	4 mg/250 mL-NaCl 0.9%	injectable
24200111810	norepinephrine Bitartrate	norepinephrine	8 mg/250 mL-NaCl 0.9%	injectable
36000016210	norepinephrine Bitartrate	norepinephrine		intravenous

National Drug Code (NDC)	Brand Name	Generic Name	Strength	Route
55390000210	norepinephrine Bitartrate	norepinephrine	1 mg/ml	intravenous
61553011511	norepinephrine Bitartrate	norepinephrine		
61553013411	norepinephrine Bitartrate	norepinephrine		
61553015311	norepinephrine Bitartrate	norepinephrine		
61553027211	norepinephrine Bitartrate	norepinephrine		injectable
66647615633	norepinephrine Bitartrate	norepinephrine		injectable
66647615733	norepinephrine Bitartrate	norepinephrine		injectable
223748605	dopamine	dopamine	80 mg/ml	intravenous
74426501	dopamine Hydrochloride	dopamine	80 mg/ml	intravenous
74581916	dopamine Hydrochloride	dopamine	40 mg/ml	intravenous
74582001	dopamine Hydrochloride	dopamine	40 mg/ml	intravenous
74582010	dopamine Hydrochloride	dopamine	40 mg/ml	intravenous
74780824	dopamine Hydrochloride	dopamine	5%-80 mg/100 ml	intravenous
74780922	dopamine Hydrochloride	dopamine	5%-160 mg/100 ml	intravenous
74780924	dopamine Hydrochloride	dopamine	5%-160 mg/100 ml	intravenous
74781022	dopamine Hydrochloride	dopamine	5%-320 mg/100 ml	intravenous
74910420	dopamine Hydrochloride	dopamine	40 mg/ml	intravenous
186063901	dopamine Hydrochloride	dopamine	40 mg/ml	intravenous

National Drug Code (NDC)	Brand Name	Generic Name	Strength	Route
264148255	dopamine Hydrochloride	dopamine	5%-160 mg/100 ml	intravenous
264514820	dopamine Hydrochloride	dopamine	5%-160 mg/100 ml	intravenous
338100502	dopamine Hydrochloride	dopamine	5%-80 mg/100 ml	intravenous
338100702	dopamine Hydrochloride	dopamine	5%-160 mg/100 ml	intravenous
338100703	dopamine Hydrochloride	dopamine	5%-160 mg/100 ml	intravenous
338100902	dopamine Hydrochloride	dopamine	5%-320 mg/100 ml	intravenous
409426501	dopamine Hydrochloride	dopamine	80 mg/mL	intravenous
409582001	dopamine Hydrochloride	dopamine	40 mg/mL	intravenous
409780822	dopamine Hydrochloride	dopamine	5%-80 mg/100 mL	intravenous
409780824	dopamine Hydrochloride	dopamine	5%-80 mg/100 mL	intravenous
409780922	dopamine Hydrochloride	dopamine	5%-160 mg/100 mL	intravenous
409780924	dopamine Hydrochloride	dopamine	5%-160 mg/100 mL	intravenous

National Drug Code (NDC)	Brand Name	Generic Name	Strength	Route
409781022	dopamine Hydrochloride	dopamine	5%-320 mg/100 mL	intravenous
409910420	dopamine Hydrochloride	dopamine	40 mg/mL	intravenous
517130525	dopamine Hydrochloride	dopamine	160 mg/ml	intravenous
517180525	dopamine Hydrochloride	dopamine	40 mg/ml	intravenous
517190525	dopamine Hydrochloride	dopamine	80 mg/ml	intravenous
590004006	Intropin	dopamine	40 mg/ml	intravenous

Supplementary 2 Table 2: Association of treatment with mortality using logistic regression model (unweighted) and propensity adjustment via inverse probability weighting (weighted) for non-traumatic SAH subjects in the subgroup that had baseline diastolic blood pressure (DBP), systolic blood pressure (SBP), heart rate (HR) (N=964), adjusted for age, race, marital status, gender, DBP, SBP and HR.

Comparison	Deaths/N (%)	Unweighted model		Weighted model	
		OR (95% CI)	P-value	OR (95% CI)	P-value
<i>dopamine vs phenylephrine</i>					
dopamine	48/88 (55)	3.66 (2.26-5.94)	<0.0001	3.79 (2.07-6.92)	<0.0001
phenylephrine	129/569 (23)	Ref	Ref	Ref	Ref
<i>norepinephrine vs phenylephrine</i>					
norepinephrine	142/283 (50)	2.83 (2.04-3.94)	<0.0001	2.01 (1.40-2.89)	0.0002
phenylephrine	129/569 (23)	Ref	Ref	Ref	Ref
<i>dopamine vs norepinephrine</i>					
dopamine	48/88 (55)	1.29 (0.78-2.15)	0.3229	1.88 (1.00-3.56)	0.0512
norepinephrine	142/283 (50)	Ref	Ref	Ref	Ref

Supplementary 2 Table 3: Association of treatment with mortality using logistic regression model (unweighted) and propensity adjustment via inverse probability weighting (weighted) for non-traumatic SAH subjects in the subgroup that did not have diastolic blood pressure (DBP), systolic blood pressure (SBP), heart rate (HR) (N=1670), adjusted for age, race, marital status, and gender.

		Unweighted model		Weighted model	
Comparison	Deaths/N	OR (95% CI)	P-value	OR (95% CI)	P-value
<i>dopamine vs phenylephrine</i>					
dopamine	203/404 (50)	2.63 (2.02-3.41)	<0.0001	1.83 (1.38-2.43)	<0.0001
phenylephrine	189/684 (28)	Ref	Ref	Ref	Ref
<i>norepinephrine vs phenylephrine</i>					
norepinephrine	169/389 (43)	2.12 (1.62-2.77)	<0.0001	1.91 (1.42-2.57)	<0.0001
Phenylephrine	189/684 (28)	Ref	Ref	Ref	Ref
<i>dopamine vs norepinephrine</i>					
dopamine	203/404 (50)	1.24 (0.93-1.65)	0.1363	0.96 (0.70-1.31)	0.7891
norepinephrine	169/389 (43)	Ref	Ref	Ref	Ref

Supplementary 2 Table 4: Association of treatment with mortality using logistic regression model and propensity adjustment via propensity score matching for non-traumatic SAH subjects in the subgroup that GCS (Glasgow Coma Score) records(N=980), adjusted for age, race, marital status, gender, and GCS.

Comparison	Deaths/N (%)	OR (95% CI)	p-Value
<i>Dopamine vs phenylephrine (96 pairs)</i>			
dopamine	52/96 (54)	3.01 (1.07-8.48)	0.036
phenylephrine	36/96 (38)	Ref	Ref
<i>Norepinephrine vs phenylephrine (296 pairs)</i>			
norepinephrine	151/296 (51)	4.31 (2.53-7.32)	<0.0001
phenylephrine	71/296(24)	Ref	Ref
<i>Dopamine vs norepinephrine (96 pairs)</i>			
dopamine	52/96(54)	0.98 (0.39-2.44)	0.96
norepinephrine	49/96 (51)	Ref	Ref

Supplementary 2 Table 5: Distribution comparison of demographic variables and Glasgow Coma Score (GCS) after propensity matching between phenylephrine and dopamine treatment groups.

Characteristics	phenylephrine	Dopamine	p-value
n	96	96	
Age (years), mean (SD)	57.88 (15.75)	59.06 (15.17)	0.595
Race (%)			0.557
African-American	20 (20.8)	22 (22.9)	
Other	8 (8.3)	12 (12.5)	
White	68 (70.8)	62 (64.6)	
Male, n (%)	34 (35.4)	33 (34.4)	1
Marital Status, n (%)			0.938
Divorced	8 (8.3)	9 (9.4)	
Married	50 (52.1)	49 (51.0)	
Single	29 (30.2)	27 (28.1)	
Unknown	2 (2.1)	4 (4.2)	
Widowed	7 (7.3)	7 (7.3)	
GCS score, mean (SD)	5.24 (3.79)	5.51 (4.12)	0.636

Supplementary 2 Table 6: Distribution comparison of demographic variables and Glasgow Coma Score (GCS) after propensity matching between phenylephrine and norepinephrine treatment groups.

Characteristics	phenylephrine	norepinephrine	p-value
n	296	296	
Age (years), mean (SD)	56.40 (15.25)	56.14 (15.68)	0.836
Race (%)			0.531
African-American	93 (31.4)	97 (32.8)	
Other	26 (8.8)	33 (11.1)	
White	177 (59.8)	166 (56.1)	
Male, n (%)	105 (35.5)	112 (37.8)	0.609
Marital Status, n (%)			1
Divorced	29 (9.8)	30 (10.1)	
Married	120 (40.5)	120 (40.5)	
Single	109 (36.8)	107 (36.1)	
Unknown	15 (5.1)	15 (5.1)	
Widowed	23 (7.8)	24 (8.1)	
GCS score, mean (SD)	5.64 (3.82)	5.46 (3.90)	0.558

Supplementary 2 Table 7: Distribution comparison of demographic variables and Glasgow Coma Score (GCS) after propensity matching between norepinephrine and dopamine treatment groups.

Characteristics	norepinephrine	dopamine	p-value
n	96	96	
Age (years), mean (SD)	57.95 (16.01)	59.06 (15.17)	0.621
Race (%)			0.968
African-American	23 (24.0)	22 (22.9)	
Other	11 (11.5)	12 (12.5)	
White	62 (64.6)	62 (64.6)	
Male, n (%)	27 (28.1)	33 (34.4)	0.436
Marital Status, n (%)			0.967
Divorced	9 (9.4)	9 (9.4)	
Married	50 (52.1)	49 (51.0)	
Single	25 (26.0)	27 (28.1)	
Unknown	6 (6.2)	4 (4.2)	
Widowed	6 (6.2)	7 (7.3)	
GCS score, mean (SD)	5.64 (4.17)	5.51 (4.12)	0.835

Supplementary 2 Table 8: Mortality by Treatment Class separated by the cumulative comorbidity of at least one of the three diseases MI, ARF, or Sepsis.

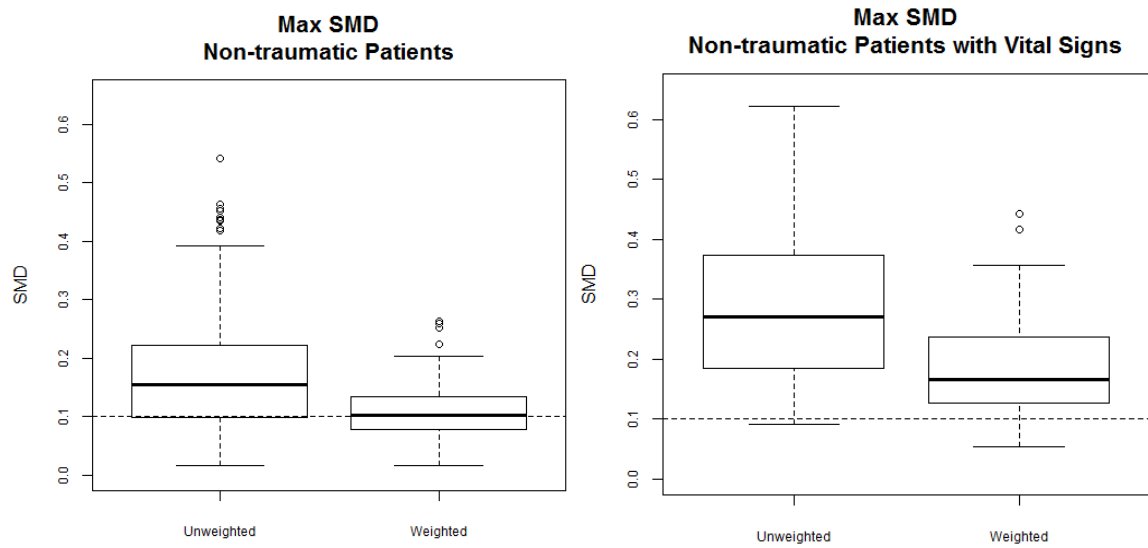
Diagnosis	dopamine	phenylephrine	norepinephrine
MI, ARF, or Sepsis	50.8	21.5	42.7
None of the three	50.0	41.4	58.5

Supplementary 2 Table 9: Mortality based on vasopressor choice for patients with only one of the three vasopressor prescription

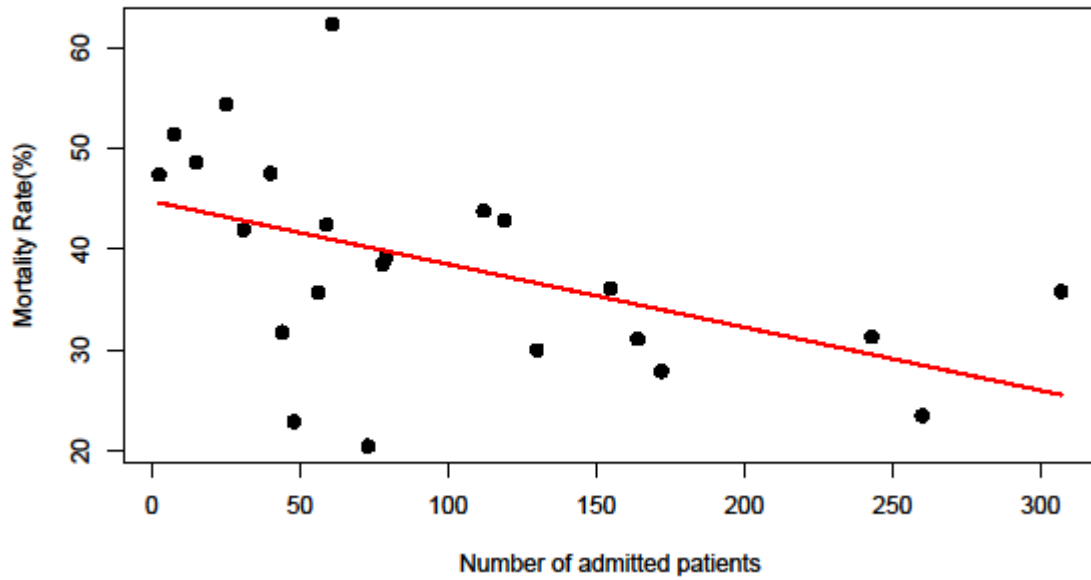
Vasopressor	percentage	Mortality
Dopamine	19.1	54.5
Phenylephrine	52.1	20.0
Norepinephrine	28.8	46.9

Supplementary 2 Table 10: Shows percentage of patients with one, two or three different vasopressors within each of the three treatment groups.

Vasopressor-Class	Number of Vasopressors		
	one	two	Three
Dopamine	60.1	33.3	3.7
Phenylephrine	68.3	25.8	6.0
Norepinephrine	69.2	27.8	3.0



Supplementary 2 Figure 1: Graphical representation of covariate balance as measured by the standardized mean difference (SMD) within the unweighted and weighted data. In both plots, the balance achieved through inverse probability weighting is better than the original unweighted data.



Supplementary 2 Figure 2: Mortality rate vs number of admitted SAH patients by hospital showing a negative association (correlation = -0.49, p-value=0.017). Hospitals with fewer than 30 subjects were pooled to estimate the mortality rates into the following bins with respect to the number of subjects: [1,5), [5,10), [10,20), and [20,30). Hospitals with 30 subjects or more were plotted individually.

2.9. References

- 1 Yao Z, Hu X, Ma L, *et al.* Timing of surgery for aneurysmal subarachnoid hemorrhage: A systematic review and meta-analysis. *Int. J. Surg.* 2017;**48**:266–74. doi:10.1016/j.ijssu.2017.11.033
- 2 Shea AM, Reed SD, Curtis LH, *et al.* Characteristics of nontraumatic subarachnoid hemorrhage in the United States in 2003. *Neurosurgery* 2007;**61**:1131–7. doi:10.1227/01.neu.0000306090.30517.ae
- 3 Mayberg MR, Batjer HH, Dacey R, *et al.* Guidelines for the management of aneurysmal subarachnoid hemorrhage. A statement for healthcare professionals from a special writing group of the Stroke Council, American Heart Association. *Stroke* 1994;**25**:2315–28. doi:10.1161/01.str.25.11.2315
- 4 Connolly ES, Rabinstein AA, Carhuapoma JR, *et al.* Guidelines for the management of aneurysmal subarachnoid hemorrhage: A guideline for healthcare professionals from the american heart association/american stroke association. *Stroke*. 2012;**43**:1711–37. doi:10.1161/STR.0b013e3182587839
- 5 Fernando SM, Perry JJ. Subarachnoid hemorrhage. *CMAJ* 2017;**189**:E1421. doi:10.1503/cmaj.170893
- 6 Budohoski KP, Guilfoyle M, Helmy A, *et al.* The pathophysiology and treatment of delayed cerebral ischaemia following subarachnoid haemorrhage. *J. Neurol. Neurosurg. Psychiatry*. 2014;**85**:1343–53. doi:10.1136/jnnp-2014-307711
- 7 Worthington JM, Goumas C, Jalaludin B, *et al.* Decreasing risk of fatal

- subarachnoid hemorrhage and other epidemiological trends in the era of coiling
Implementation in Australia. *Front Neurol* 2017;**8**. doi:10.3389/fneur.2017.00424
- 8 Roy B, McCullough LD, Dhar R, *et al*. Comparison of Initial Vasopressors Used for Delayed Cerebral Ischemia after Aneurysmal Subarachnoid Hemorrhage. *Cerebrovasc Dis* 2017;**43**:266–71. doi:10.1159/000458536
 - 9 Lin CL, Dumont AS, Zhang JH, *et al*. Cerebral Vasospasm after Aneurysmal Subarachnoid Hemorrhage: Mechanism and Therapies. *Biomed Res. Int.* 2014;**2014**. doi:10.1155/2014/679014
 - 10 Francoeur CL, Mayer SA. Management of delayed cerebral ischemia after subarachnoid hemorrhage. *Crit. Care.* 2016;**20**. doi:10.1186/s13054-016-1447-6
 - 11 Kosnik EJ, Hunt WE. Postoperative hypertension in the management of patients with intracranial arterial aneurysms. *J Neurosurg* 1976;**45**:148–54. doi:10.3171/jns.1976.45.2.0148
 - 12 Diringer MN, Bleck TP, Hemphill JC, *et al*. Critical care management of patients following aneurysmal subarachnoid hemorrhage: Recommendations from the neurocritical care society’s multidisciplinary consensus conference. *Neurocrit. Care.* 2011;**15**:211–40. doi:10.1007/s12028-011-9605-9
 - 13 Reynolds MR, Buckley RT, Indrakanti SS, *et al*. The safety of vasopressor-induced hypertension in subarachnoid hemorrhage patients with coexisting unruptured, unprotected intracranial aneurysms. *J Neurosurg* 2015;**123**:862–71. doi:10.3171/2014.12.JNS141201

- 14 Sakr Y, Dünisch P, Santos C, *et al.* Poor outcome is associated with less negative fluid balance in patients with aneurysmal subarachnoid hemorrhage treated with prophylactic vasopressor-induced hypertension. *Ann Intensive Care* 2016;**6**:25. doi:10.1186/s13613-016-0128-6
- 15 Rose JC, Mayer SA. Optimizing blood pressure in neurological emergencies. *Neurocrit. Care.* 2004;**1**:287–99. doi:10.1385/NCC:1:3:287
- 16 Muehlschlegel S, Dunser MW, Gabrielli A, *et al.* Arginine vasopressin as a supplementary vasopressor in refractory hypertensive, hypervolemic, hemodilutional therapy in subarachnoid hemorrhage. *Neurocrit Care* 2007;**6**:3–10. doi:10.1385/NCC:6:1:3
- 17 Crispo JAG, Willis AW, Thibault DP, *et al.* Associations between anticholinergic burden and adverse health outcomes in Parkinson disease. *PLoS One* 2016;**11**. doi:10.1371/journal.pone.0150621
- 18 Rasmy L, Wu Y, Wang N, *et al.* A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. *J Biomed Inform* 2018;**84**:11–6. doi:10.1016/j.jbi.2018.06.011
- 19 Oshiro EM, Walter KA, Piantadosi S, *et al.* A new subarachnoid hemorrhage grading system based on the Glasgow Coma Scale: A comparison with the Hunt and Hess and World Federation of Neurological Surgeons Scales in a clinical series. *Neurosurgery.* 1997;**41**:140–8. doi:10.1097/00006123-199707000-00029

- 20 ROSENBAUM PR, RUBIN DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70**:41–55.
doi:10.1093/biomet/70.1.41
- 21 Greene TJ, DeSantis SM, Swartz MD. A Novel Non-Parametric Method for Ordinal Propensity Score Stratification and Matching. In: *A Novel Non-Parametric Method for Ordinal Propensity Score Stratification and Matching*. 2017.
- 22 Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med* 2015;**34**:3661–79.
doi:10.1002/sim.6607
- 23 Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Med Decis Mak* 2009;**29**:661–77. doi:10.1177/0272989X09341755
- 24 Mccaffrey DF, Griffin BA, Almirall D, *et al*. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med* 2013;**32**:3388–414. doi:10.1002/sim.5753
- 25 Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. *J R Stat Soc Ser B Stat Methodol* 2007;**69**:659–77. doi:10.1111/j.1467-9868.2007.00607.x
- 26 Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika* 2000;**87**:706–10. doi:10.1093/biomet/87.3.706

- 27 Feng P, Zhou XH, Zou QM, *et al.* Generalized propensity score for estimating the average treatment effect of multiple treatments. *Stat Med* 2012;**31**:681–97. doi:10.1002/sim.4168
- 28 Menard S. *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE Publications, Inc. 2014. doi:10.4135/9781483348964
- 29 Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr* 1974;**19**:716–23. doi:10.1109/TAC.1974.1100705
- 30 Miller JA, Dacey RG, Diringer MN. Safety of hypertensive hypervolemic therapy with phenylephrine in the treatment of delayed ischemic deficits after subarachnoid hemorrhage. *Stroke* 1995;**26**:2260–6. doi:10.1161/01.str.26.12.2260
- 31 Joseph M, Ziadi S, Nates J, *et al.* Increases in cardiac output can reverse flow deficits from vasospasm independent of blood pressure: a study using xenon computed tomographic measurement of cerebral blood flow. *Neurosurgery* 2003;**53**:1044–51; discussion 1051-2. doi:10.1227/01.neu.0000088567.59324.78
- 32 Muizelaar JP, Becker DP. Induced hypertension for the treatment of cerebral ischemia after subarachnoid hemorrhage. Direct effect on cerebral blood flow. *Surg Neurol* 1986;**25**:317–25. doi:10.1016/0090-3019(86)90205-3
- 33 Meyer R, Deem S, David Yanez N, *et al.* Current practices of triple-H prophylaxis and therapy in patients with subarachnoid hemorrhage. *Neurocrit Care* 2011;**14**:24–36. doi:10.1007/s12028-010-9437-z

- 34 Delgado T, Wolfe B, Davis G, *et al.* Safety of peripheral administration of phenylephrine in a neurologic intensive care unit: A pilot study. *J Crit Care* 2016;**34**:107–10. doi:10.1016/j.jcrc.2016.04.004
- 35 Madaelil TP, Dhar R. Posterior reversible encephalopathy syndrome with thalamic involvement during vasopressor treatment of vertebrobasilar vasospasm after subarachnoid hemorrhage. *J Neurointerv Surg* 2016;**8**:e45. doi:10.1136/neurintsurg-2015-012103.rep
- 36 Ract C, Vigué B. Comparison of the cerebral effects of dopamine and norepinephrine in severely head-injured patients. *Intensive Care Med* 2001;**27**:101–6. doi:10.1007/s001340000754
- 37 Vespa P, Diringier MN. High-volume centers. *Neurocrit. Care.* 2011;**15**:369–72. doi:10.1007/s12028-011-9602-z
- 38 Zacharia BE, Hickman ZL, Grobelny BT, *et al.* Epidemiology of Aneurysmal Subarachnoid Hemorrhage. *Neurosurg. Clin. N. Am.* 2010;**21**:221–33. doi:10.1016/j.nec.2009.10.002
- 39 Etminan N, Chang HS, Hackenberg K, *et al.* Worldwide Incidence of Aneurysmal Subarachnoid Hemorrhage According to Region, Time Period, Blood Pressure, and Smoking Prevalence in the Population: A Systematic Review and Meta-analysis. *JAMA Neurol* 2019;**76**:588–97. doi:10.1001/jamaneurol.2019.0006

Chapter 3: Representation of EHR data for predictive modeling: a comparison between UMLS and other terminologies

This chapter is adapted from:

Rasmy L, Tiriyaki F, Zhou Y, Xiang Y, Tao C, Xu H, Zhi D. Representation of EHR data for predictive modeling: a comparison between UMLS and other terminologies. *Journal of the American Medical Informatics Association*. 2020 Oct 1;27(10):1593-1599. doi: 10.1093/jamia/ocaa180. PMID: 32930711; PMCID: PMC7647355.

Title: Representation of EHR Data for Predictive Modeling: A Comparison between UMLS and Other Terminologies

Authors: Laila Rasmy, MSc¹, Firat Tiriyaki¹, Yujia Zhou, MSc¹, Yang Xiang, PhD¹, Cui Tao, PhD¹, Hua Xu, PhD^{1*}, Degui Zhi, PhD^{1*}

¹ School of Biomedical Informatics University of Texas Health Science Center, Houston, Texas, USA

*Co-Senior Authors

Corresponding Author:

Degui Zhi, Ph.D

UTHealth School of Biomedical Informatics

Keywords: UMLS, Terminology representation, Predictive modeling, Electronic health records

3.1. Abstract

Objective: Predictive disease modeling using electronic health record (EHR) data is a growing field. Although clinical data in their raw form can be used directly for predictive modeling, it is a common practice to map data to standard terminologies to facilitate data aggregation and reuse. There is, however, a lack of systematic investigation of how different representations could affect the performance of predictive models, especially in the context of machine learning and deep learning.

Methods: We projected the input diagnoses data in the Cerner HealthFacts® database to UMLS and five other terminologies, including CCS, CCSR, ICD-9, ICD-10, and PheWAS, and evaluated the prediction performances of these terminologies on two different tasks: the risk prediction of heart failure in diabetes patients (DHF) and the risk prediction of pancreatic cancer (PC). Two popular models were evaluated: logistic regression (LR) and a recurrent neural network (RNN).

Results: For LR, using UMLS delivered the optimal AUROC results in both DHF (81.15%) and PC (80.53%) tasks. For RNN, UMLS worked best for PC prediction (AUROC 82.24%), second only (AUROC 85.55%) to PheWAS (AUROC 85.87%) for DHF prediction.

Discussion/Conclusion: In our experiments, terminologies with larger vocabularies and finer-grained representations were associated with better prediction performances. In particular, UMLS is consistently one of the best-performing ones. We believe that our work may help to inform better designs of predictive models, although further investigation is warranted.

3.2. Background

In the current big data era of biomedical informatics, abundant electronic health record (EHR) data are becoming available, leading to the development of predictive modeling algorithms. In the past five years, thousands of predictive modeling-related studies have utilized a variety of methods, such as logistic regression (LR) or deep learning, to predict the patient's risk of developing such diseases as heart failure[1–5] and pancreatic cancer.[6,7] An important, but unaddressed, research question in regard to predictive modeling is how to efficiently feed the EHR data to models[8–10].

Structured diagnosis data in EHR datasets are usually heterogeneous, leading to challenges in data analysis, including interpretability and generalizability issues. For example, different hospitals and departments use different terminologies; thus, to develop a generalizable model, researchers either train the model on all of the different terminologies in use or introduce a standardizer that can normalize the data into a single terminology.

Terminology standards are evolving constantly, and newer versions will introduce additional levels of data redundancy. For example, patient diagnosis information was commonly stored in the International Classification of Diseases-ninth revision (ICD-9) format before 2015, but then, for billing purposes, hospitals had to upgrade it to the tenth revision (ICD-10), which introduced a higher level of details. Currently, an even newer revision, ICD-11, is being released. Further, in many cases, the coding system in EHRs is a mix of multiple ICD terminologies. As a result, it is difficult to organize information represented in heterogeneous formats and for models trained on older terminologies (e.g., ICD-9 codes) to generalize to new terminologies without proper normalizations.

EHR vendors also are introducing internal codes that can be mapped to different standard terminologies in a one-to-one manner to facilitate various system functionalities. Using such codes for predictive model training may restrict the generalizability of such models to vendor-specific solutions or even to a single hospital if the mappings are different between sites. In addition, many of the existing terminology mappings are in many-to-many styles, which might hinder the accuracy and the interpretability of the model.

Terminology normalization involves assigning a unique standard medical term to a health condition.[11] Most terminology mapping and normalization-related studies concern the development of mappings between different terminologies,[12–14] the tools developed for automated mapping suggestions, or the development of concept embeddings based on different terminologies.[15–19] There are, however, several practical questions on terminology normalization that have not been addressed. The first is how to find the optimal level of granularity required for predictive modeling, assuming that the data source is homogeneous. For example, it is not known whether we should use the diagnosis information as originally recorded in the dataset or group similar or relevant codes to reduce the input dimension.

The second is how important terminology normalization is when the data source is heterogeneous. Rajkomar et al.[10] described the advantage of using the Fast Healthcare Interoperability Resources (FHIR) format for interchangeable information representation but acknowledged that the limited semantic consistency from unharmonized data may have a negative impact on the model performance. In our previous work,[4] we compared the use of Clinical Classifications Software (CCS) codes with the raw data from Cerner

HealthFacts® and found that grouping diagnosis codes was not helpful, which conflicts with the findings of other studies[2,9] that the CCS grouping was helpful. Notably, our findings also were supported by those of other studies.[18,20] Unified Medical Language System (UMLS) provides a multipurpose knowledge source and attracts more research attention, as it includes mappings to almost all clinical terminologies at different hierarchical levels.[21] It also has been broadly used in concept normalizations in the natural language processing NLP domain;[17,22,23] thus, we selected it as our most expressive terminology.

3.3. Objective

In this study, our objectives are twofold. The first objective is to compare simply feeding the models with the raw data, as they were originally collected, versus preprocessing the data when mapping it to a single terminology. The second is to evaluate the performance of predictive models using UMLS and five other terminologies commonly used in the healthcare analytics domain. We used two clinical prediction tasks: predicting the risk to develop heart failure among a cohort of type-II diabetes mellitus (DMII) patients and the risk to develop pancreatic cancer. Our study cohorts were extracted from the Cerner HealthFacts® database, a de-identified EHR database extracted from over 600 hospitals with which Cerner has a data use agreement. The original diagnosis data are coded with a unique diagnosis identifier (Cerner-Diagnosis ID) that can be mapped to ICD-9, ICD-10-CM, or ICD-10-CA codes in a one-to-one manner. For comparison, we further mapped the diagnoses codes to six terminologies, including UMLS concept unique identifier (CUI)[24], ICD-9[25], ICD-10[25], PheWAS[26], and CCS codes in both the single-

level[27,28] and its refined version (CCSR).[29] We compared the performances using L2 penalized LR ($L2LR$) and a bidirectional recurrent neural network (RNN)-based predictive model.

3.4. Methods

3.4.1. Prediction tasks and cohort description

We evaluated the use of patients' diagnosis information in different terminology representations on two different prediction tasks. The first task is to predict the development of heart failure in patients with DMII after at least one month of their first DMII diagnosis. The second task is to predict whether the patient will be diagnosed with pancreatic cancer in the next visit. The second task is more like a diagnosis aid, as we did not specify a prediction window.

We extracted our cohorts from the Cerner HealthFacts® dataset v.2017,[30] which includes de-identified patient information from more than 600 hospitals for more than a 15-year period. The full cohort for the heart failure prediction in DMII patients consists of 70,782 cases and 1,095,412 controls denoted as the DHF full cohort, out of which we randomly selected a sample of 60,000 cases and 60,000 controls for terminology evaluations further denoted as the DHF cohort. Table 3.5 shows the descriptive analysis of the selected sample versus the full cohort. For pancreatic cancer prediction, we found 11,486 eligible cases in the population who were 45 years or older and did not report any other cancer diseases before their first pancreatic cancer diagnosis. From a pool of more than 25 million matched controls, we randomly selected 17,919 controls to build our pancreatic cancer experimental cohort, which was denoted as the PC cohort. We further

randomly split each sample cohort into training, validation, and test sets using the ratio of 7:1:2.

Table 3.5. Description of cohorts.

Characteristic	DHF full cohort		DHF cohort (Study sample)		PC cohort (Study sample)	
	Case	Controls	Case	Controls	Case	Controls
Cohort size (<i>n</i>)	70,782	1,095,412	60,000	60,000	11,486	17,919
Male %	49%	47%	49%	46%	47%	43%
Age (mean (std. dev.))	70 (12)	60 (14)	70 (12)	60 (14)	69 (19)	63 (13)
Race						
White (%)	76%	70%	77%	71%	80%	75%
African American (%)	17%	16%	16%	16%	14%	12%
Average number of visits	13	16	14	15	7	7
Average number of codes	28	32	30	31	23	21

We used the patients' diagnosis information only before the index visit, which is commonly the last eligible visit before prediction, to train the predictive models. Details of the cohorts' composition are presented in Appendix A.

3.4.2. Diagnosis terminology

Cerner HealthFacts® v. 2017 includes 17,629 ICD-9 codes, 94,044 ICD-10-CM codes, and 16,044 ICD-10-CA codes, each of which is mapped to a unique Cerner-Diagnosis ID that is used to unify the representation of diagnosis among all hospitals' diagnoses data within the Cerner HealthFacts® database. The patient's diagnosis information is stored mainly through the use of Cerner-Diagnosis ID. The main advantage of using this raw data is that they include the information of the original code types used for documentation and can be

directly used without any further processing. This dilutes the actual value of the patient diagnosis, however, as the same diagnosis may be represented by multiple codes. We also included the raw data using the Cerner-Diagnosis ID as a baseline terminology. Figure 3.1 shows our diagnosis terminology mapping roadmap.

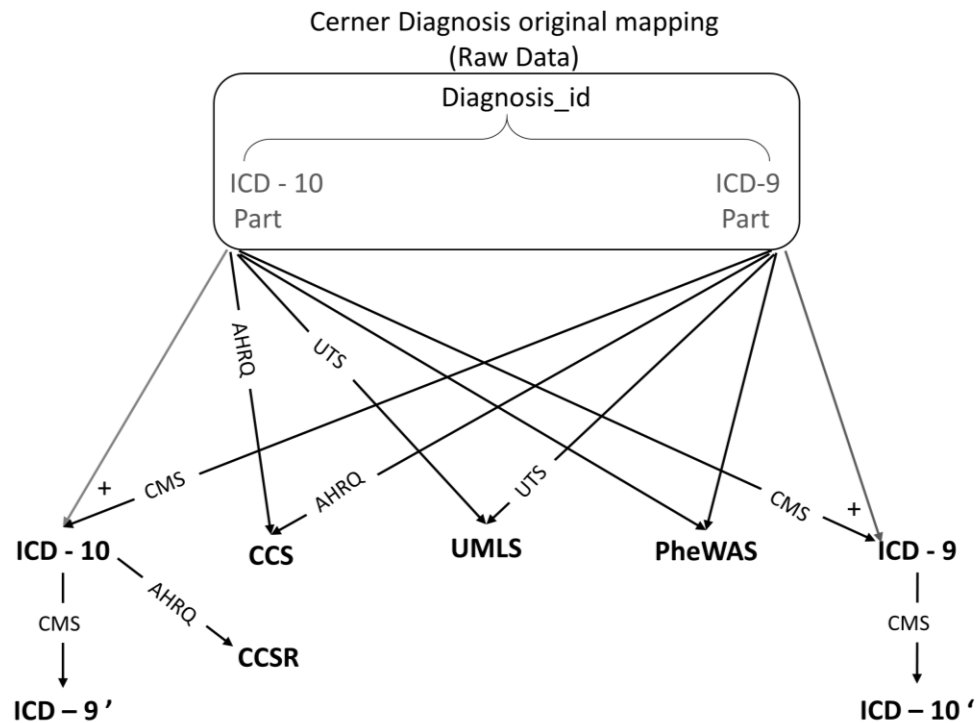


Figure 3.2. Terminology conversion roadmap

We used official resources for code mappings. For example, we used the Center of Medicare and Medicaid Services' (CMS) most recent general equivalence mapping (GEM) version 2018[25] to map between ICD-9 and ICD-10 codes. For UMLS mapping, we used the UMLS knowledge resources available on the UMLS Terminology Services (UTS)

website,[24] and we used the latest version of ICD-9 and ICD-10 to CCS single-level mapping available on the Agency of Healthcare Research and Quality (AHRQ) website[27,28] as well as the CCS Refined version.[29] For PheWAS mapping, we used the Phecode maps available in the PheWAS Catalog.[26] We had to review the mappings to the raw data for some typos in the Cerner diagnosis dictionary table that led to ICD-9/10 codes as not exactly matching the corresponding records in different mapping files, due mainly to their missing the last digit, which can be either 0 or 9.

All of the Cerner-Diagnosis IDs in our cohort were successfully mapped to CCS and ICD-9 codes, regardless of those that were mapped to ‘noDx’ for no mapping as existing in the original mapping files. There were approximately 300 ICD-10 codes in our cohort that were not mapped that were associated with approximately 100 UMLS codes. We decided to ignore those codes, as they appeared only a maximum of 10 times in our cohort. Details of the different terminologies used and the mapping are provided in Appendix B of the supplementary materials.

To understand whether the difference in the predictive model accuracy is due to the terminology representation itself or to the information loss induced by the mapping process, we focused on the ICD-9/ICD-10 conversion as an example. We converted the previously converted ICD-9 codes to ICD-10 and named them ICD-10 revert prime (ICD-10’). We did the same for the previously converted ICD-10 codes and converted them back to ICD-9 and named them ICD-9 revert prime (ICD-9’). For the revert prime mappings, we used only the original mapping files provided by CMS without any further review or

improvement. For reproducibility, we share our codebase and mappings on https://github.com/ZhiGroup/terminology_representation.

3.4.3. Tasks and models

We evaluated the usefulness of the terminologies described above for two tasks. The first task is the prediction of the DMII patients' risk to develop heart failure after 30 days from their first diabetes diagnosis. The second task is the calculation of the risk score of the patient to be diagnosed with pancreatic cancer based on the patient's history until the most recent visit.

For both tasks, we evaluated two models: *L2LR* and RNN. LR is a popular model for its accuracy and interpretability. The majority of currently implemented predictive models are based on LR. We used one-hot encoding for the presence/absence of any diagnosis code as input for LR. We used the default LR implementation available in the Scikit-Learn[31] package which includes L2 penalty for regularization. We also experimented hyperparameter grid search for the L2 penalty. In addition, we evaluated a bidirectional RNN. RNNs are appropriate for modeling the sequential nature of patient medical records and have been shown to provide high predictive accuracy in the healthcare domain.[1,2,4,32] Following Choi et al.,[2–4,33] we represented a patient record as a sequence of visits (encounters) and each encounter as a set of diagnosis codes. We used an embedding layer to transform one-hot input diagnosis vectors into dense vectors and then used a bidirectional gated recurrent unit (GRU) for propagating information across visits and a fully connected layer for the output label. Hyperparameters were chosen by Bayesian optimization. This architecture was shown to be very competitive in our previous

benchmark.[34] We used our previously published code on https://github.com/ZhiGroup/pytorch_ehr. A detailed description of our model implementation is available in Appendix C.

3.4.4. Statistical analysis for model comparison

We used the area under the receiver operating characteristic curve (AUROC) as the evaluation metric for the model prediction accuracy. For deterministic methods, such as *L2LR*, we apply the Delong test[35] to calculate the significance of the difference between different models' AUROC. For probabilistic methods, such as RNN (due to random initialization of model parameters), we repeated the analyses for RNN models of each terminology 10 times, and multi-group one-way ANOVA tests (unpaired *t*-tests for two groups) were used for comparing the means of each terminology. All-pairwise Tukey-Kramer analysis was used to identify significant group-wise differences.

3.5. Results

As noted, the description of both cohorts is presented in Table 3.1. Also as noted, we lost some patient information for the incomplete terminology mapping, mainly for the primary ICD-10 to 9 code mappings and the reversed prime conversions. Nevertheless, that rarely leads to loss of a complete patient sample for the initial rules of the minimum number of visits, and original diagnosis codes were redefined before the random sample selection in the DHF cohort. Thus, our test set of 24,000 patients remains consistent along with the evaluations of all of our models. For the PC cohort, only a couple of patients from our test

set of 5,881 patients were not included in the ICD-9 mappings; those patients were excluded from the reported results.

As shown in Table 3.6, for the DHF prediction, the test AUROC ranges between 78% and 81% for *L2LR* and between 83% and 85% for RNN. For the PC prediction, the test AUROC ranges between 77% and 80.5% for *L2LR* and between 79% and 82.5% for RNN. The difference between the RNN and *L2LR* AUROCs remains nearly the same among all diagnosis terminologies, approximately 4.9% on average for DHF and 1.5% for PC. The best *L2LR* models' AUROC is associated with the use of UMLS-CUI on both tasks, whereas single-level CCS shows the worst AUROC in all tasks and models. The findings remain consistent even with the DHF full cohort (Appendix D), for which UMLS showed the highest AUROC (82%). Also, our results remained consistent using LR with different regularization hyperparameters for both L1 and L2 regularizations (Appendix E).

Using the Delong test to understand the difference in the AUROC significance and with a *p*-value of 0.0024 after Bonferroni correction (Figure 3.2(A)), we find that the UMLS results are significantly better than those for the other terminologies except for the raw data in PC prediction and PheWAS in DHF prediction. For RNN models, UMLS showed the highest AUROC for PC prediction, whereas PheWAS was associated with the best AUROC for DHF prediction. These pairwise comparison results are statistically significant based on the Tukey-Kramer procedure, as shown in Figure 3.2(B) and Appendix F. We train and test the RNN models only once on the DHF full cohort (Appendix D). UMLS was the second-best performer, with an AUROC of 85.52%, which is 0.34% less than that

of the raw data, which showed the highest AUROC, at 85.86%. The AUROC of PheWAS was lower, at 85.07%.

Table 3.6. Prediction performance of different diagnosis terminologies for the DHF and PC tasks

Diagnosis terminology	Diabetes heart failure cohort (DHF)			Pancreatic cancer cohort (PC)		
	Number of unique codes	<i>L2LR</i>	RNN	Number of unique codes	<i>L2LR</i>	RNN
Raw data (Cerner-Diagnosis ID)	26,427	80.61	85.48 (0.10)	13,071	80.30	81.43 (0.37)
CCS-single level	284	78.07	82.96 (0.15)	253	77.23	79.03 (0.36)
CCSR	538	78.87	84.17 (0.21)	538	77.92	79.63 (0.34)
ICD-9	11,187	80.12	85.20 (0.13)	7,055	79.15	80.78 (0.32)
ICD-10	22,893	79.78	84.35 (0.20)	13,620	78.95	79.27 (0.44)
PheWAS	1,820	80.71	85.87 (0.10)	1,715	78.82	81.15 (0.31)
UMLS CUI	29,491	81.15	85.55 (0.06)	14,551	80.53	82.24 (0.29)

Note. *L2LR* and RNN show the average and the standard deviation for AUROC on the test set. Bold indicates the values with the highest AUROC per task/model.

A)	PC cohort		Raw	CCSR	ICD-10	ICD-9	PheWAS	UMLS	
	CCS		2x10 ⁻⁷	0.0884	0.003	8x10 ⁻⁴	0.0024	2x10 ⁻⁸	
	DHF cohort	Raw		3x10 ⁻⁵	1x10 ⁻⁵	6x10 ⁻⁵	0.0008	0.285	
		CCS		CCSR	0.0605	0.025	0.0601	2x10 ⁻⁶	
		Raw		9x10 ⁻²³	Raw	ICD-10	0.503	0.8361	3x10 ⁻⁶
		CCSR		1x10 ⁻⁸	9x10 ⁻¹³	CCSR	ICD-9	0.4895	2x10 ⁻⁵
		ICD-10		6x10 ⁻¹³	7x10 ⁻⁸	4x10 ⁻⁵	ICD-10	PheWAS	0.0001
		ICD-9		2x10 ⁻¹⁷	0.001	2x10 ⁻⁸	0.0154	ICD-9	
		PheWAS		2x10 ⁻³⁹	0.524	1x10 ⁻²³	6x10 ⁻⁸	5x10 ⁻⁴	PheWAS
		UMLS		4x10 ⁻³⁵	1x10 ⁻⁸	4x10 ⁻²²	4x10 ⁻¹⁸	4x10 ⁻¹¹	0.03432

B)	DHF cohort				PC cohort			
	Level		Mean		Level		Mean	
	PheWAS	A		85.87%	UMLS	A		82.24%
	UMLS	B		85.55%	Raw	B		81.43%
	Raw	B		85.48%	PheWAS	B C		81.15%
	ICD-9	C		85.20%	ICD-9	C		80.78%
	ICD-10		D	84.35%	CCSR		D	79.63%
	CCSR		D	84.17%	ICD-10		D E	79.27%
	CCS		E	82.96%	CCS		E	79.03%

Figure 3.3. Significance of AUROC difference. (A) Logistic regression pairwise AUROC difference significance calculated using Delong test; P values less than .0024 are significantly different. (B) For the Tukey-Kramer honest significant difference test value, levels not connected by the same letter are significantly different.

The mapping to ICD-9 is always better than mapping to ICD-10, although those differences were not significant for *L2LR* models, based on the Delong test, but were significant for RNN models. We hypothesize that the result is due to the majority of the original data's being recorded in ICD-9, and, thus, mapping to ICD-10 will incur a loss of information during the terminological translation. We further investigated this loss-in-translation effect and report the results in the next section.

3.5.1. Effect of information loss due to terminology mapping

Mapping back from earlier converted ICD-10 codes to ICD-9 was associated with clear information loss that can be seen in the difference in the number of codes in our cohort; for example, our cohort originally had a 26,427-diagnosis code that mapped to a combination

of ICD-9 and ICD-10 codes (Table 3.6). Approximately 70% of our patient diagnosis data already were coded in ICD-9 codes, so, after mapping the ICD-10 codes to ICD-9 and combining the codes with those data originally mapped to ICD-9 codes, we had 11,187 ICD-9 codes in our cohort. Thus, we can explain the decrease in the number of codes as a result of the grouping effect of the lower dimension ICD-9 codes, but, on mapping back to ICD-10 codes, the number of codes increases only to 14,644 codes, which is approximately 50% of the number of original diagnosis codes, or a little higher percentage of the primarily converted ICD-10 codes (22,893 codes). Such information loss may explain the significant decrease in AUROC, using the ICD-9' and ICD-10' sets (Table 3.7).

Table 3.7. Difference in AUROC between primary mapping to ICD-9/10 codes and reversed mapping to ICD-9'/10' codes

	Number of Codes	<i>L2LR</i> AUROC	Delong <i>p</i> -value	RNN AUROC	Unpaired <i>t</i> -test <i>p</i> -value
ICD-9	11,187	80.12	$p < 0.0001$	85.20 (0.13)	$p < 0.0001$
ICD-9'	9,063	79.28		84.18 (0.09)	
ICD-10	22,893	79.78	$p < 0.0001$	84.35 (0.20)	$p < 0.0001$
ICD-10'	14,644	79.23		83.12 (0.21)	

3.6. Discussion

For *L2LR* models, the results were consistent between the two prediction tasks. UMLS showed the best performance, whereas CCS single-level mapping was associated with the lowest AUROC on both prediction tasks and on both models, which is consistent with our

previous experiments.[4] There were no significant differences between ICD-9 and ICD-10 code mapping, although ICD-9 mappings are always higher in our experiments. The findings remain consistent even when we evaluated the differences using the DHF full cohort (Appendix D). There were no significant differences between CCS and CCSR codes in PC prediction, but the difference was significant for DHF prediction, which can be explained by the larger test set in the DHF cohort. In general, although LR models are not longitudinal, they are simple to use and have been the most commonly used models in EHR predictive modeling. Our results indicated that UMLS is often the top choice for predictive modeling when using LR models in our datasets.

For RNN models, the results vary between the different prediction tasks or between different cohort sizes. Whereas UMLS and PheWAS were the top-performing terminologies, their relative rankings change, depending on the tasks. PheWAS was the best-performing model for DHF in the selected sample cohort, whereas UMLS was the best performing for PC prediction. When evaluated using the DHF full cohort, raw data were associated with the best AUROC.

We note that it is not our main goal to benchmark models for realistic clinical tasks; therefore, the performance documented here does not necessarily translate to applicability in the real world. For example, PC risk prediction is a notoriously difficult task. Our PC performance may be due to biases in the data preparation. Nonetheless, our reported AUROCs are consistent with the range reported in previous studies for both DHF prediction[1,2,4] and PC prediction.[6,7]

We admit that, although the differences among groups are often statistically significant, the actual effect sizes are not necessarily large. For RNN models, the maximum difference in mean AUROC among UMLS, PheWAS, and raw data in the DHF and the PC cohorts were 0.4% and 1% respectively. The lower difference seen in the DHF cohort were owing to the larger cohort size, as RNN models can easily overfit on smaller cohorts. Nonetheless, the effect of terminologies appears independent of model architecture, and, thus, terminology choice has a real impact on predictive modeling.

Although the choice of model architecture (LR vs. RNN) has a major impact on prediction performance, the choice of terminologies also has a small but significant impact. Moreover, this impact is on top of the performance difference for model architectures. Therefore, terminology choice is a decision that has real-world impact.

To understand the key factors of terminologies that have an impact on prediction performance, we look at the characteristics of the best- and the worst-performing terminology mappings. There are two factors related to terminology mapping's influence on the accuracy of clinical prediction models from EHRs: quality of the terminology and the quality of mapping. Although mapping to more expressive terminology is a common practice for expressive deep-learning models, it is common for traditional machine-learning methods, such as logistic regression, to reduce dimensionality in search of a parsimonious model. Our results showed that, for both *L2LR* and RNN, large vocabulary sizes are associated with better performance. UMLS showed both the best performance in logistic regression models and high performance with deep-learning models; it is the vocabulary with the highest number of codes and has the advantage of better semantic

consistency and hierarchical relationships. Surprisingly, PheWAS, with a vocabulary size of only 1,820, showed good performance compared to other terminologies with higher levels of granularity. This can be attributed to the careful definitions of the mapping, as it was revised based on statistical co-occurrence, code frequency, and human review[12,13,26]. While the performance of the CCS- and CCSR-trained models were suboptimal during our experiments - mainly due to their smaller vocabulary sizes (284 and 538, respectively) - they may be still a good choice in practice due to their human readability. Not surprisingly, the use of raw data provides one of the best results as compared to other terminology-mapping exercises. Such a conclusion can give us assurance that models can learn from the current data without any further preprocessing. We can explain the good performance of the raw data-based models through two factors. First, the original coding type includes a level of important information for our prediction tasks. Second, the preprocessing and mapping exercise, although of high quality and including attention to detail, introduces some noise that may have impact on the model's learning ability. In our study, the raw data are represented by the Cerner-Diagnosis ID that maps to different terminologies, such ICD-9 and ICD-10.

Mapping structured raw data to UMLS-CUI can lead to better integration with diagnosis information extracted from the unstructured text as well as data recorded in other terminologies, such as SNOMED-CT. Further, it will be easier to embed knowledge about relationships between different clinical entities, including diseases, medications, procedures, laboratory tests, and so forth.

There are several limitations to this study. The first is the lack of measurement of the quality of the codes' mapping. We had observed a few incorrect ICD-9/ICD-10 codes in the Cerner diagnosis dictionary table, which could be due to data-entry typos. In addition, as we kept the hierarchical mapping, especially when using the UMLS codes, the one-to-many mappings of codes may require extra scrutiny. The second is that the prediction labels are derived from the raw data that are coded in either ICD-9 or ICD-10. This is reflected by the superior performance of ICD-9 over ICD-10, as the majority of data were coded in ICD-9. In addition, this may create a bias that favors ICD-9 or ICD-10 over other terminologies. The third is that our findings are shown to be valid for only the tested terminologies, tasks, models, and data sets. The generalizability of our results to other scenarios warrants further study. The fourth is that, for the sake of simplicity, we focused on only a single element of the EHR data: the diagnosis information. Future work that evaluates the terminology representation on other elements, including medication, procedures, and laboratory tests, as well as the interactions between the terminology of those elements is warranted. We also plan to evaluate the same on different tasks to validate the generalizability of our conclusion.

3.7. Conclusion

Through benchmarking, we found that the normalization of EHR diagnosis data to the UMLS standard was the best (or second best) performing among tested terminologies for both prediction tasks and both prediction models. For research purposes or local model development, raw data, when the sample size is large enough, are often sufficient to

achieve decent accuracy. If there is a need for diagnosis code grouping for dimension reduction, however, PheWAS, with fewer than 2,000 codes, is the best option. The quality of mapping had an impact on our study findings. In our data set, ICD-9 had better results than ICD-10 mainly because a larger proportion of the raw data was coded in ICD-9.

For a real-world project, when generalizability is a priority and the quality of terminology mapping is assured, we recommend normalization of terminologies to an expressive common terminology, such as UMLS. Due to information loss in translation in existing mapping tools, however, evaluation of mapping quality may be needed before determining the optimal target terminology for predictive modeling.

3.8. Acknowledgments

We would like to acknowledge the use of the Cerner HealthFacts® dataset and the assistance provided by the University of Texas Health Science Center in Houston (UTHealth) School of Biomedical Informatics (SBMI) Data Service team.

3.8.1. Funding

This work was supported by the Cancer Prevention and Research Institute of Texas (CPRIT) Grant No. RP170668; UTHealth Innovation for Cancer Prevention Research Training Program Pre-Doctoral Fellowship (CPRIT Grant No. RP160015); the National Cancer Institute (NCI) Grant No. 1U24CA194215; and the National Institutes of Health (NIH) Grant No. R01AI130460.

3.8.2. Authors' contribution

L.R. carried out the experiments and led the writing of the manuscript. F.T. developed the Cerner UMLS mappings. L.R. and Y.Z. extracted the EHR data. Y.X. and C.T. participated in the writing. H.X. and D.Z. conceived the original idea, contributed to the writing, and supervised the project. L.R. and D.Z. finalized the manuscript.

3.8.3. Competing interests

The authors have no competing interests to declare.

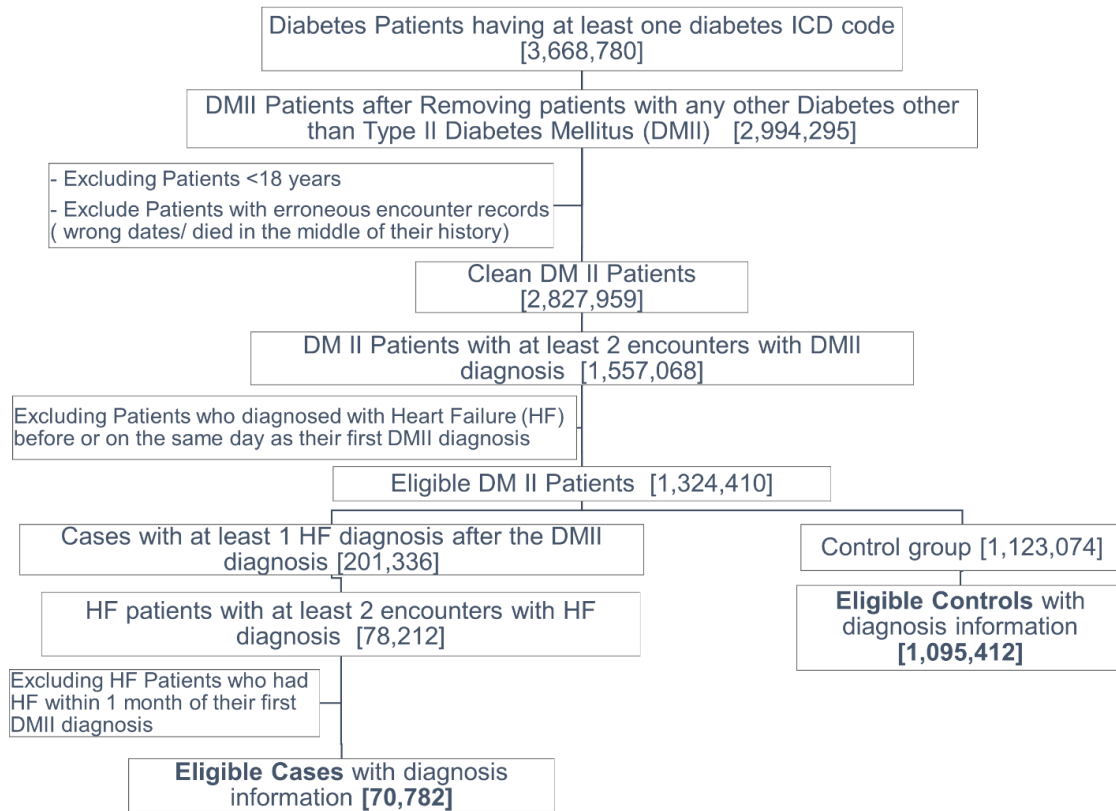
3.9. Supplementary Material

Appendix A. Cohort Definitions

We extracted our cohort from the Cerner Health Facts® dataset[26] which includes de-identified patient information from more than 600 hospitals for more than fifteen years period.

I) Diabetes Heart Failure Cohort (DHF)

The first prediction task is to predict diabetes patient risk to develop heart failure after at least one month from their diabetes diagnosis. We first identified patients who reported type II Diabetes Mellitus (DM II) for at least two encounters using their diagnosis ICD-9/10 codes, we excluded patients who reported any other form of diabetes including gestational or secondary DM from our cohort in order to avoid any chance of mislabeling.



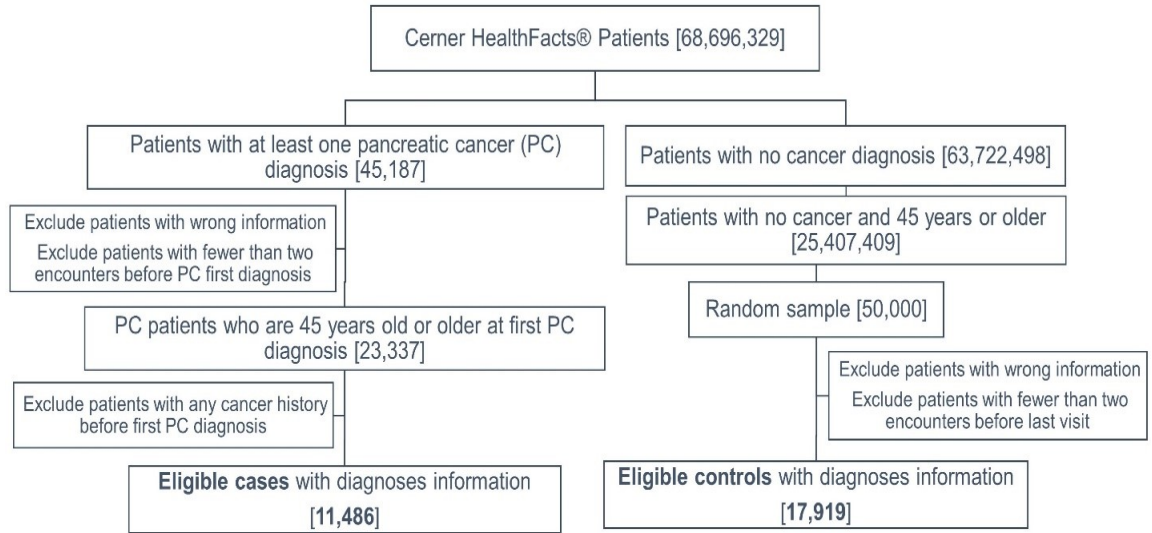
Supplementary 3 Figure 3: Flowchart for DHF cohort definition

For Cases, we identified patients with incidences of heart failure (HF) reported at least 30 days after their first DMII encounter and have at least two encounters with an HF ICD code. Cases should have at least two encounters with HF related ICD9 or ICD10 codes using the following condition: ICD-9_codes like '428%' or ICD-9_codes in ('404.03', '404.13', '402.11', '404.11', '402.01', '404.01', '402.91', '398.91', '404.93', '404.91') or ICD-10_codes like 'I50%' or ICD-10_codes in ('I11.0', 'I09.81', 'I13.2', 'I97.13', 'I97.131', 'I13.0', 'I97.130').

Further data cleaning, include the exclusion of patients with incorrect data, for example, patients who were recorded as expired in the middle of their encounters or had encounters with admission date after the discharge date. As a result, we had a full cohort of 70,782 cases and 1,095,412 Controls, out of which we randomly selected a sample of 60,000 cases and 60,000 controls for this study. Further details for cohort extraction in the chart below. For this study purpose as we focus on comparing different terminology normalization of diagnosis information using the same training and test sets. We restricted our sample to have at least two visits and no more than a hundred visits, and to have at least three Cerner diagnosis codes and no more than two hundred unique Cerner diagnosis code in their full history. Table 3.1 is showing the descriptive analysis of both full cohort and selected sample.

II) Pancreatic cancer cohort (PC)

Using ICD-9 codes starting with 157 and ICD-10 codes starting with C25, we originally identified around 45,000 pancreatic cancer (PC) patients out of which 11,486 cases, who are 45 years or older and didn't report any other cancer disease before their first pancreatic cancer diagnosis were eligible for inclusion in this cohort. Further details of the cohort definition in the figure below



Supplementary 3 Figure 4: Flowchart for PC cohort definition

Appendix B. Terminology mappings

I) Mapping between ICD-10 and ICD-9

It is a common belief that, for semantic consistency, mapping to a single coding system should be a good practice[16]. The most commonly used terminologies in clinical practice, for billing purposes, are ICD-9 and ICD-10. We assumed that mapping from ICD-10 to ICD-9 codes should provide a more compact representation with a relatively fewer number of codes and should be easy to map, but the performance may be compromised by the associated information loss. On the other hand, if mapping from ICD-9 to ICD-10, while keeping the level of granularity, the accuracy of the super-resolution mapping remains questionable.

The Center of Medicare and Medicaid Services(CMS) provides good tools to map ICD-9 to ICD-10 and vice versa. So we downloaded the most recent general equivalence mapping (GEM) version 2018 from the CMS website[20]. We used both the ICD-10 to ICD-9 and ICD-9 to ICD-10 files, we mapped those to our Cerner diagnosis table, except for those codes that are explicitly documented as ‘NoDx’ in the mapping column which is around 350 on both sides. All the ICD-9 codes used in our cohort were perfectly mapped to ICD-10 codes. On the other side, mapping ICD-10 to ICD-9 was not as perfect as we originally assumed, there were around 300 ICD-10 codes in our data were not mapped to a corresponding ICD-9 code. We decided to ignore those codes as they were appearing for a maximum of 10 times in our cohort. In order to understand if the difference in the predictive model accuracy using the ICD-9 or ICD-10 codes is due to the terminology representation itself or due to the information loss induced by the mapping process, we

simply converted the previously converted ICD-9 codes to ICD-10 and named it ICD-10' and did the same for the previously converted ICD-10 codes and converted them back to ICD-9'. For that revert prime mapping, we used only the original mapping files provided by CMS without any further review or improvement.

II) Unified Medical language systems

Unified Medical language systems (UMLS) is a very effective tool when collecting clinical data from different resources like literature, clinical notes, EHR structured data, insurance, and billing information, as it integrates key medical terminology and coding standards, like SNOMED-CT, ICD-9/10, RxNorm, Loinc, and MESH terms. In order to test the model performance using UMLS codes, we downloaded the UMLS knowledge sources[21] and mainly use the information available at the MRSTY and MRCONSO data files. We extracted the UMLS_CUI, ICD Code, ICD Type from the MRCONSO file. We considered expanding the ICD code hierarchy for the entries that appear in MRCONSO as an interval of ICD codes. In order to avoid confusion between diagnoses and procedures codes that might overlap, we added the Semantics column to using the MRSTY file. Finally, we used the UMLS-CUI to ICD-9 and 10 codes to map it to the original Cerner diagnosis id. There were only a hundred Cerner diagnosis id not mapped to UMLS-CUI, but each of those codes appeared at maximum three times in our cohort so we decided to ignore those.

III) Clinical Classification Software

Unlike UMLS, the Clinical Classification Software (CCS) codes provided by the agency of healthcare research and quality (AHRQ) as a part of the Healthcare Cost and Utilization Project (HCUP), maps ICD-9 /10 codes to a higher level of details. Original CCS used to

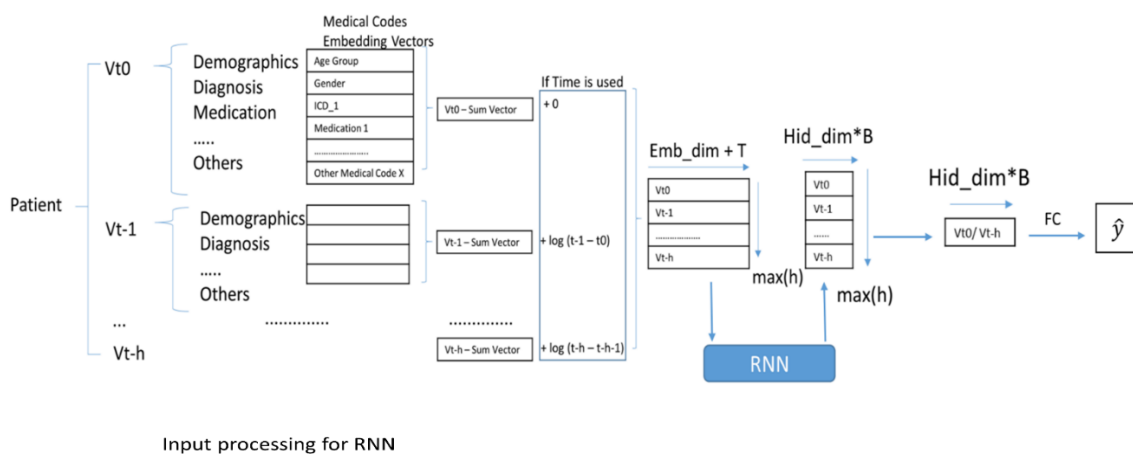
have two formats, a single level and a multilevel. Both have nearly the same number of codes around 284, but the multilevel includes more hierarchical information. For simplicity, we decided to use the single level CCS mappings especially it perfectly matches both ICD-9 and ICD-10 codes which is not the case with the multilevel mapping as the mapping to ICD-10 was replaced by the CCS Refined (CCSR) version. For our mapping, we used the latest version of ICD-9 to CCS mapping and the ICD-10 to CCS mapping available on the AHRQ website[23,24]. Additionally, we evaluated the refined CCSR version[25] which aggregates more than 70,000 ICD-10 codes to around 540 CCSR codes.

IV) The Phenome-wide association studies

The Phenome-wide association studies (PheWAS) defined around 1866 Phecode that represents mainly diseases. Those codes are following the three-digit grouping of the ICD-9 codes. PheWAS code grouping has been revised based on statistical co-occurrence, code frequency, and human review[22]. We downloaded the Phecode map with ICD-9, ICD-10, and ICD10-CM codes and used it to map the raw data to the PheWAS codes.

Appendix C. Recurrent Neural Network model architecture

Recurrent Neural Network (RNN) models are appropriate for modeling the sequential nature of patient medical records and have been shown to provide high predictive accuracy in the healthcare domain.



Supplementary 3 Figure 5: RNN based model training

As appears in Supplementary Figure 3, we represented a patient record as a sequence of visits and each visit as a set of clinical codes. In this study, we only used ‘diagnosis’ data for terminology comparison. Our codebase available on https://github.com/ZhiGroup/pytorch_ehr facilitates the evaluation of different RNN based models architectures.

For RNN evaluation in this study, we used a basic single layer bi-directional gated recurrent unit (GRU) with a hidden dimension of 64. Our input is structured as mini-batches each of 128 patients, each patient is a sequence of visits where each visit is the

sum of the diagnosis codes embeddings vectors within this visit. We also included the time difference between visits and fed to the model as appear in Supplementary Figure 3.

For comparison purposes, We fixed the hyperparameters for all experiments, we used equal embedding and hidden dimensions of 64. We used Adamax optimizer with a learning rate of 0.01 and an L2 penalty of 0.00001.

Appendix D. Diabetes Heart Failure full Cohort Results

The Diabetes Heart Failure (DHF) full cohort is an unbalanced set and it doesn't include any restrictions on the maximum number of codes or visits per patient.

Supplementary 3 Table 11: LR and RNN results of DHF full cohort

Diagnosis Terminology	Number of unique codes	<i>L2LR</i>	RNN
Raw Data (ICD-9 + ICD-10)	26,427	81.49	85.86
CCS – Single Level	284	77.75	81.97
CCSR	538	78.92	83.02
ICD-9	11,187	80.87	85.20
ICD-10	22,893	80.67	84.2
PheWAS	1,820	80.69	85.07
UMLS CUI	29,491	81.96	85.52

Appendix E. DHF LR Additional Results

We run different variations of LR models on the DHF cohort to further validate our findings. The penalty parameters tuning was done using the grid search cross-validation model selection function from the Scikit-learn package (GridSearchCV).

Supplementary 3 Table 12: Results of different variations of LR models the DHF cohort

Diagnosis Terminology	LR (No Penalty)	<i>L1</i> /LR (C=0.1)	<i>L2</i> LR (Tuned) (C=0.01)	<i>L2</i> LR* (C =1)
Raw Data (ICD-9 + ICD-10)	78.93%	82.30%	82.28% (C=0.01)	80.61%
CCS-single level	78.07%	78.09%	78.07% (C=0.01)	78.07%
CCSR	78.86%	78.90%	78.92% (C=0.1)	78.87%
ICD-9	79.12%	81.56%	81.64% (C=0.1)	80.12%
ICD-10	78.53%	81.24%	81.04% (C=0.1)	79.78%
PheWAS	80.63%	81.09%	81.11% (C=0.01)	80.71%
UMLS CUI	81.12%	82.81%	82.88% (C=0.01)	81.15%

Where C is the inverse of the regularization strength associated with the best performance based on cross-validation results. * *L2*LR with C=1 is the default parameter used in this study


The best L1 penalty hyperparameter was consistent among all terminologies (C=0.1), while the L2 penalty hyperparameter varied among different terminologies. UMLS showed the best performance among all tested variations.

Appendix F. Statistical Significance using Tukey-Kramer HSD






Below are the pairwise significance for RNN models for both cohorts










I) *DHF Cohort* $\alpha=0.05$

Level	- Level	Diff.	Std Err Diff	Lower CL	Upper CL	p-Value	
PHEWAS	CCS	0.0291	0.0007	0.0271	0.0311	<.0001*	
UMLS	CCS	0.0259	0.0007	0.0239	0.0279	<.0001*	
Raw	CCS	0.0251	0.0007	0.0231	0.0271	<.0001*	
ICD9	CCS	0.0224	0.0007	0.0204	0.0243	<.0001*	
PHEWAS	CCSR	0.0170	0.0007	0.0150	0.019	<.0001*	
PHEWAS	ICD10	0.0152	0.0007	0.0133	0.0172	<.0001*	
ICD10	CCS	0.0139	0.0007	0.0119	0.0158	<.0001*	
UMLS	CCSR	0.0138	0.0007	0.0118	0.0158	<.0001*	
Raw	CCSR	0.0131	0.0007	0.0111	0.0150	<.0001*	
CCSR	CCS	0.0121	0.0007	0.0101	0.0141	<.0001*	
UMLS	ICD10	0.0120	0.0007	0.0101	0.0140	<.0001*	
Raw	ICD10	0.0113	0.0007	0.0093	0.0133	<.0001*	

ICD9	CCSR	0.0103	0.0007	0.0083	0.0123	<.0001*	
ICD9	ICD10	0.0085	0.0007	0.0065	0.0105	<.0001*	
PHEWAS	ICD9	0.0067	0.0007	0.0047	0.0087	<.0001*	
PHEWAS	Raw	0.004	0.0007	0.002	0.006	<.0001*	
UMLS	ICD9	0.0035	0.0007	0.0015	0.0055	<.0001*	
PHEWAS	UMLS	0.0032	0.0007	0.0012	0.0052	0.0001*	
Raw	ICD9	0.0028	0.0007	0.0008	0.0048	0.0014*	
ICD10	CCSR	0.0018	0.0007	-0.0002	0.0038	0.1087	
UMLS	Raw	0.0008	0.0007	-0.001	0.0027	0.9013	

II) PC Cohort $\alpha=0.05$

Level	- Level	Difference	Std Err	Lower CL	Upper CL	p-Value	
			Dif	CL			
UMLS	CCS	0.0321966	0.0015791	0.027387	0.0370060	<.0001*	
UMLS	ICD10	0.0297051	0.0015791	0.024896	0.0345145	<.0001*	
UMLS	CCSR	0.0261337	0.0015791	0.021324	0.0309430	<.0001*	
Raw	CCS	0.0240043	0.0015791	0.019195	0.0288137	<.0001*	
Raw	ICD10	0.0215128	0.0015791	0.016703	0.0263221	<.0001*	

PheWAS	CCS	0.0212722	0.0015791	0.016463	0.0260815	<.0001*	
PheWAS	ICD10	0.0187807	0.0015791	0.013971	0.0235900	<.0001*	
Raw	CCSR	0.0179414	0.0015791	0.013132	0.0227507	<.0001*	
ICD9	CCS	0.0175817	0.0015791	0.012772	0.0223911	<.0001*	
PheWAS	CCSR	0.0152092	0.0015791	0.010400	0.0200186	<.0001*	
ICD9	ICD10	0.0150902	0.0015791	0.010281	0.0198996	<.0001*	
UMLS	ICD9	0.0146149	0.0015791	0.009806	0.0194242	<.0001*	
ICD9	CCSR	0.0115188	0.0015791	0.006709	0.0163281	<.0001*	
UMLS	PheWAS	0.0109244	0.0015791	0.006115	0.0157338	<.0001*	
UMLS	Raw	0.0081923	0.0015791	0.003383	0.0130017	<.0001*	
Raw	ICD9	0.0064226	0.0015791	0.001613	0.0112319	0.0025*	
CCSR	CCS	0.0060629	0.0015791	0.001254	0.0108723	0.0051*	
PheWAS	ICD9	0.0036905	0.0015791	-0.00112	0.0084998	0.2432	
CCSR	ICD10	0.0035714	0.0015791	-0.00124	0.0083808	0.2788	
Raw	PheWAS	0.0027321	0.0015791	-0.00208	0.0075415	0.5989	
ICD10	CCS	0.0024915	0.0015791	-0.00232	0.0073009	0.6967	

3.10. References

- 1 Maragatham G, Devi S. LSTM Model for Prediction of Heart Failure in Big Data. *J Med Syst* 2019;**43**:111. doi:10.1007/s10916-019-1243-3
- 2 Choi E, Bahadori MT, Kulas JA, *et al.* RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. *Adv Neural Inf Process Syst* 2016;;3504–12.<http://papers.nips.cc/paper/6321-retain-an-interpretable-predictive-model-for-healthcare-using-reverse-time-attention-mechanism> (accessed 29 Dec 2017).
- 3 Choi E, Schuetz A, Stewart WF, *et al.* Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc* 2017;**24**:361–70. doi:10.1093/jamia/ocw112
- 4 Rasmy L, Zheng WJ, Xu H, *et al.* A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. *J Biomed Inform* 2018;**84**. doi:10.1016/j.jbi.2018.06.011
- 5 Jin B, Che C, Liu Z, *et al.* Predicting the Risk of Heart Failure with EHR Sequential Data Modeling. *IEEE Access* 2018;**6**:9256–61. doi:10.1109/ACCESS.2017.2789324
- 6 Muhammad W, Hart GR, Nartowt B, *et al.* Pancreatic Cancer Prediction Through an Artificial Neural Network. *Front Artif Intell* 2019;**2**:2.

doi:10.3389/frai.2019.00002

- 7 Hsieh MH, Sun L-M, Lin C-L, *et al.* Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. *Cancer Manag Res* 2018;**Volume 10**:6317–24. doi:10.2147/CMAR.S180791
- 8 Ayala Solares JR, Diletta Raimondi FE, Zhu Y, *et al.* Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J. Biomed. Inform.* 2020;**101**:103337. doi:10.1016/j.jbi.2019.103337
- 9 Min X, Yu B, Wang F. Predictive Modeling of the Hospital Readmission Risk from Patients' Claims Data Using Machine Learning: A Case Study on COPD. *Sci Rep* 2019;**9**. doi:10.1038/s41598-019-39071-y
- 10 Rajkomar A, Oren E, Chen K, *et al.* Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018;**1**:18. doi:10.1038/s41746-018-0029-1
- 11 Subramanyam KK, S S. Deep Contextualized Medical Concept Normalization in Social Media Text. *Procedia Comput Sci* 2020;**171**:1353–62. doi:10.1016/j.procs.2020.04.145
- 12 Wei W-Q, Bastarache LA, Carroll RJ, *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* 2017;**12**:e0175508. doi:10.1371/journal.pone.0175508
- 13 Wu P, Gifford A, Meng X, *et al.* Developing and Evaluating Mappings of ICD-10

- and ICD-10-CM codes to Phecodes. *bioRxiv* 2018;;462077. doi:10.1101/462077
- 14 Thompson WK, Rasmussen L V, Pacheco JA, *et al.* An evaluation of the NQF Quality Data Model for representing Electronic Health Record driven phenotyping algorithms. *AMIA . Annu Symp proceedings AMIA Symp* 2012;**2012**:911–20.<http://www.ncbi.nlm.nih.gov/pubmed/23304366> (accessed 13 Mar 2019).
 - 15 Choi E, Xiao C, Stewart W, *et al.* MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare. 2018;;4547–57.<http://papers.nips.cc/paper/7706-mime-multilevel-medical-embedding-of-electronic-health-records-for-predictive-healthcare> (accessed 21 Nov 2019).
 - 16 Beam AL, Kompa B, Fried I, *et al.* Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. Published Online First: 4 April 2018.<http://arxiv.org/abs/1804.01486> (accessed 13 Mar 2019).
 - 17 Alawad M, Hasan SMS, Blair Christian J, *et al.* Retrofitting Word Embeddings with the UMLS Metathesaurus for Clinical Information Extraction. In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE 2018. 2838–46. doi:10.1109/BigData.2018.8621999
 - 18 Xiang Y, Xu J, Si Y, *et al.* Time-sensitive clinical concept embeddings learned from large electronic health records. *BMC Med Inform Decis Mak* 2019;**19**:58. doi:10.1186/s12911-019-0766-3
 - 19 Feng Y, Min X, Chen N, *et al.* Patient outcome prediction via convolutional neural networks based on multi-granularity medical concept embedding. *ieeexplore.ieee.org*<https://ieeexplore.ieee.org/abstract/document/8217753/>

- (accessed 30 Sep 2019).
- 20 Jung K, Sudat SEK, Kwon N, *et al.* Predicting need for advanced illness or palliative care in a primary care population using electronic health record data. *J Biomed Inform* 2019;**92**:103115. doi:10.1016/j.jbi.2019.103115
 - 21 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267-70. doi:10.1093/nar/gkh061
 - 22 Choi Y, Chiu CY-I, Sontag D. Learning Low-Dimensional Representations of Medical Concepts. *AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci* 2016;**2016**:41–50.<http://www.ncbi.nlm.nih.gov/pubmed/27570647> (accessed 2 Oct 2018).
 - 23 Maldonado R, Yetisgen M, Harabagiu SM. Adversarial Learning of Knowledge Embeddings for the Unified Medical Language System. *AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci* 2019;**2019**:543–52.<http://www.ncbi.nlm.nih.gov/pubmed/31259009> (accessed 8 Jun 2020).
 - 24 UMLS Knowledge Sources: File Downloads. <https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html> (accessed 13 Mar 2019).
 - 25 2018-ICD-10-CM-and-GEMs. Published Online First: 2017.<https://www.cms.gov/medicare/coding/icd10/2018-icd-10-cm-and-gems.html> (accessed 13 Mar 2019).
 - 26 PheWAS - Phenome Wide Association Studies.

- <https://phewascatalog.org/phecodes> (accessed 13 Mar 2019).
- 27 Beta Clinical Classifications Software (CCS) for ICD-10-CM/PCS.
<https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp> (accessed 13 Mar 2019).
- 28 Elixhauser A, Steiner C, Palmer L. Clinical Classifications Software (CCS). 2015.
- 29 Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses.
https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp (accessed 30 Mar 2020).
- 30 Cerner - Cerner Health Facts ® - Data Sets - SBMI Data Service - The University of Texas Health Science Center at Houston (UTHealth) School of Biomedical Informatics. <https://sbmi.uth.edu/sbmi-data-service/data-set/cerner/> (accessed 25 Nov 2018).
- 31 `sklearn.linear_model.LogisticRegression` — scikit-learn 0.20.3 documentation.
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed 13 Mar 2019).
- 32 Ma F, Chitta R, Zhou J, *et al.* Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. Published Online First: 18 June 2017. doi:10.1145/3097983.3098088
- 33 Ma F, Chitta R, Zhou J, *et al.* Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. Published Online First: 18 June 2017. doi:10.1145/3097983.3098088

- 34 Rasmy, L., Zhu, J., Li, Z., Tran, HTN., Wu, Y., Zhou, Y., Tiryaki, F., Xiang, Y., Xu, H. and Zhi, D. 2018. Medinfo 2019 (podium abstract submitted Nov 2018). Simple Recurrent Neural Networks is all we need for clinical events predictions using EHR data. In: *MedInfo 2019*. 2019.
- 35 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;**44**:837–
- 45.<http://www.ncbi.nlm.nih.gov/pubmed/3203132> (accessed 13 Mar 2019).

Chapter 4: Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction

This chapter is adapted from:

Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*. 2021 May 20;4(1):86. doi: 10.1038/s41746-021-00455-y. PMID: 34017034; PMCID: PMC8137882.

Title: Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction

Authors: Laila Rasmy^{1†}, M.S., Yang Xiang^{2†*}, Ph.D., Ziqian Xie^{1†}, Ph.D., Cui Tao¹, Ph.D., Degui Zhi^{1*}, Ph.D.

¹ School of Biomedical Informatics, University of Texas Health Science Center at Houston, U.S.

² Peng Cheng Laboratory, Shenzhen, China

†Co-first authors with equal contributions

*Corresponding authors: Degui.Zhi@uth.tmc.edu or xiangy@pcl.ac.cn

Keywords: Pretrained Contextualized Embeddings, Disease Prediction, Electronic Health Records, Structured Data, Deep Learning, Transformers, BERT, Transfer Learning

4.1. Abstract

Deep learning (DL) based predictive models from electronic health records (EHRs) deliver impressive performance in many clinical tasks. Large training cohorts, however, are often required by these models to achieve high accuracy, hindering the adoption of DL-based models in scenarios with limited training data. Recently, bidirectional encoder representations from transformers (BERT) and related models have achieved tremendous successes in the natural language processing domain. The pre-training of BERT on a very large training corpus generates contextualized embeddings that can boost the performance of models trained on smaller datasets. Inspired by BERT, we propose Med-BERT, which adapts the BERT framework originally developed for the text domain to the structured EHR domain. Med-BERT is a contextualized embedding model pre-trained on a structured EHR dataset of 28,490,650 patients. Fine-tuning experiments showed that Med-BERT substantially improves the prediction accuracy, boosting the area under the receiver operating characteristics curve (AUC) by 1.21-6.14% in two disease prediction tasks from two clinical databases. In particular, pre-trained Med-BERT obtains promising performances on tasks with small fine-tuning training sets and can boost the AUC by more than 20% or obtain an AUC as high as a model trained on a training set 10 times larger, compared with deep learning models without Med-BERT. We believe that Med-BERT will benefit disease-prediction studies with small local training datasets, reduce data collection expenses, and accelerate the pace of artificial intelligence aided healthcare.

4.2. Introduction

Artificial intelligence (AI)-aided disease prediction has undergone considerable development in recent years[1-3]. At present, it can improve the precision of diagnosis, enable disease prevention by early warning, streamline clinical decision making, and reduce healthcare costs[4-7]. Powerful AI tools, advanced conventional machine learning[8-10], and deep-learning[11-14] approaches also have been widely applied in clinical predictive modeling and have gained numerous successes. Given enough training samples, deep-learning models can achieve comparable or even better performance than domain experts in the diagnosis of certain diseases[15-19]. One prerequisite of typical deep-learning-based methods is the availability of large and high-quality annotated datasets, which are used to model the underlying complex semantics of the input domain as much as possible and to avoid under-fitting of model training[20 21]. Big EHR data, however, often are not accessible for numerous reasons, including the limited number of cases for new or rare conditions; difficulty in data cleaning and annotation, especially if collected from different sources; and governance issues that hinder the data acquisition[22].

Transfer learning was developed to address the issue whereby some representations were first pre-trained on large volumes of unannotated datasets and then further adapted to guide other tasks[23]. A recent trend in transfer learning is to use self-supervised learning over large general datasets to derive a general-purpose pre-trained model that captures the intrinsic structure of the data, which can be applied to a specific task with a specific dataset by fine-tuning. This pre-training-fine-tuning paradigm has been proven to be

extremely effective in natural language processing (NLP)[24-30] and, recently, computer vision[31 32]. Bidirectional encoder representations from transformers (BERT) is one of the most popular models for handling sequential inputs, e.g., text, with numerous variations[33-40]. BERT also has been embraced by the clinical domain[33 34 41]. However, these models were pre-trained on clinical text and are only for clinical NLP tasks.

Structured EHR, as a primary input source for disease prediction, offers rich and well-structured information that reflects the disease progression of each patient and is one of the most valuable resources for health data analysis[42 43]. Adapting the transfer learning framework to structured EHR is a natural idea based on the analogy between natural language text and EHR, i.e., both are sequential modalities for tokens from a large vocabulary. However, a one-to-one mapping between the elements of natural language and structured EHR is not available.

There is a growing literature on transfer learning for EHR. Some researchers directly repurpose internal layers of trained deep models (e.g., RNN) for an existing task to a new task[44] but these transfer learning might be too tightly coupled with specific tasks and its generalizability has not been well established. For the pre-training style transfer learning, previous studies on structured EHR showed some successes[45 46] but they mainly focused on static embeddings such as word2vec[24] and GloVe[47], which failed to capture deep context information.

In this work, we choose the BERT framework, including its architecture and its training methodology, for training models on large EHR data. Notably, other contextualized pre-

trained embedding frameworks from the NLP domain, such as ULMFiT[48] and ELMo[49], could also be tested in the EHR domain. However, we choose BERT in this work because it is widely adopted with proven success.

To the best of our knowledge, there are only two relevant studies in the literature of the clinical domain: BEHRT[50] and G-BERT[51]. These models, however, have the following limitations. BEHRT aims to develop pre-trained models to predict the existence of any medical codes in certain visits. It uses positional embeddings to distinguish different visits and adds an *age* layer to imply temporal orders. The authors' definition of the area under receiver operating characteristics (AUC), however, was a non-standard one, making it difficult to compare their results with previous studies. G-BERT applied a graph neural network (GNN) model to expand the context of each clinical code through ontologies and jointly trained the GNN and BERT embeddings. It modified the masked language model (Masked LM) pre-training task into domain-specific ones, including maximizing the gap between the existing and non-existing codes and using different types of codes to predict each other. However, G-BERT's inputs are all single-visit samples, which are insufficient to capture long-term contextual information in EHR. In addition, the size of their pre-training dataset is not large, making it difficult to evaluate its full potential. Furthermore, neither BEHRT nor G-BERT uses disease-prediction tasks as the evaluation of their pre-trained model by fine-tuning. To alleviate the aforementioned issues and to evaluate a pre-trained contextualized embedding model specific to disease prediction, we designed Med-BERT, an adaption of the BERT methodology for the structured EHR modality. Med-BERT is trained on

structured diagnosis data coded using the International Classification of Diseases (ICD) codes, unlike the original BERT and most of its variations that were trained on free text. Note that we can also include other types of codes such as medications and laboratory tests, and we leave its investigation as future work.

We compare Med-BERT with BEHRT and G-BERT in Table 4.1. Remarkably, Med-BERT has a much larger vocabulary and a much larger pre-training cohort than the other two models, which help to provide a reality check of EHR BERT-based models. The larger cohort size and longer visit sequences in Med-BERT’s pre-training set will greatly benefit the model in learning more comprehensive contextual semantics. We also believe that, by using a large and publicly accessible vocabulary, i.e., ICD-9 and ICD-10, and pre-training the model on a multi-institutional dataset (Cerner), Med-BERT will likely be easily deployable to different institutions and clinical scenarios. Further, among all these pre-trained models, only Med-BERT has been successfully cross-tested by a fine-tuning task on an external data source (Truven).

Table 4.8: Comparison of Med-BERT with BEHRT and G-BERT from multiple perspectives.

Criteria	BEHRT	G-BERT	Med-BERT
Type of input code	Caliber code for diagnosis developed by a college in London	Selected ICD-9 code for diagnosis + ATC code for medication	ICD-9 + ICD-10 code for diagnosis
Vocabulary size	301	<4K	82K

Pre-training data source	CPRD (primary care data) ⁶¹	MIMIC III (ICU data) ⁶²	Cerner HealthFacts (general EHR)
Input structure	Code + visit + age embeddings	Code embeddings from ontology + visit embeddings	Code + visit + code serialization embeddings
Pre-training sample unit	Patient’s visit sequence	Single visit	Patient’s visit sequence
Total number of pre-training patients	1.6M	20K	20M
Average number of visits for each patient for pre-training	Not reported but > 5	<2	8
Pre-training task	Masked LM	Modified Masked LM	Masked LM + prediction of prolonged length of stay in hospital
Evaluation task	Diagnosis code prediction in different time windows	Medication code prediction	Disease predictions according to strict inclusion/exclusion criteria
Total number of patients in evaluation tasks	699K, 391K, and 342K for different time windows	7K	50K, 20K, and 20K for three task cohorts

Similar to BEHRT and G-BERT, Med-BERT made several modifications to the overall BERT methodology to fit the EHR data modality. Med-BERT used code embeddings to represent each clinical code, visit embeddings to differentiate visits, and the transformer structure to capture the inter-correlations between codes. Within each visit, we defined *serialization embeddings* to denote the relative order of each code, whereas neither

BEHRT nor G-BERT introduced code ordering within a visit. In addition, we designed a domain-specific pre-training task *prediction of prolonged length of stay in hospital* (Prolonged LOS), which is a popular clinical problem that requires contextual information modeling to evaluate the severity of a patient’s health condition according to the disease progression and requires no human annotation. We expect that the addition of this task can help the model to learn more clinical and more contextualized features for each visit sequence and facilitate certain tasks.

The usefulness of the pre-trained Med-BERT was evaluated by fine-tuning on the following two disease-prediction tasks: *the prediction of heart failure among patients with diabetes* (DHF) and *the prediction of onset of pancreatic cancer* (PaCa), using three patient cohorts from two different EHR databases, Cerner Health Facts® and Truven Health MarketScan®. These tasks are different from the pre-training prediction tasks (Masked LM and Prolonged LOS) and, thus, are good evaluation tasks to test the generalizability of the pre-trained model. In addition, we chose these tasks because they capture more complexity than merely the existence of certain diagnosis codes, and are based on established phenotyping algorithms that further integrate multiple pieces of information beyond diagnosis codes, such as constraints on time window, diagnosis occurrence times, medications, and laboratory test values.

Fine-tuning experiments were conducted for the following purposes: (1) to test the performance gains by adding Med-BERT on three state-of-the-art predictive models; (2) to compare Med-BERT with a pre-trained non-contextualized embedding, the clinical

word2vec-style embedding[52]; and (3) to see how much Med-BERT would contribute to disease predictions with different fine-tuning training sizes.

Our primary contributions are summarized as follows:

1. This work is the first proof-of-concept demonstration that a BERT-style model for structured EHR can deliver a meaningful performance boost in real-world-facing predictive modeling tasks.
2. We innovatively designed a domain-specific cross-visit pre-training task that is prevalent among EHR data and is effective in capturing contextual semantics.
3. This work is the first demonstration of significantly boosted performance over state-of-the-art methods on multiple clinical tasks with phenotyped cohorts.
4. This work is the first that presents the generalizability of EHR BERT models by boosting the performance in a dataset (Truven) other than the training dataset (Cerner).
5. The performance boost of Med-BERT is observed across all sample sizes, demonstrating the enabling power of pre-trained models for clinical tasks for which only limited training data are available.
6. We provided a visualization tool to demonstrate the dependency semantics in EHRs, facilitating the interpretability of the model.
7. We made our pre-trained models and code available, enabling its applications by other researchers.

4.3. Methods

4.3.1. Cohort Definition

4.3.1.1. *Med-BERT pre-training cohort*

Cerner Health Facts® (version 2017) is a de-identified EHR database that consists of over 600 hospitals and clinics in the United States. It represents over 68 million unique patients and includes longitudinal data from 2000 to 2017. The database consists of patient-level data, including demographics, encounter meta-information, diagnoses, procedures, lab results, medication orders, medication administration, vital signs, microbiology, surgical cases, other clinical observations, and health systems attributes. Data in Health Facts® are extracted directly from the EMRs of hospitals with which Cerner has a data use agreement. Encounter meta-information includes the identification of pharmacy, clinical and microbiology laboratory, and admission and billing information from affiliated patient care locations. All admissions, medication orders and dispensing, laboratory orders, and specimens are date and time-stamped, providing a temporal relationship between treatment patterns and clinical information. The Cerner Corporation has established Health Insurance Portability and Accountability Act-compliant operating policies to establish de-identification for Health Facts®.

During the data preprocessing phase for pretraining, for each patient in the cohort, we organized the visits in a temporal order and ranked the diagnosis codes within each visit according to three criteria: (1) the diagnosis was flagged as *present* on admission; (2) the diagnosis was captured during the visit (e.g., hospitalization) or only at the billing phase; and (3) the diagnosis priority is provided by the Cerner database, indicating some priorities

of the diagnoses, e.g., principal/secondary diagnosis (the priority is provided by the database, but it might not be a perfect priority ranking)

For each visit, we extracted the diagnosis codes (represented by ICD, Ninth Revision, Clinical Modification (ICD-9) and ICD, Tenth Revision, Clinical Modification (ICD-10)) and the length of stay in hospital. We then ranked the codes in each visit according to the above three criteria and determined the order by using (1) \rightarrow (2) \rightarrow (3) in sequence. We observed only very limited performance gains, however, by adding the code order during the evaluation, compared with randomly scattering the codes. Hence, we set it as a placeholder here and assume that more effective orders could be defined in the future.

Patients with fewer than three diagnosis codes in their records as well as those with wrong recorded time information, e.g., discharge date before admission date, were removed from the population. In total, we had 28,490,650 unique patients (Figure 4.1), which were further separated into training, valid, and testing sets by the ratio of 7:1:2 on both the pre-training and evaluation phases.

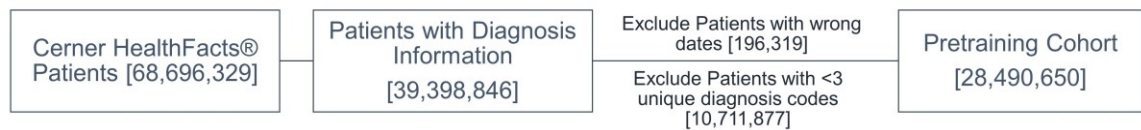


Figure 4.4. Selection pipeline for the pre-training cohort.

4.3.1.2. Diabetes heart failure cohort (DHF)

We originally identified 3,668,780 patients with at least one encounter with a diabetes diagnosis, based on the associated ICD-9/10 codes. We decided to exclude patients with any history of diabetes insipidus, gestational diabetes, secondary diabetes, neonatal diabetes mellitus, or type I diabetes mellitus (DM) from our cohort, as we focus on patients with type II DM and need to avoid any chance of wrong coding, taking into consideration that most of the EHR data are based on user manual entries and that there is a high associated chance of data entry mistakes. For the same reason, we decided to include patients who have more than one encounter with a diabetes diagnosis code. In addition, for type II DM patients, we verified that the patients' A1C reading is ≥ 6.5 or that they are taking an antidiabetic agent, including metformin, chlorpropamide, glimepiride, glyburide, glipizide, tolbutamide, tolazamide, pioglitazone, rosiglitazone, sitagliptin, saxagliptin, alogliptin, linagliptin, repaglinide, nateglinide, miglitol, acarbose, or insulin.

For these cases, we identified patients with incidences of heart failure (using ICD-9 code equivalents, such as 428, or in 404.03, 404.13, 402.11, 404.11, 402.01, 404.01, 402.91, 398.91, 404.93, and 404.91, or ICD-10 code equivalents, such as I50%, or in I11.0, I09.81, I13.2, I97.13, I97.131, I13.0, and I97.130). In addition, we verified that the eligible cases are either prescribed a diuretic agent, had high B-type natriuretic peptide (BNP) or had been subjected to relevant procedures, including dialysis or an artificial heart-associated procedure following⁶³. We included only those patients who reported heart failure (HF) at least 30 days after their first encounter with a type II DM code and excluded patients with only one HF encounter.

Further data cleaning included the exclusion of patients with incorrect or incomplete data, for example, patients who were recorded as expired in between their first encounter and our event (first HF encounter for cases or last encounter for controls) as well as patients who are younger than 18 years old at their first diabetes diagnosis. The final cohort is shown in Supplementary Figure 1 and includes 39,727 cases and 632,920 controls.

4.3.1.3. Pancreatic cancer cohort (PaCa)

Using ICD-9 codes that start with 157 and ICD-10 codes that start with C25, we originally identified around 45,000 pancreatic cancer patients from the Cerner HealthFacts dataset, of which 11,486 cases of individuals of 45 years or older did not report any other cancer disease before their first pancreatic cancer diagnosis were eligible for inclusion in this cohort. Further details of the cohort definition are shown in Supplementary Figure 2.

Similarly, we extracted a PaCa cohort from Truven Health MarketScan® Research Databases for evaluation purposes. The Truven Health MarketScan® Research Databases (version 2015) are a family of research data sets that fully integrate de-identified patient-level health data (medical, drug, and dental), productivity (workplace absence, short- and long-term disability, and workers' compensation), laboratory results, health risk assessments, hospital discharges, and electronic medical records into datasets available for healthcare research. It captures person-specific clinical utilization, expenditures, and enrollment across inpatient, outpatient, prescription drug, and carve-out services. The annual medical databases include private-sector health data from approximately 350 payers. Historically, more than 20 billion service records are available in the MarketScan databases. These data represent the medical experience of insured employees and their

dependents for active employees, early retirees, Consolidated Omnibus Budget Reconciliation Act (COBRA) continuees, and Medicare-eligible retirees with employer-provided Medicare Supplementary plans. Most of the diagnosis codes in Truven are ICD-9 codes, as the version of the database that we used is 2015, but the implementation of ICD-10 started in October 2015⁶⁴.

4.3.1.4. On Ethical Data Use Related to this Manuscript

The IBM[®] MarketScan[®] Research Databases (Formerly, Truven[®]) contain individual-level, de-identified, healthcare claims information from employers, health plans, hospitals, and Medicare and Medicaid programs. The data in Health Facts[®] are extracted directly from the EMR of hospitals with which Cerner has a data use agreement. Both IBM and Cerner Corporation have established Health Insurance Portability and Accountability Act-compliant operating policies to establish de-identification for IBM[®] MarketScan[®] Research Databases and Health Facts[®]. The use of IBM[®] MarketScan[®] Research Databases and Cerner Health Facts[®] mandates compliance with all vendor contractual obligations; of specific ethical relevance is the legally binding directive that no user of these data may attempt to re-identify the de-identified data. As an additional safeguard, at an institutional level, UTHealth researchers employing the IBM[®] MarketScan[®] Research Databases and Cerner Health Facts[®] for their studies are subject to oversight and approval by the Committee for the Protection of Human Subjects (UTHSC-H IRB) under protocol HSC-SBMI-13-0549. The use of the IBM[®] MarketScan[®] Research Databases and Cerner Health Facts[®] for this study is

covered by the approval by the Committee for the Protection of Human Subjects (UTHSC-H IRB) under protocol HSC-SBMI-13-0549.

4.3.1.5. Data Availability

The data that supports the findings of this study are available from the Data Service office at the University of Texas Health Science Center at Houston School of Biomedical Informatics (SBMI) but restrictions apply to the availability of these data, which were used under license from the data provider.

4.3.2. The Data Modality of Structured EHR

We define structured EHR data of each patient as a sequence of visits, each as a list of codes. This is a classic formulation commonly used in the literature[53-56]. The codes within a visit can be either ordered or unordered. If unordered, the EHR data for each patient can be reduced to a sequence of sets. The Med-BERT framework can handle both ordered and unordered codes inside a visit. In this paper, we have access to the priority of the diagnosis codes as coded by billers, e.g., the primary diagnosis is mostly assigned the first priority followed by the second most important diagnosis and so on, and thus we encode that information to introduce order.

Both structured EHR and natural language text are sequential data with tokens.

Therefore, the data modality of EHR is similar to text in many ways. However, EHR data has distinct characteristics (Figure 4.2). A direct comparison between the data modalities of the structured EHR data with the natural language text is shown in Table 4.2.

Table 4.9: Comparison of characteristics of EHR data versus Natural language data

Criteria	Natural language	EHR
<i>Token granularity</i>	The basic token is a word, which is a compressed semantic unit in language and can express some basic meaning. But in many cases, an integrated semantic unit (e.g., a named entity or a prepositional phrase) requires the combination of multiple tokens.	The basic token is a clinical code, which can represent an integrated semantic unit, e.g., a disease description, a drug, or a procedure.
<i>Syntactic: Hierarchical structure</i>	A paragraph (document) contains multiple sentences, and a sentence contains multiple words.	More complex, a patient's information contains multiple visits, and a visit contains multiple codes of different categories.
<i>Syntactic: Sequential order</i>	Simple and clear.	The visits are sorted sequentially according to time but the codes within a visit may be unordered or with certain prioritized orders.
<i>Semantic</i>	Dependency relations among sentences (e.g. discourse relations) as well as words within each sentence (e.g. syntactic dependency, semantic roles) are clear.	Dependency relationships are not always clear, e.g., adjacent visits may be of little relevance owing to large time intervals.
<i>Time interval</i>	Regular, one between adjacent words.	Usually no explicit intervals between codes, and irregular intervals between adjacent visits.
<i>Data completeness</i>	Relatively complete for regular texts such as written language.	Usually incomplete and sometimes erroneous due to the nature of EHR.

Criteria	Natural language	EHR
<i>Sequence length</i>	Within a relatively narrow range: The maximum sequence length of words in a sentence rarely reaches a hundred.	More variable: A patient's medical records can include anywhere from one to hundreds of visits. In a single visit, a patient can have hundreds of medical codes.

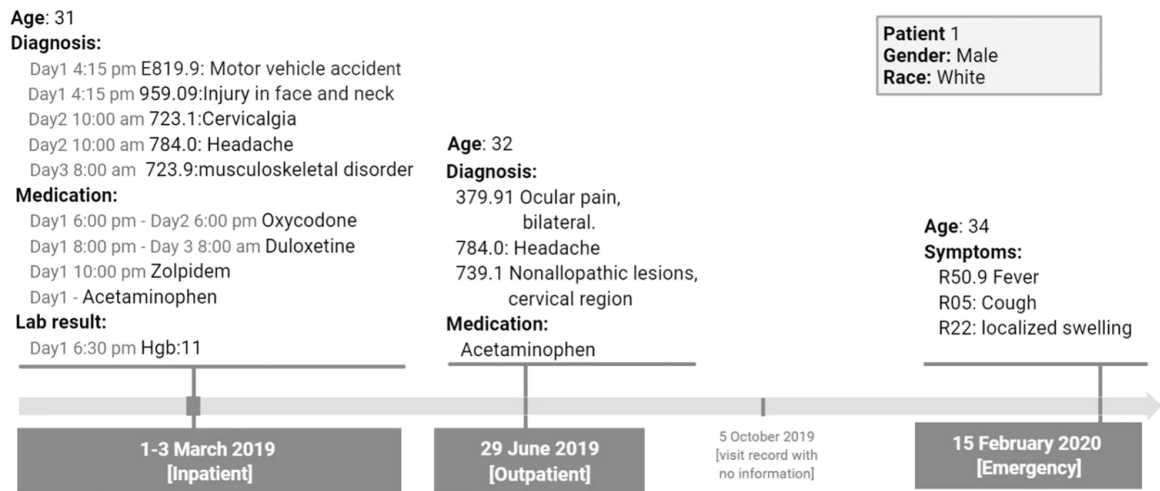


Figure 4.5. An example of structured EHR data of a hypothetical patient as it would be available from a current EHR system (e.g., Cerner or Truven)

4.3.3. Med-BERT Architecture

In this work, we utilized essentially the same transformer architecture as that in the original BERT paper[57], including multi-level embeddings and bidirectional transformers. We also adopted similar pre-training techniques (same loss function on masking and classification pre-training tasks). Still, given the semantic differences between EHR and text, adapting the BERT methodology to structured EHR is non-trivial. For example, while the input modality of the original BERT was a 1-D sequence of

words, our input modality is structured EHR which is recorded in a multilayer and multi-relational style. There are no clear rules on how to flatten the structured EHR into a 1-D sequence and how to encode the “structures” of the structured EHR in the BERT transformer architecture. In addition, it is unclear how to organize the EHR data efficiently to match the structured inputs of a pre-trained model such as BERT, and what are the appropriate domain-specific tasks for pre-training.

Figure 4.3 introduced our design of the Med-BERT embedding layers to accommodate the new modality. Specifically, three types of embeddings were taken as inputs for Med-BERT. These embeddings were projected from diagnosis codes, the order of codes within each visit, and the position of each visit and named, respectively, *code embeddings*, *serialization embeddings*, and *visit embeddings*. Code embeddings are the low-dimensional representations of each diagnosis code; serialization embeddings denote the relative order, in our case, the priority order, of each code in each visit; and visit embeddings are used to distinguish each visit in the sequence.

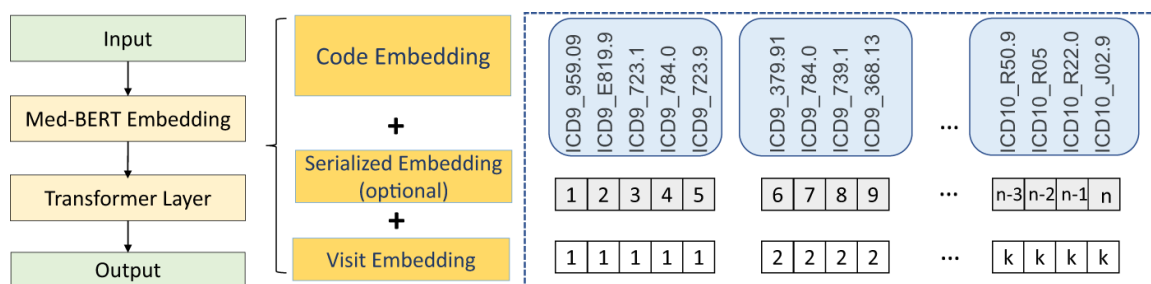


Figure 4.6. Med-BERT structure.

Unlike BERT, we did not use the specific tokens $[CLS]$ and $[SEP]$ at the input layer. Our choice is mainly due to the differences in the input formats of EHR and text. In BERT, only two adjacent sentences are fed for each input sample, and the token $[SEP]$ serves as a separator of the two sentences for the pre-training task of *next sentence prediction*. Next sentence prediction, however, was not involved in our tasks (as explained in the next subsection). We reasoned that the visit embeddings can separate well each visit and that adding $[SEP]$ would only be redundant. In BERT, the token $[CLS]$ was used mainly to summarize the information from the two sentences. However, EHR sequences are usually much longer; e.g., a sequence may contain 10 or more visits, and simply using one summarization token will inevitably lead to information loss. Therefore, for the classification tasks, either our prolonged LOS pre-training task or the downstream disease-prediction tasks, where the information of a long-range sequence is usually needed, we added a feed-forward layer (FFL) to the sum of the output from all of the codes within visits to represent a sequence, instead of using only a single token. Of course, it is also possible to use an RNN prediction layer instead of a simple FFL on top of Med-BERT.

4.3.4. Pre-training Med-BERT

We utilized the same optimization algorithm and recommended hyperparameters (*See Implementation Details*) of the original BERT model[57] during our Med-BERT pre-training phase. We trained the parameters of the Med-BERT model parameters on the diagnosis information of a cohort of 20 million patients using the following tasks.

4.3.4.1. Masked Language Model (Masked LM)

This task was directly inherited from the original BERT paper, which was used to predict the existence of any code, given its context. In detail, there was an 80% chance that a code was replaced by *[MASK]*, a 10% chance that the code was replaced by a random code, and another 10% chance that it was kept unchanged. This task is the core of the contextualized embedding model.

4.3.4.2. Prediction of Prolonged Length of Stay (Prolonged LOS) in Hospital

For the classification task, instead of using the question-answer pairs as in BERT, we decided to choose a clinical problem with a relatively high prevalence in our pre-training dataset and one that is not disease-specific to ensure better generalizability of our pre-trained model. The three most commonly used quality-of-care indicators, mortality, early readmission, and prolonged length of stay in hospital (LOS), were selected and tested. Through comparison, we found that the mortality and the early readmission tasks are relatively easy: the model quickly converges to >99% accuracy. Therefore, we chose prolonged LOS, the task of assessing each patient for whether an incident of prolonged hospital visit (LOS >7 days) had ever occurred throughout the entire EHR sequence of the patient, as a pre-training task. We used this simplified version of prolonged LOS prediction by targeting at the patient level rather than the visit level to reduce the pre-training complexity. Also, similar to the Masked LM task, we are not aiming to define a real future predicting task during the pre-training phase.

We found that the prolonged LOS task for pre-training leverages the bidirectional structure of Med-BERT. A prolonged LOS not only reflects the patient's health status

recorded in the past visits but also has an impact on the subsequent visits. On the other hand, tasks such as disease onset prediction or mortality always will be terminated at the last visit of the patient sequence, the input data of which can be constructed in only one direction.

4.3.5. Applying Med-BERT for Downstream Prediction Tasks by Fine-tuning

Med-BERT, similar to BERT, follows the pre-training-fine-tuning paradigm. The pre-trained model itself only generates contextualized embedding for each input token. The model outputs a general purpose embedding and does not directly output any prediction labels. For any specific downstream prediction task, a classification layer (prediction head) needs to be added on top of the Med-BERT model. One can use a simple prediction head such as FFL on top of the sequential output from the final Med-BERT layer. For EHR predictive models, a commonly used prediction head is the RNN rolling over the output of token embeddings.

During fine-tuning, following the original BERT, we attached a prediction head on top of the Med-BERT architecture. The parameters of the Med-BERT part were loaded and initialized from the pre-trained model, and then the parameters of both the Med-BERT part and the prediction head were updated by gradient descent. The input of the model was data from a disease-specific training cohort, which we referred to as the fine-tuning cohort. To understand the added values by the pre-trained Med-BERT (especially the usefulness of big training data), we compared the results of fine-tuning the pre-trained model and the un-trained model (same architecture with a randomly initialized token+segment+position embedding layers and the multi-head transformer layers). All

models were fine-tuned on a validation set (part of the fine-tuning cohort) and the reported numbers are the results on the test set.

4.3.6. Evaluation of Med-BERT

We conducted evaluations on two disease-prediction tasks on three cohorts from two databases. The two tasks are *heart failure in diabetes patients (DHF)* and *pancreatic cancer (PaCa)*. We used Cerner for both tasks, forming the *DHF-Cerner* and *PaCa-Cerner* cohort; and used Truven for only the pancreatic cancer prediction task, forming the *PaCa-Truven* cohort, for generalizability evaluation. The detailed cohort definitions are presented in the Methods section. Unlike BEHRT and G-BERT, whose evaluation tasks are simply the prediction of certain codes which are similar to the tasks in pre-training, our definition of disease prediction tasks is more complex, as it requires the phenotyping from multiple perspectives, e.g., the existence of certain diagnosis codes, drug prescriptions, procedures, laboratory test results, and, sometimes, the frequency of events in predefined time windows. Therefore, we claim that our evaluation tasks are more realistic (compared with BEHRT) and more helpful in establishing the generalizability of Med-BERT.

For all three tasks, we conducted three experiments: (1) Ex-1: to evaluate how Med-BERT can contribute to state-of-the-art methods; (2) Ex-2: to compare Med-BERT with one state-of-the-art static clinical word2vec-style embedding, t-W2V (trained on the full Cerner cohort)[52]; and (3) Ex-3: to investigate how much the pre-trained model can help in transfer learning with various training sample sizes.

For each fine-tuning task, we randomly selected a subset of the original cohort and further split it into training, validation, and testing sets with the ratio of 7:1:2. Since we have enough patients that are not included in the pre-training, we prioritized the assignment of samples to the test set to ensure that our test sets did not include any patient previously included in the Med-BERT pre-training set. For performance measurement, we used the area under the receiver operating characteristics curve (AUC) as our primary evaluation metric, which has been widely adopted by many previous studies of disease prediction[14 53 58]. Additional performance evaluation metrics are reported in Supplementary Table 1 and Supplementary Table 2.

For Ex-1, to evaluate the augmented power of pre-trained Med-BERT on top of state-of-the-art base models, we compare the performances of the base models only and the performance of the base models on top of Med-BERT. We use GRU[59], Bi-GRU[60], and RETAIN[53] as our base recurrent neural networks (RNN) models. While GRUs were shown to be very competitive baseline models, we also included RETAIN, a popular disease prediction model with double GRUs with attention. We also presented the results by using Med-BERT only; i.e., only FFL was added on top of the last layer of Med-BERT. This Med-BERT only model will provide an evaluation beyond RNN-based models. In addition, to evaluate the effect of pre-training using big data, we compare the performance of pre-trained Med-BERT with the untrained Med-BERT architecture. For the sake of completeness, we also included L2 regularized Logistic Regression (*L2LR*) and Random Forest (RF), two popular non-deep learning methods, using standard multi-hot input format, as baseline models.

For Ex-2, to compare Med-BERT against static embeddings, we chose the t-W2V model. Our decision to use t-W2V to represent non-contextualized static embeddings was based on a previous study[61] where different static embedding techniques including word2vec[24], fasttext[62], and pointwise positive mutual information-singular value decomposition (PPMI-SVD)[63] were compared and t-W2V was found to perform best in the evaluated disease prediction task. Notably, Glove[64] is a competent alternative of word2vec (w2c) for static EHR concept embedding but it was documented as having a comparable performance with w2c. Therefore, we selected t-W2V as our baseline for static embedding for the sake of convenience.

For Ex-3, to evaluate the value-added of Med-BERT with various fine-tuning training sizes, we selected samples with increasing sizes from the training data for each cohort for fine-tuning. Intuitively, the pre-trained model would be more helpful when the training size is smaller, as it helps inject a broader scope of knowledge.

For Ex1 and Ex2, where we used the full finetuning training cohorts, we reported the average AUC and standard deviation for each model, based on 10 runs with randomly initialized prediction head weights. For all iterations in Ex3, we conducted a random bootstrap sampling 10 times and reported the average AUC and standard deviation for each cohort.

4.3.7. Implementation Details

For the transformer architecture of Med-BERT, we used 6 layers, 6 attention heads, and a hidden dimension of 192 (L=6, H=192, A=6). We set the feed-forward/filter size to be 64.

For pre-training, we set the maximum sequence length as 512 tokens. We masked one diagnosis code per patient during Masked LM. We used the default BERT optimizer, AdamWeight decay optimizer. We used the recommended learning rate of $5e-5$, and a dropout rate of 0.1. We used the TensorFlow code of the original BERT from <https://github.com/google-research/bert> (February 2019 version). We used a single Nvidia Tesla V100 GPU of 32GB graphics memory capacity, and we trained the model for a week for more than 45 million steps, for which each step consists of 32 patients (batch size).

Before fine-tuning, we first converted the pre-trained model to the PyTorch version, using the HuggingFace package (version 2.3)[65]. For fine-tuning, we utilized our established codebase https://github.com/ZhiGroup/pytorch_ehr for the implementation of BERT_only, GRU, bi-GRU, and RETAIN models with minor modification to implement multi-layer embeddings instead of visit-level embeddings. We used the Adam optimizer and a learning rate of $1e-5$ for most of the models except for unidirectional GRU with static embedding for which a learning rate of 0.001 was associated with the best results. For the evaluation tasks, we used Nvidia GeForce RTX 2080 Ti GPUs of 12GB memory. For *L2LR* and RF, we used the scikit-learn package version 0.24. We used the default hyperparameters for both the logistic regression and the random forest classifiers.

4.3.8. Code Availability

To facilitate reproducibility and benefit other EHR-based studies, we shared our source code as well as our visualization tool on <https://github.com/ZhiGroup/Med-BERT>. The

pre-trained models are available from the authors upon request and with permission of the SBMI Data Service office.

4.4. Results

4.4.1. Data Source

We extracted our cohorts from two databases: Cerner Health Facts® (Cerner) and Truven Health MarketScan® (Truven). We defined one cohort for Med-BERT pre-training from Cerner and three phenotyped cohorts for fine-tuning, two of which were from Cerner (DHF-Cerner and PaCa-Cerner) and one from Truven (PaCa-Truven). The descriptive analysis of the cohorts is shown in Table 4.3. See Methods: Cohort definition for details.

Table 4.10. Descriptive analysis of the cohorts.

Characteristic	Pre-training	DHF-Cerner	PaCa-Cerner	PaCa-Truven
Cohort size (<i>n</i>)	28,490,650	672,647	29,405	42,721
Percent of Patients with the event ¹	15%	14%	0.07%	0.06%
Average Age on last/index encounter (<i>std</i>)	41	61	65	63
Gender - <i>Male (%)</i>	45%	47%	45%	48%
Race:				
White (%)	68%	72%	77%	NA
African American (%)	15%	16%	13%	
Asian / Pacific Islander (%)	2%	2%	2%	
African American (%)	2%	2%	1%	

Average number of visits per patient	8	17	7	19
Average number of codes per patient	15	33	14	18
Vocabulary size	82,603	26,427	13,071	7,002
ICD-10 codes (%)	33.8%	13.3%	20.7%	0%

¹The event for pretraining is a prolonged hospitalization >7 days. The event for PaCa-Cerner and PaCa-Truven is the diagnosis of pancreatic cancer, the event for DHF-Cerner is the development of heart failure for diabetic patients

4.4.2. Performance Boost of Med-BERT on Fine-tuning Tasks

Table 4.4 presents the AUCs for Ex-1 on the three fine-tuning evaluation tasks. The trends of additional performance evaluation metrics (Supplementary Table 1 and Supplementary Table 2) are largely consistent with that of AUC shown in Table 4.4 and Figure 4.4. For DHF-Cerner, it is notable that Bi-GRU+Med-BERT and RETAIN+Med-BERT obtain the best results and perform comparably, followed by Med-BERT_only and GRU+Med-BERT. For each base model, adding t-W2V (except GRU) will generally achieve better results, but adding Med-BERT improves the results much further. It is remarkable that those powerful deep-learning based models, such as GRU, Bi-GRU, and RETAIN that already obtain over 0.83 on AUC with relatively large training data, e.g., 50K samples, adding Med-BERT still makes a considerable performance boost.

Table 4.11. Average AUC values and standard deviations (in parentheses) for the different methods for the three evaluation tasks.

Model	DHF-Cerner	PaCa-Cerner	PaCa-Truven
<i>GRU</i>	83.93 (0.13)	78.26 (0.84)	78.17 (0.21)
<i>GRU+t-W2V</i>	83.95 (0.24)	80.08 (1)	77.54 (0.27)
<i>GRU+Med-BERT</i>	85.14 (0.06)	82.13 (0.24)	80.37 (0.12)
<i>Bi-GRU</i>	82.82 (0.17)	76.09 (0.61)	76.79 (0.29)
<i>Bi-GRU+t-W2V</i>	84.23 (0.06)	79.35 (0.27)	77.44 (0.22)
<i>Bi-GRU+Med-BERT</i>	85.39 (0.05)	82.23 (0.29)	80.57 (0.21)
<i>RETAIN</i>	83.28 (0.16)	79.68 (0.32)	78.02 (0.19)
<i>RETAIN+t-W2V</i>	84.98 (0.02)	81.8 (0.17)	79.46 (0.18)
<i>RETAIN+Med-BERT</i>	85.33 (0.09)	81.3 (0.55)	79.98 (0.17)
<i>Med-BERT_only (FFL)</i>	85.18 (0.12)	81.67 (0.31)	79.98 (0.26)
<i>untrained Med-BERT only</i>	82.76 (0.13)	75.16 (0.77)	75.9 (0.18)
<i>Logistic Regression (LR)¹</i>	81.01 (0)	79.94 (0)	77.28 (0)
<i>Random Forest (RF)¹</i>	81.88 (0.08)	79.48 (0.31)	77.00 (0.12)

¹LR and RF input is one hot representation while other models using embeddings.

For PaCa-Cerner, similar trends also were observed, whereby Bi-GRU+Med-BERT, Med-BERT_only, and GRU+Med-BERT generally outperform methods without Med-BERT and adding Med-BERT enhanced the AUCs of the base models by 1.62–6.14%. For PaCa-Truven, the best AUC was obtained by GRU+Med-BERT, whereas the other Med-BERT-related models also have better results than those without Med-BERT. On this Truven dataset, we still observe performance gains of 1.96–3.78%, although the average improved AUCs appear to be a bit lower than those on PaCa-Cerner.

Nevertheless, it is reassuring to see that Med-BERT can be generalized well to a different dataset whose data distributions might be quite different from Cerner, the one it was pre-trained on.

As an ablation experiment, we also made a comparison between the result of pre-trained Med-BERT and that of untrained Med-BERT, where “untrained” means we did not feed the model with large EHR for a self-supervised pre-training but only took advantage of its structure. Table 4.4 shows that untrained Med-BERT performs much worse than Med-BERT only and does not even outperform the baseline method of logistic regression for PaCa prediction tasks. Therefore, we can conclude that the pre-training phase plays a more important role for the boosted performance. Cases where untrained Med-BERT does not outperform the baseline logistic regression are likely due to overfitting, although we used the standard practice of both early stopping and dropout to reduce the likelihood of overfitting during the model training. This is possibly due to the fact that the untrained Med-BERT is an over-parameterized model (around 17 million parameters) with a huge

number of configurations, so it might overfit to the training data[66]. On the other hand, the pre-trained model started with a good configuration that is robust to a very large data set for the pre-training, and thus is likely to generalize well.

It is a standard practice that the pre-trained BERT model is not used on its own for prediction, rather a prediction head is needed for the fine-tuning tasks[57]. Since Med-BERT is an unsupervised pre-training model, fine-tuning should be done with certain configurations for different tasks, especially on the input data formats. However, in Table 4.4, we observed that a Med-BERT model with only an FFL on top of the last layer (*Med-BERT_only (FFL)*) can also obtain competitive performances.

In Figure 4.4 we show how much Med-BERT can help boost the prediction performance of the base deep-learning models by incorporating contextual information through pre-training. In the line chart of DHF-Cerner, we notice that, without Med-BERT, it is difficult for GRU only to have an AUC exceeding 0.65 when given fewer than 1,000 training samples. The addition of Med-BERT, however, greatly increases the AUCs by about 20% and helps the model to reach 0.75, even when training on 500 samples. For Bi-GRU, considerable improvements also can be observed, but they are not as high as those for GRU. For RETAIN, Med-BERT seems to be more helpful when the training set contains more than 500 samples.

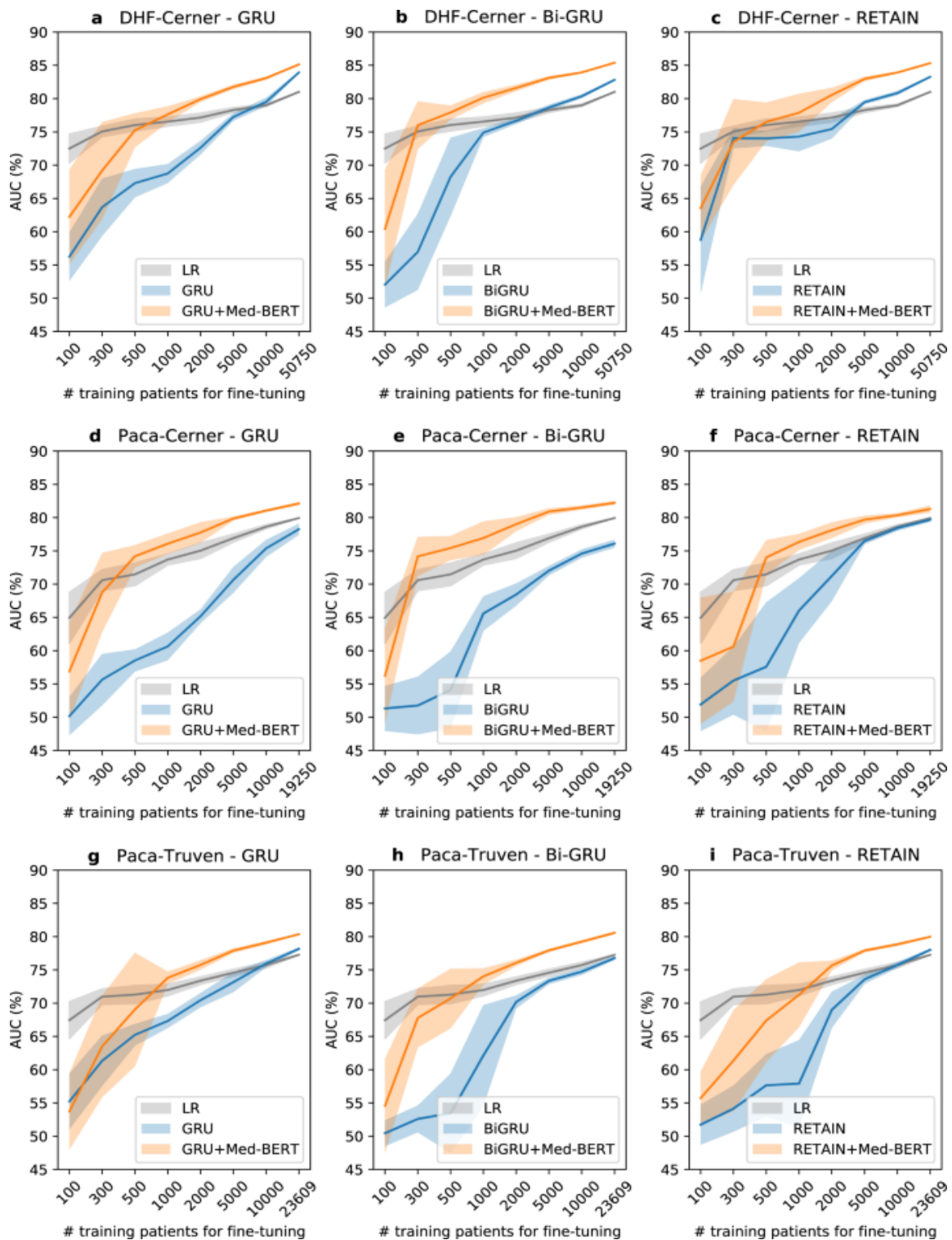


Figure 4.4. Comparison of prediction AUC for the test sets by training on different sizes of data on various Cohorts between the methods with or without the pre-trained Med-BERT layer. Logistic regression (LR) results are included as a baseline. (a) Cohort: DHF-

Cerner, Method: GRU; (b) Cohort: DHF-Cerner, Method: bidirectional GRU; (c) Cohort: DHF-Cerner, Method: RETAIN; (d) Cohort: PaCa-Cerner, Method: GRU; (e) Cohort: PaCa-Cerner, Method: bidirectional GRU; (f) Cohort: PaCa-Cerner, Method: RETAIN; (g) Cohort: PaCa-Truven, Method: GRU; (h) Cohort: PaCa-Truven, Method: bidirectional GRU; (i) Cohort: PaCa-Truven, Method: RETAIN. The shadows indicate the standard deviations.

For PaCa-Cerner, large improvements by adding Med-BERT to GRU and Bi-GRU were demonstrated for almost all training sizes. In particular, for Bi-GRU, Med-BERT enables the AUC to reach 0.75 when training on only 300 samples. The charts for PaCa-Truven show similar trends, but the overall AUC values are lower compared to those on PaCa-Cerner when training on smaller sample sizes.

Logistic regression, a popular non-DL machine learning algorithm, serves consistently as a competitive baseline model, especially on small datasets. Indeed, for smaller training sizes as 500 or less in our experiment, *L2LR* (L2 regularized logistic regression) showed decent performances. However, Med-BERT outperforms *L2LR* in all prediction tasks when the sample size is over 1,000.

4.4.3. Visualization of Attention Patterns in Med-BERT

Med-BERT not only offers improvement for prediction accuracy but also enables prediction interpretation. It is interesting and meaningful to explore how the pre-trained model has learned using the complex structure and a huge volume of data. We show several examples of how codes are connected with each other according to the attention weights from the transformer layers, the core component of Med-BERT.

The bertviz tool[66] was adapted and improved to better visualize the attention patterns in each layer of the pre-trained model. We added "SEP" tokens between visits only for visualization purposes. We observed distinct patterns in different layers of the model. In the pre-trained model, among the six layers of the BERT transformer model, the connections of the first two layers are mostly syntactic, some attention heads are restricted within a visit, and some point to the same codes across different visits. In the middle two layers, some medically meaningful attention patterns that capture contextual and visit-dependent information emerge. For the final couple of layers, the attention patterns become diffused and difficult to interpret.

Figure 4.5 is an example of the same code in different visits, showing different attention patterns. This demonstrates the ability of Med-BERT to learn contextualized representations. The earlier code for type 2 diabetes mellitus focuses mainly on the code for the long-term use of insulin within the same visit, but the later diabetes code focuses on the insulin code, both in the current and the previous visits. This could potentially indicate that the model learns the temporal relationship between visits through the segment embedding. More examples are provided in Supplementary Figure 3.

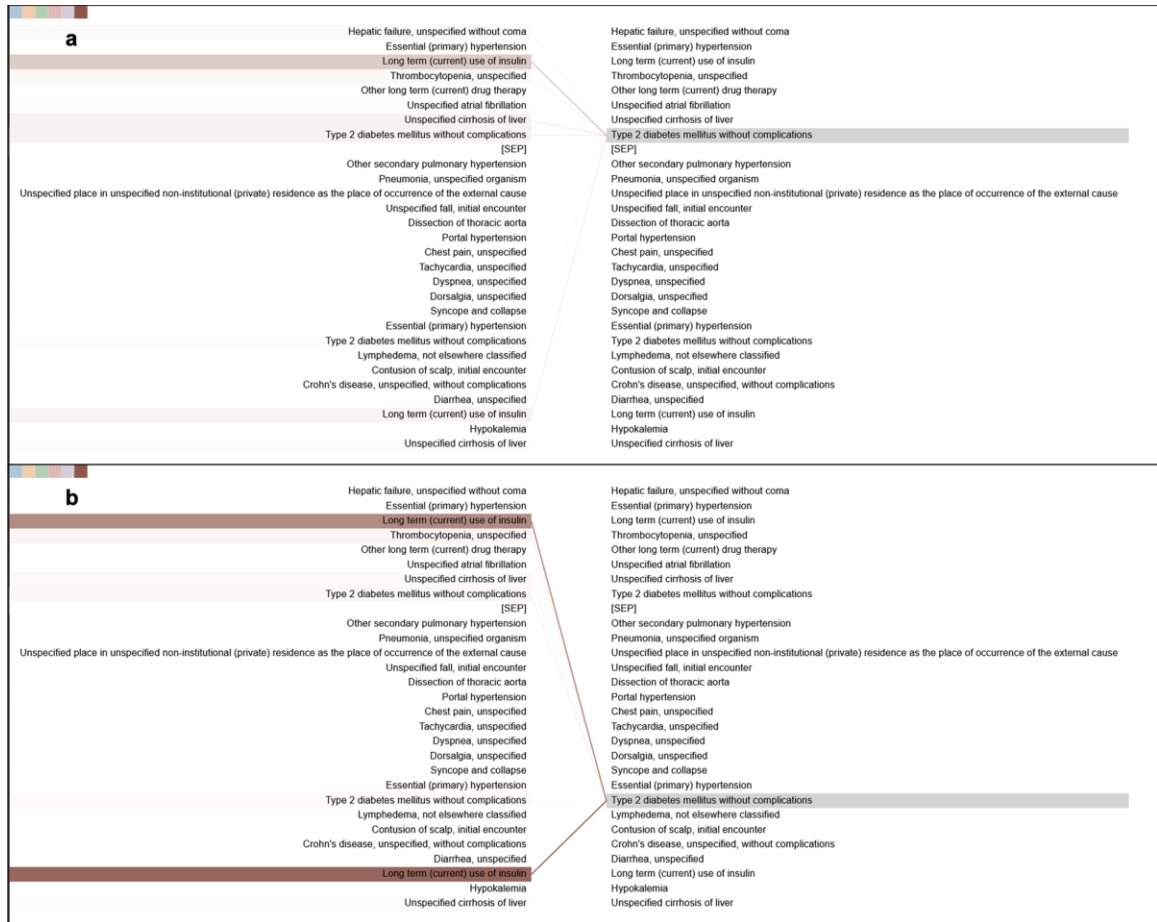


Figure 4.5. Example of different connections of the same code, “type 2 diabetes mellitus,” in different visits.

The attention patterns of the fine-tuned model are different. The fine-tuned models express distinct task-dependent patterns across different layers, showing the generalizability and adaptability of the model for learning different levels of knowledge in real-world scenarios. Figure 4.6 provides an example of the Med-BERT model fine-tuned on the DHF-Cerner dataset with attention converging onto several related codes in the second layer. Figure 4.7 is an example of the attention pattern in the fourth layer of the Med-BERT model fine-tuned on the PaCa-Cerner dataset, capturing the relevant

correlation between diagnosis codes. Additional visualization patterns can be seen in the Supplementary Figure 3. We believe that these kinds of visualization patterns can help us to better understand the inner mechanism of the neural network model and to build trusting and better communications of health information.



Figure 4.6. Example of the dependency connections in the DHF-Cerner cohort.



Figure 4.7. Example of the dependency connections in the PaCa-Cerner cohort.

4.5. Discussion

Med-BERT shows its power in helping to improve the prediction performance on multiple tasks with different configurations, and it is particularly effective in the “extreme transfer learning” paradigms, i.e., fine-tuning on only several hundreds of samples. Deep-learning-based predictive models usually require at least thousands of samples. These models need to learn complex semantics through feeding samples that convey different underlying disease progressions and variational context information so that they can be capable of dealing with intricate unseen cases. However, most deep-learning algorithms are insufficient in modeling the data comprehensively due to their limitation in an in-depth understanding of the inputs. Pre-trained models can well address this issue by using more sophisticated structures to better capture the complex semantics of inputs, behaving as a knowledge container, and injecting the knowledge into new tasks. Similar to pre-trained models on other domains, Med-BERT, by using its bidirectional transformer and deep structure as well as big data, also have been shown in this study to be extremely helpful when transferring to new tasks.

Masked LM and Prolonged LOS were designed and included to reinforce the modeling of contextual information and to help collect sequential dependencies. Labels for both can be generated in an unsupervised way, i.e., without human annotations. In Masked LM, the goal is to predict a masked code using the sequential information from the forward and the backward directions. In Prolonged LOS, the goal is to determine whether a patient is associated with any visit that is a prolonged stay, which also relies on cumulative contexts. We believe that, by including the prediction tasks from both the

code level and the patient (sequence) level, Med-BERT can further strengthen the representation learning of EHR sequences from different granularities.

Intuitively, a better parameter initialization of deep-learning models could lead to better performance and faster convergence. However, these benefits would gradually diminish with the growth of training samples. We consider 50K and 20K as acceptable scales of samples for training satisfactory (converging) deep-learning models. When we added Med-BERT, however, considerable improvements also could be observed. For example, RETAIN obtains satisfactory performances on all the three tasks, but adding Med-BERT brings further improvements by 1.62–2.05%. In addition, for GRU and Bi-GRU, whose model structures are simpler than that of RETAIN, the improvements can be much larger, which bring these simple models to a comparable level of or even better than RETAIN. Further, according to the results of Med-BERT_only, which also achieves good performance, we may conclude that Med-BERT will potentially release researchers from developing complex models for disease-prediction problems.

Similar to Med-BERT, static embedding method t-W2V also can serve as a good performance booster to the base deep-learning models. However, the improvements of t-W2V are smaller compared to Med-BERT in most cases. A probable explanation is that t-W2V has limitations in modeling long-sequential information, considering its shallow structure and the limited size of the context window which cannot be guaranteed to act well in all situations.

In practice, Med-BERT will significantly help to reduce the burden of data labeling, which can be seen through comparing the sizes of training samples required to achieve

certain AUC levels. Ex-3 proved the effectiveness of transferring Med-BERT into realistic disease-prediction tasks. Most of the charts in Figure 4.4 reflect that Med-BERT can substantially boost the performance of base models on small samples. For example, in the first sub-chart of PaCa-Cerner in Figure 4.4, if we draw a horizontal line across the y-tick of 0.75, we will see a requirement of 1,000 samples for GRU+Med-BERT and over 10,000 samples for GRU only. Similarly, we can see the Bi-GRU+Med-BERT trained on 5,000 samples can provide slightly better performance than Bi-GRU only trained on more than 50,000 samples as appears in Supplementary Table 2-A.

Thus, Med-BERT brought the model performance on par with a training set almost 10 times larger. The data acquisition cost of these over 9,000 samples, which sometimes can be quite expensive, will be substantially saved by using Med-BERT. In this situation, with Med-BERT, researchers and clinicians are able to quickly get a general and acceptable understanding of the progressions of new diseases before collecting enough annotated samples.

Admittedly, although Med-BERT empowers deep-learning models throughout all training sample sizes tested, Med-BERT powered models still do not outperform the non-deep learning baseline model logistic regression (LR) for the smallest training sample sizes ($n < 500$). This is consistent with the literature that LR remains a competitive predictive model for small training sample sizes in a number of studies[14]. LR benefits from its simple and shallow structure, which is much easier to fit based on even only a few samples compared with the complex structure and immense parameter space of deep-learning models. However, this advantage is gradually weakened as the training size

grows. Therefore, for practice, we would recommend the use of Med-BERT fine-tuning for the scenarios where the training sample size is sufficiently large (e.g. $n > 500$).

The vocabulary of the current version of Med-BERT is the union of ICD-9 and ICD-10 codes with 82,000 tokens. Compared with BEHRT and G-BERT, our vocabulary has broader coverage and is widely adopted in practice. We believe that it will greatly facilitate the transferability of the model, as ICD is a global health information standard recommended by the World Health Organization and is used by different institutions from over 100 countries around the world. This can be demonstrated in our PaCa-Truven evaluation, in which we tested our models' efficacy using a cohort extracted from a health insurance dataset.

In this work, we chose BERT, an advanced contextualized embedding methodology in NLP, for EHR modality. However, there are alternative ideas: such as ULMFiT[48], ELMo[49] GPTs[28 67 68], etc. It is probably necessary to evaluate these alternatives for pre-training and fine-tuning on EHR. We will leave it as future work.

There are still several limitations of the current work. First, we used only the diagnosis information in the ICD format. Second, we did not include the length of time intervals between visits in this study, which may cause some temporal information loss. Third, we did not fully explore the order of concepts within each visit, and the current setting based on code priorities might not be sufficiently reliable. In the future, more research on designing different pre-training tasks will be conducted, and different types of fine-tuning tasks beyond disease prediction also will be tested. We also plan to include other sources, such as time, medications, procedures, and laboratory tests, as inputs of Med-BERT. In

addition, task-specific visualizations and interpretations are other areas that we plan to explore.

In conclusion, we proposed Med-BERT, a contextualized embedding model pre-trained on a large volume of structured EHR data, and further evaluated the model in disease-prediction tasks. Domain-specific input formats and pre-trained tasks were designed.

Extensive experiments demonstrated that Med-BERT has the capacity to help boost the prediction performance of baseline deep-learning models on different sizes of training samples and can obtain promising results. The visualization module enabled us to look deeper into the underlying semantics of the data and working mechanisms of the model, in which we observed meaningful examples. Those examples were further verified by clinical experts, indicating that Med-BERT can capture the semantics among EHRs during both pre-training and fine-tuning. Methodologically, our work establishes the feasibility and usefulness of contextualized embedding of structured EHR data.

Practically, our pre-trained model enables training powerful deep learning predictive models with limited training sets.

4.6. Acknowledgments

We are grateful for our collaborators, David Aguilar, MD, Masayuki Nigo, MD, and Bijun S. Kannadath, MBBS, MS, for the helpful discussions on cohorts' definitions and results' evaluation. This research was undertaken with the assistance of resources and services from the School of Biomedical Informatics Data Service, which is supported in part by CPRIT Grant RP170668. Specifically, we would like to acknowledge the use of

Cerner HealthFacts® and the IBM Truven MarketScan™ datasets as well as the assistance provided by the UTHealth SBMI Data Service team to extract the data. The Nvidia GPU hardware is partly supported through Xiaoqian Jiang’s UT star award. We are also grateful to the NVIDIA Corporation for supporting our research by donating a Tesla GPU.

CT and DZ are supported by the American Heart Association under award number 19GPSGC35180031 and partly supported by the Cancer Prevention and Research Institute of Texas (CPRIT) Grant RP170668. LR is supported by UTHealth Innovation for Cancer Prevention Research Training Program Pre-Doctoral Fellowship (CPRIT Grant RP160015). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Cancer Prevention and Research Institute of Texas.

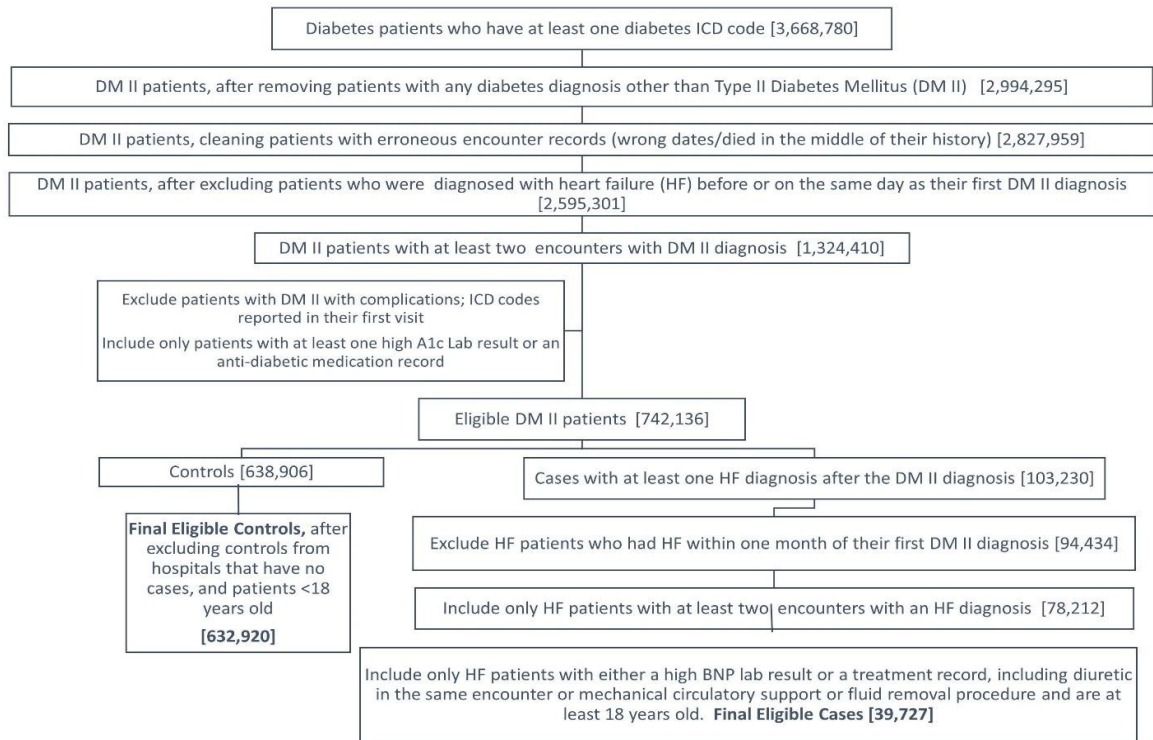
4.6.1. Competing Interests

The authors have no competing interests to declare.

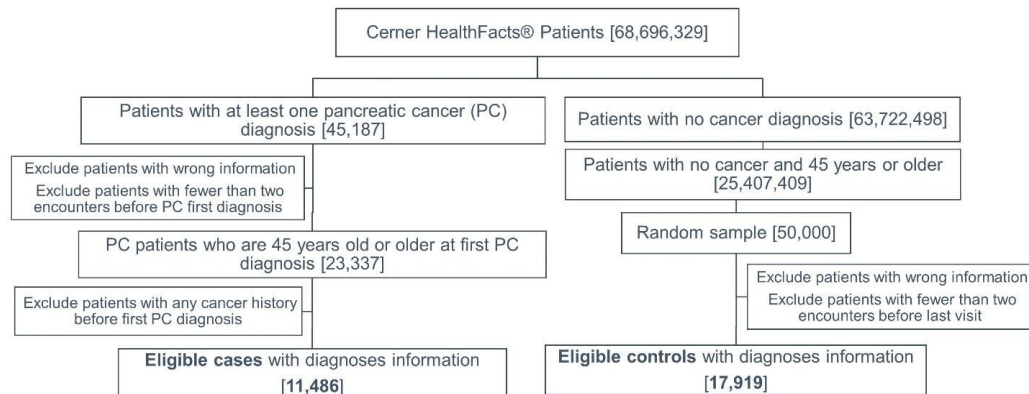
4.6.2. Authors Contributions

LR, YX, and ZX are co-first authors. DZ initialized the conceptualization of the project. LR, YX, ZX, and DZ designed the methods. LR led the implementation of the methods, with substantial inputs from YX and ZX. YX and DZ led the design of experiments. LR conducted the experiments and produced results. ZX led the visualization. YX, LR, and DZ led the writing, with substantial inputs from ZX and CT. YX, DZ, and CT supervised the execution of the project.

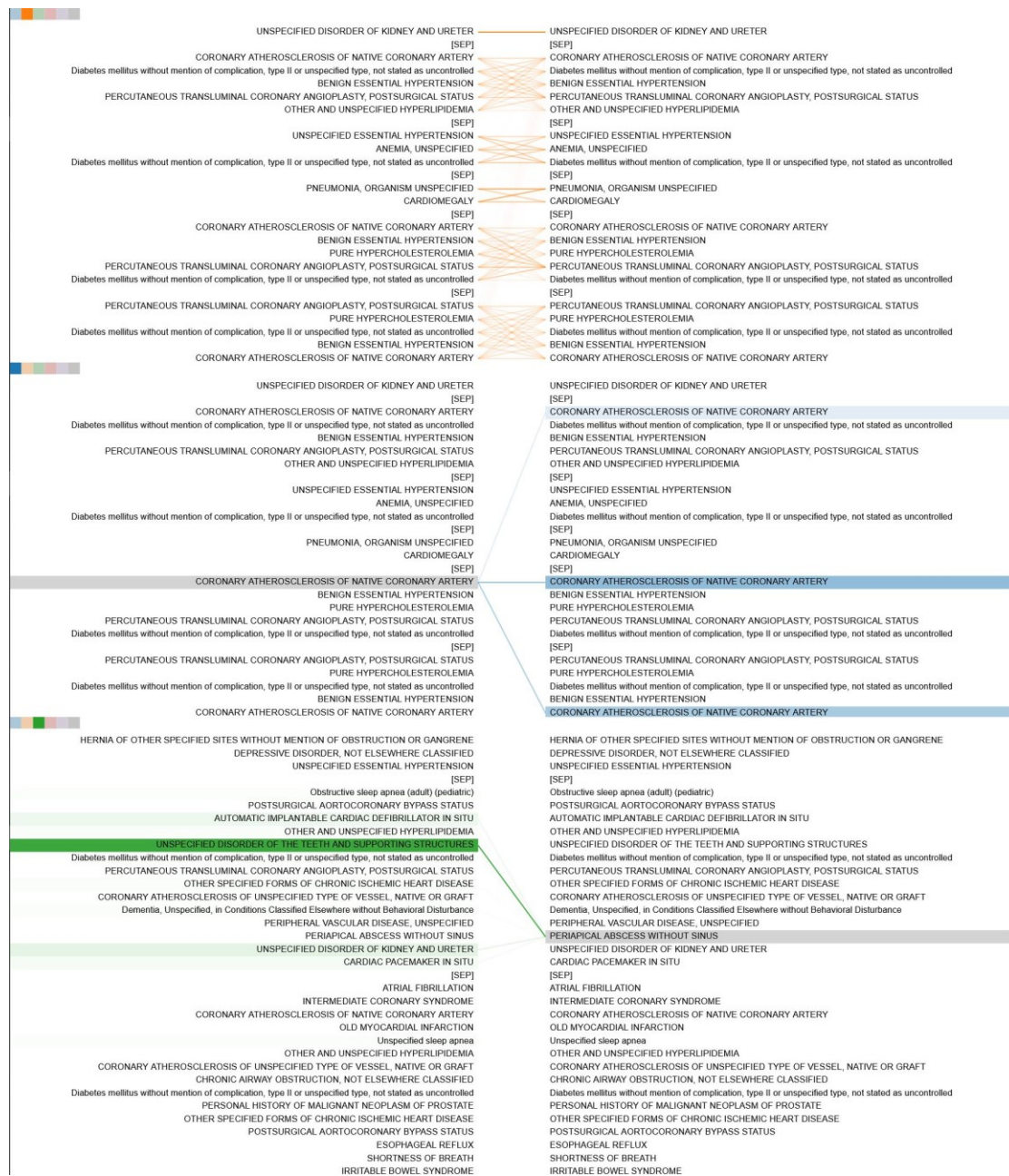
4.7. Supplementary material



Supplementary 4 Figure 6. Flowchart for the DHF cohort definition.



Supplementary 4 Figure 7. Flowchart for the PaCa cohort definition.



Supplementary 4 Figure 8. Attention connections from the first three transformer layers (a top-down direction) of a sample patient sequence. In the first layer, several heads show short-range attention patterns, and each token attends mainly to the nearby tokens that are within the same visit. In the second layer, some attention heads learn to make the correspondence between the same tokens. The third layer has the most interpretable patterns. A token in the third layer will focus strongly on other relevant tokens but mostly within the same visit. After the third layer, the attention becomes more diffuse and less explainable; however, there are still some heads that show long-range attention patterns.

Supplementary 4 Table 1: Additional benchmark results.

Supplementary 4 Table 1A: Average values and standard deviations (in parentheses) of additional evaluation metrics for DHF prediction

Model	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1- score
<i>GRU</i>	84.22 (0.09)	82.07 (0.19)	78.26 (4.37)	73.28 (4.35)	74.53 (2.16)	76.22 (1.1)
<i>GRU+t-W2V</i>	84.25 (0.15)	82.23 (0.26)	78.18 (3.41)	73.39 (3.28)	74.51 (1.5)	76.23 (0.85)
<i>GRU+Med-BERT</i>	85.29 (0.12)	83.44 (0.13)	81.25 (1.64)	72.37 (2.09)	74.47 (1.06)	77.69 (0.25)
<i>Bi-GRU</i>	83.04 (0.18)	81.32 (0.18)	77.07 (1.15)	72.72 (1.14)	73.68 (0.56)	75.33 (0.34)
<i>Bi-GRU+t-W2V</i>	84.59 (0.12)	82.66 (0.1)	80.60 (1.50)	71.70 (1.53)	73.85 (0.7)	77.06 (0.34)
<i>Bi-GRU+Med-BERT</i>	85.39 (0.07)	83.87 (0.05)	79.51 (2.50)	75.21 (2.53)	76.12 (1.3)	77.74 (0.56)
<i>RETAIN</i>	83.44 (0.25)	81.35 (0.16)	77.20 (0.62)	73.70 (0.63)	74.41 (0.33)	75.77 (0.21)
<i>RETAIN+t-W2V</i>	85.17 (0.06)	83.34 (0.05)	79.79 (0.78)	73.76 (0.85)	75.08 (0.43)	77.36 (0.16)
<i>RETAIN+Med-BERT</i>	85.36 (0.11)	83.63 (0.11)	78.07 (2.73)	76.09 (2.49)	76.44 (1.26)	77.2 (0.74)
<i>Med-BERT_only (FFL)</i>	85.25 (0.14)	83.67 (0.18)	78.09 (3.83)	75.87 (3.82)	76.35 (2.01)	77.12 (0.88)
<i>untrained Med-BERT only</i>	83.10 (0.22)	81.15 (0.17)	76.67 (2.73)	72.94 (2.37)	73.78 (1.03)	75.15 (0.81)
<i>Logistic Regression (LR)*</i>	81.22 (0)	78.52 (0)	77.12 (0)	70.83 (0)	72.36 (0)	74.66 (0)
<i>Random Forest (RF)*</i>	81.91 (0.35)	79.89 (0.17)	77.51 (0.11)	70.94 (0.43)	72.54 (0.28)	74.94 (0.15)

Supplementary 4 Table 1B: Average values and standard deviations (in parentheses) of additional evaluation metrics for PaCa prediction using Cerner cohort

Model	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1- score
<i>GRU</i>	81.63 (0.30)	71.40 (0.90)	58.42 (2.53)	83.54 (2.61)	62.76 (2.99)	60.41 (0.98)
<i>GRU+t-W2V</i>	83.41 (0.13)	72.18 (1.22)	58.27 (4.40)	84.41 (5.14)	64.53 (5.43)	60.92 (1.63)
<i>GRU+Med-BERT</i>	84.06 (0.21)	74.84 (0.19)	64.52 (1.71)	82.86 (1.63)	63.99 (1.61)	64.22 (0.4)
<i>Bi-GRU</i>	79.65 (0.56)	69.05 (0.56)	57.09 (1.67)	82.36 (1.27)	60.41 (1.22)	58.67 (0.74)
<i>Bi-GRU+t-W2V</i>	82.75 (0.12)	71.98 (0.35)	59.49 (1.36)	84.53 (1.02)	64.45 (1.08)	61.85 (0.44)
<i>Bi-GRU+Med-BERT</i>	84.32 (0.13)	75.08 (0.36)	63.82 (2.75)	83.59 (2.81)	64.89 (2.86)	64.25 (0.69)
<i>RETAIN</i>	80.99 (0.32)	72.02 (0.30)	52.78 (1.76)	88.35 (0.79)	68.1 (1.03)	59.45 (1.04)
<i>RETAIN+t-W2V</i>	84.60 (0.18)	74.88 (0.20)	61.98 (1.62)	85.86 (1.43)	67.42 (1.71)	64.55 (0.61)
<i>RETAIN+Med-BERT</i>	83.34 (0.13)	71.78 (3.10)	59.24 (9.26)	84.94 (6.43)	66.36 (6.51)	61.71 (2.43)
<i>Med-BERT_only (FFL)</i>	83.96 (0.23)	73.91 (0.53)	65.03 (4.80)	80.53 (5.27)	61.79 (5.04)	63.03 (0.43)
<i>untrained Med-BERT only</i>	79.56 (0.57)	67.97 (0.91)	54.81 (4.19)	82.01 (4.61)	59.47 (4.61)	56.78 (1.38)
<i>Logistic Regression (LR)</i>	79.45 (0)	73.59 (0)	56.82 (0)	89.49 (0)	71.79 (0)	63.43 (0)
<i>Random Forest (RF)</i>	79.05 (0.08)	65.65 (0.33)	63.83 (0.27)	79.96 (0.53)	59.99 (0.66)	61.85 (0.40)

Supplementary 4 Table 1C: Average values and standard deviations (in parentheses) of additional evaluation metrics for PaCa prediction using Truven cohort

Model	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1- score
<i>GRU</i>	77.31 (0.35)	68.93 (0.19)	49.54 (4.02)	88.06 (2.67)	68.2 (2.92)	57.2 (1.61)
<i>GRU+t-W2V</i>	77.19 (0.28)	67.90 (0.33)	43.30 (4.78)	90.98 (2.16)	71.35 (2.82)	53.64 (2.99)
<i>GRU+Med-BERT</i>	79.33 (0.17)	71.21 (0.28)	56.81 (2.00)	86.07 (1.32)	67.56 (1.31)	61.68 (0.66)
<i>Bi-GRU</i>	76.66 (0.21)	66.87 (0.32)	46.65 (0.81)	88.61 (0.63)	67.61 (1.05)	55.2 (0.58)
<i>Bi-GRU+t-W2V</i>	77.21 (0.29)	67.35 (0.27)	46.63 (1.81)	89.38 (1.07)	69.17 (1.36)	55.67 (0.93)
<i>Bi-GRU+Med-BERT</i>	79.45 (0.22)	71.54 (0.45)	56.80 (1.50)	86.02 (0.96)	67.45 (0.99)	61.65 (0.54)
<i>RETAIN</i>	77.80 (0.20)	68.93 (0.35)	45.74 (0.63)	90.24 (0.54)	70.5 (0.96)	55.48 (0.4)
<i>RETAIN+t-W2V</i>	79.58 (0.29)	70.36 (0.34)	51.41 (0.90)	88.76 (0.48)	69.99 (0.61)	59.27 (0.49)
<i>RETAIN+Med-BERT</i>	79.20 (0.16)	69.39 (0.97)	33.43 (8.33)	95.49 (2.83)	80.56 (5.16)	46.4 (7.02)
<i>Med-BERT_only (FFL)</i>	79.26 (0.19)	71.16 (0.59)	50.68 (5.43)	88.78 (3.33)	70.33 (4.23)	58.55 (2.45)
<i>untrained Med-BERT only</i>	75.89 (0.49)	65.56 (0.52)	47.75 (5.50)	86.23 (3.68)	64.34 (3.16)	54.5 (2.26)
<i>Logistic Regression (LR)</i>	77.11 (0)	67.33 (0)	45.52 (0)	89.17 (0)	68.16 (0)	54.58 (0)
<i>Random Forest (RF)</i>	76.36 (0.07)	64.62 (0.23)	43.28 (0.34)	89.71 (0.17)	68.17 (0.52)	52.94 (0.41)

Supplementary 4 Table 2: Additional performance results

Supplementary 4 Table 2A. Experiment 3 - Additional Metrics for DHF prediction evaluation using smaller training set size

Threshold used for Sensitivity, Specificity, Precision and F1-score is 0.5

Model	Training Set size	TEST AUC	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1-score
GRU	100	56.24 (3.65)	65.52 (24.07)	54.78 (3.42)	58.08 (8.72)	51 (10.08)	54.18 (3.31)	55.74 (4.16)
	200	62.31 (3.87)	70.79 (12.39)	60.73 (3.36)	60.62 (6.09)	57.75 (7.07)	58.83 (3.06)	59.55 (3.48)
	300	63.67 (4.32)	69.79 (9.37)	61.67 (3.83)	61.94 (6.07)	58.26 (7.56)	59.7 (3.63)	60.62 (3.66)
	400	64.34 (4.26)	68.52 (5.67)	62.26 (4.25)	58.77 (4.52)	62.09 (6.13)	60.7 (3.41)	59.62 (3.19)
	500	67.29 (2.13)	70.88 (6.25)	65.27 (1.92)	61.1 (5.71)	64.22 (5.22)	62.95 (2.15)	61.85 (2.97)
	1000	68.74 (1.43)	70 (4.82)	66.53 (1.54)	64.18 (6)	62.82 (5.89)	63.26 (1.92)	63.52 (2.46)
	2000	72.62 (1.03)	74.1 (4.03)	70.27 (0.96)	67.89 (2.06)	65.49 (3.28)	66.14 (1.45)	66.97 (0.66)
	5000	77.24 (0.58)	78.23 (1.07)	74.48 (0.5)	71.5 (5.88)	69.69 (4.72)	70.17 (1.55)	70.66 (2.16)
	10000	79.51 (0.63)	80.11 (1.32)	76.92 (0.85)	73.05 (4.94)	72.17 (4.39)	72.37 (1.82)	72.57 (1.64)
	Full Cohort (50750)	83.93 (0.13)	84.22 (0.09)	82.07 (0.19)	78.26 (4.37)	73.28 (4.35)	74.53 (2.16)	76.22 (1.1)
GRU+Med-BERT	100	62.21 (7.12)	71.1 (14.17)	60 (6.99)	65.58 (39.99)	42.27 (38.69)	50.43 (19.05)	50.45 (27.67)
	200	68.19 (7.47)	78.01 (9.82)	65.62 (6.79)	28.21 (33.51)	86.43 (16.3)	57.96 (21.78)	29.7 (32.39)
	300	69.18 (7.3)	73.69 (10.49)	65.67 (7.29)	75.08 (30.49)	44.23 (38.81)	59.25 (9.88)	61.28 (21.39)
	400	72.76 (3.97)	78.67 (6.02)	69.72 (4.35)	47.73 (31.95)	77.73 (17.05)	69.57 (5.55)	48.87 (29.09)
	500	75.24 (2.56)	79.11 (4.91)	71.96 (2.29)	73.76 (10.13)	62.18 (17.56)	67.2 (6.19)	69.53 (3.22)
	1000	77.63 (1.22)	76.6 (3.87)	74.55 (1.13)	74.93 (2.82)	67.66 (3.19)	69.71 (1.62)	72.18 (1.21)
	2000	79.92 (0.42)	80.43 (2.74)	77.51 (0.55)	75.31 (2.95)	70.37 (3.41)	71.65 (1.61)	73.37 (0.69)
	5000	81.76 (0.34)	81.92 (1.42)	79.36 (0.52)	78.61 (2.49)	69.76 (2.45)	72.07 (1.04)	75.16 (0.66)

Model	Training Set size	TEST AUC	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1-score
<i>Bi-GRU</i>	10000	83.08 (0.24)	83.5 (1.22)	81.01 (0.31)	79.32 (1.9)	71.16 (2.28)	73.18 (1.07)	76.1 (0.37)
	Full Cohort (50750)	85.14 (0.06)	85.29 (0.12)	83.44 (0.13)	81.25 (1.64)	72.37 (2.09)	74.47 (1.06)	77.69 (0.25)
	100	52.04 (3.44)	54.83 (21.76)	51.73 (2.92)	54.34 (27.3)	48.41 (26.9)	51.24 (2.44)	49.02 (16.09)
	200	58.24 (3.43)	67.59 (7.24)	56.43 (2.67)	51.89 (16.31)	59.46 (12.33)	55.78 (1.97)	52.44 (10.56)
	300	56.96 (5.72)	61.76 (14.29)	55.36 (5.21)	59.42 (17.69)	50.13 (18.55)	54.54 (3.91)	55.35 (10.09)
	400	65.03 (4.94)	70.22 (10.63)	61.92 (4.49)	61.26 (6.88)	60.18 (7.57)	60.56 (3.68)	60.69 (4.08)
	500	68.24 (5.87)	70.51 (8.77)	65.39 (5.36)	61.76 (13.31)	64.48 (5.31)	62.86 (3.76)	61.82 (8.77)
	1000	74.89 (0.67)	72.33 (3.93)	72.7 (1.05)	70.2 (2.91)	66.32 (2.54)	67.4 (0.81)	68.73 (1.08)
	2000	76.66 (0.46)	78.54 (3.89)	74.75 (0.5)	70.79 (1.69)	69.02 (1.91)	69.38 (0.93)	70.06 (0.67)
	5000	78.66 (0.41)	79.71 (1.14)	77.03 (0.53)	73.06 (1.9)	69.58 (1.99)	70.43 (0.86)	71.7 (0.57)
	10000	80.36 (0.31)	80.72 (1.31)	78.68 (0.33)	74.58 (1.67)	71.15 (1.28)	71.92 (0.52)	73.21 (0.67)
	Full Cohort (50750)	82.82 (0.17)	83.04 (0.18)	81.32 (0.18)	77.07 (1.15)	72.72 (1.14)	73.68 (0.56)	75.33 (0.34)
	100	60.44 (8.88)	59.43 (17.91)	57.24 (7.71)	64.09 (42.77)	44.24 (39.98)	50.74 (10.71)	49.12 (28.56)
	200	74.44 (3.09)	81.64 (10.38)	71.26 (3.61)	59.09 (28.99)	72.01 (16.31)	69.44 (4.27)	58.15 (24.11)
	300	76 (3.59)	77.19 (8.64)	72.97 (4.21)	66.84 (24.67)	70.03 (13.24)	72.3 (10.12)	64.17 (22.67)
	400	75.85 (3.5)	79.73 (5.3)	73.07 (4.12)	62.58 (33.76)	67.1 (26.57)	68.49 (9.81)	57.27 (29.63)
	500	77.92 (1.04)	79.48 (6.68)	74.99 (0.9)	69.65 (17.45)	70.91 (8.77)	71.15 (3.49)	68.5 (12.69)
	1000	80.14 (0.85)	78.99 (4.03)	77.88 (0.72)	76.57 (2.43)	68.92 (3.37)	71.01 (1.69)	73.64 (0.65)
	2000	81.61 (0.49)	82.02 (2.46)	79.55 (0.71)	74.26 (2)	73.47 (1.95)	73.52 (1)	73.86 (0.71)
	5000	83.12 (0.27)	83.37 (1.54)	81.21 (0.3)	77.17 (1.91)	73.05 (1.59)	73.95 (0.72)	75.51 (0.64)
	10000	83.94 (0.16)	84.15 (1.06)	82.17 (0.16)	78.79 (2.5)	72.78 (2.31)	74.18 (1.05)	76.38 (0.68)
<i>Bi-GRU+Med-BERT</i>								

Model	Training Set size	TEST AUC	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1-score
	Full Cohort (50750)	85.39 (0.05)	85.39 (0.07)	83.87 (0.05)	79.51 (2.5)	75.21 (2.53)	76.12 (1.3)	77.74 (0.56)
RETAIN	100	58.75 (8.01)	70.28 (20.26)	56.13 (6.8)	56.67 (27.96)	54.51 (23.81)	55.69 (8.54)	52.95 (14.8)
	200	66.94 (6.04)	75.83 (12.86)	63.27 (6.06)	51.74 (21.84)	69.76 (17.61)	64.55 (6.8)	54.46 (13.97)
	300	74.04 (1.53)	79.36 (8.61)	71.23 (2.23)	46.6 (23.14)	80.09 (13.58)	72.55 (6.01)	52.65 (16.74)
	400	71.88 (4.19)	72.93 (7.49)	68.91 (4.78)	55.42 (22.33)	72.34 (15.15)	67.92 (6.11)	58.04 (14.28)
	500	74.02 (1.13)	79.37 (6.39)	70.96 (1.85)	56.28 (25.6)	73.02 (17.42)	70.51 (6.8)	58 (16.37)
	1000	74.29 (2.21)	76.04 (1.91)	71.59 (2.62)	61.94 (13.33)	72.31 (9.68)	69.64 (4.13)	64.52 (7)
	2000	75.43 (1.44)	76.16 (2.68)	72.88 (1.7)	60.05 (15.83)	74.65 (9.82)	70.98 (3.36)	63.55 (9.58)
	5000	79.47 (0.33)	79.39 (1.29)	77.41 (0.39)	74.75 (2.28)	69.52 (2.28)	70.87 (0.95)	72.73 (0.71)
	10000	80.84 (0.32)	81.27 (1.18)	78.92 (0.36)	75.45 (1.39)	71.29 (1.6)	72.26 (0.77)	73.81 (0.38)
	Full Cohort (50750)	83.28 (0.16)	83.44 (0.25)	81.35 (0.16)	77.2 (0.62)	73.7 (0.63)	74.41 (0.33)	75.77 (0.21)
RETAIN+Med-BERT	100	63.53 (4.91)	80.16 (13.14)	60.44 (5.45)	47.29 (49.01)	56.69 (47.55)	45.35 (25.99)	34.78 (33.34)
	200	70.89 (4.24)	81.44 (9.7)	67.6 (4.58)	62.34 (36.47)	59.18 (30.44)	58.12 (22.43)	54.44 (25.57)
	300	73.49 (6.46)	77.22 (7.27)	70.5 (6.49)	43.83 (35.27)	76.91 (22.42)	71.76 (10.6)	43.93 (29.78)
	400	76.37 (2.96)	78.24 (5.23)	72.99 (3.94)	73.89 (11.69)	63.67 (19.74)	68.31 (6.31)	70.1 (3.43)
	500	76.5 (2.91)	79.55 (5.74)	73.35 (3.51)	41.1 (24.07)	84.79 (13.26)	76.61 (7.25)	48.6 (19.6)
	1000	77.88 (2.85)	79.37 (3.6)	74.03 (3.41)	72.95 (7.46)	68.79 (10.58)	70.5 (4.7)	71.28 (2.26)
	2000	80.55 (1.09)	80.56 (2.06)	77.29 (1.83)	68.87 (6.22)	76.38 (4.51)	74.48 (2.19)	71.34 (2.83)
	5000	82.96 (0.35)	83.41 (1.15)	80.63 (0.59)	76.35 (2.31)	73.97 (2.01)	74.42 (0.94)	75.34 (0.72)
	10000	83.92 (0.16)	84.06 (0.86)	81.95 (0.22)	77.74 (2.56)	73.82 (2.4)	74.67 (1.15)	76.14 (0.67)
	Full Cohort (50750)	85.33 (0.09)	85.36 (0.11)	83.63 (0.11)	78.07 (2.73)	76.09 (2.49)	76.44 (1.26)	77.2 (0.74)

Model	Training Set size	TEST AUC	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1-score
<i>Logistic Regression (LR) *</i>	100	72.49 (2.16)	68.75 (16.44)	70.43 (2.09)	71.57 (9.7)	61.54 (9.25)	65.13 (2.52)	67.78 (3.81)
	200	74.05 (0.44)	76.06 (9.62)	71.6 (0.45)	70.3 (4.29)	65.37 (3.62)	66.85 (1.08)	68.44 (1.63)
	300	75.06 (0.78)	76.01 (8)	72.91 (0.82)	70.69 (3.37)	66.6 (2.97)	67.75 (1.16)	69.13 (1.35)
	400	75.68 (0.76)	74.33 (8.35)	73.4 (1.09)	69.59 (3.57)	68.7 (3.77)	68.86 (1.55)	69.15 (1.22)
	500	76.03 (0.89)	76.21 (6.15)	73.48 (1.06)	72.03 (2.76)	66.84 (3.19)	68.33 (1.49)	70.08 (0.9)
	1000	76.54 (0.68)	75.51 (3.46)	74.06 (0.65)	71.4 (2.06)	68.1 (2.49)	68.95 (1.23)	70.13 (0.88)
	2000	77.15 (0.68)	77.93 (3.63)	74.67 (0.85)	71.4 (2.06)	69.33 (1.36)	69.75 (0.58)	70.55 (0.99)
	5000	78.28 (0.37)	78.4 (2.16)	75.93 (0.44)	72.47 (1.27)	70.29 (0.71)	70.73 (0.35)	71.58 (0.66)
	10000	79 (0.18)	79.33 (0.93)	76.47 (0.23)	74.06 (0.74)	69.84 (0.61)	70.87 (0.31)	72.43 (0.34)
	Full Cohort (50750)	81.01 (0)	81.22 (0)	78.52 (0)	77.12 (0)	70.83 (0)	72.36 (0)	74.66 (0)

Supplementary 4 Table 2B. Experiment 3 - Additional Metrics for PaCa-Cerner prediction evaluation using smaller training set size

Threshold used for Sensitivity, Specificity, Precision and F1-score is 0.5

Model	Training Set size	TEST AUC	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1-score
<i>GRU</i>	100	50.16 (2.88)	73.2 (13.19)	34.47 (3.26)	22.76 (16.48)	80.74 (15.01)	35.63 (3.66)	24.38 (13.91)
		56.59 (3.64)	70.43 (10.56)	41.46 (4.39)	42.49 (14.27)	69.23 (13.15)	40.25 (4.61)	39.52 (9.9)
	200	55.64 (3.86)	68.85 (5.42)	40.88 (4.89)	36.48 (15.51)	73.89 (12.78)	40.21 (3.58)	35.9 (11.8)
		57.79 (3.06)	67.77 (10.9)	42.56 (3.81)	42.36 (9.67)	70.26 (7.94)	40.39 (3.56)	40.75 (5.47)
	400	58.53 (1.67)	67.04 (6.47)	43.04 (1.88)	43.61 (8.31)	70.02 (8.05)	41.09 (3.02)	41.73 (4.15)
	500							

Model	Training Set size	TEST AUC	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1-score
<i>GRU+Med-BERT</i>	1000	60.64 (2.07)	69.24 (3.59)	47.2 (1.9)	44.2 (5.59)	72.33 (4.88)	43.09 (2.99)	43.44 (3.24)
	2000	65.17 (1.04)	70.89 (4.31)	53.54 (1.69)	52.45 (4.18)	70.37 (4.55)	45.64 (2.29)	48.66 (1.56)
	5000	70.64 (1.94)	75.55 (2.33)	61.86 (2.28)	52.2 (7.42)	78.61 (4.53)	53.75 (2.23)	52.59 (3.18)
	10000	75.38 (1.28)	79.33 (1.27)	68.21 (1.32)	54.4 (5.91)	83.39 (4.7)	61.31 (4.67)	57.26 (1.94)
	Full Cohort (19250)	78.26 (0.84)	81.63 (0.3)	71.4 (0.9)	58.42 (2.53)	83.54 (2.61)	62.76 (2.99)	60.41 (0.98)
	100	56.89 (7.3)	73.78 (17.02)	36.65 (5.53)	23.8 (40.9)	76.64 (41.1)	24.95 (19.48)	14.6 (20.13)
	200	66.67 (4.51)	72.71 (8.9)	48.29 (6.68)	13.03 (31.15)	89.49 (31.11)	36.05 (35.04)	9.83 (17.92)
	300	68.74 (5.94)	77.36 (11.94)	36.25 (11.94)	73.91 (41.26)	50.07 (39.85)	27.08 (34.92)	27.08 (23.76)
	400	71.99 (6.7)	73.58 (7.04)	57.57 (9.93)	47.39 (26.63)	77.5 (28.67)	53.21 (23.31)	45.49 (17.36)
	500	74.19 (1.69)	77.79 (7.43)	61.4 (2.96)	38.51 (22.09)	87.06 (13.06)	68.65 (15.93)	42.68 (17.01)
	1000	76.08 (1.54)	80.11 (3.93)	63.46 (2.8)	52.34 (24.16)	79.71 (11.67)	52.45 (21.28)	49.45 (20.05)
	2000	77.76 (1.55)	79.7 (2.68)	65.51 (2.43)	52.77 (12.82)	83.89 (7.47)	62.4 (6.07)	55.69 (6.15)
	5000	79.86 (0.31)	82.5 (1.91)	69.78 (1.24)	65.88 (3.34)	77.67 (3.14)	58.31 (2.33)	61.75 (0.43)
	10000	81.05 (0.22)	82.92 (0.9)	72.81 (0.51)	66.95 (2.73)	78.31 (2.99)	59.39 (2.3)	62.86 (0.51)
<i>Bi-GRU</i>	Full Cohort (19250)	82.13 (0.24)	84.06 (0.21)	74.84 (0.19)	64.52 (1.71)	82.86 (1.63)	63.99 (1.61)	64.22 (0.4)
	100	51.31 (3.39)	54.54 (11.33)	33.88 (2.3)	41.8 (21.32)	61.12 (21.84)	33.95 (3.15)	35.37 (7.93)
	200	49.78 (4.08)	52.19 (12.94)	32.78 (3.08)	40.44 (23.04)	58.7 (26.69)	33.03 (4.04)	32.65 (9.77)
	300	51.75 (4.33)	60.1 (10.63)	34.74 (4.02)	24.23 (23.16)	78.98 (24.35)	37.52 (7.09)	24.15 (13.48)
	400	54.86 (4.4)	59.55 (11.64)	37.68 (5.29)	42.58 (20.52)	64.1 (23.65)	38.23 (6.39)	37.12 (8.09)
	500	54.1 (5.77)	60.01 (10.33)	37.57 (6.25)	19.83 (18.7)	85.85 (16.72)	42.52 (7.44)	22.26 (15.59)
	1000	65.58 (2.59)	69.59 (5.23)	52.48 (3.45)	46.46 (3.28)	76.44 (4.68)	48.54 (4.67)	47.34 (2.95)
	2000	68.44 (1.67)	73.07 (3.63)	57.15 (3.09)	46.93 (8.5)	79.76 (6.1)	52.97 (4.24)	49.04 (4.68)

Model	Training Set size	TEST AUC	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1-score
	5000	72.03	76.16	62.94	52.77	79.65	55.09	53.82
		(0.7)	(2.57)	(1.1)	(2.98)	(2.57)	(2.04)	(1.25)
		74.6		66.74	56.76	80.07	57.29	57
	10000	(0.63)	77 (0.92)	(0.67)	(2.16)	(1.2)	(0.81)	(0.96)
		76.09	79.65	69.05	57.09	82.36	60.41	58.67
		(0.61)	(0.56)	(0.56)	(1.67)	(1.27)	(1.22)	(0.74)
<i>Bi-GRU+Med-BERT</i>	Full Cohort (19250)	56.23	75.83	37.31	12.45	88.31	28.32	8.58
		(7.13)	(18.61)	(6.72)	(31.06)	(31.19)	(33.26)	(16.26)
	100	66.49	78.29	50.56	9.5	96.34	42.73	12.85
		(6.99)	(9.58)	(10.61)	(14.68)	(7.03)	(41.03)	(17.66)
	200	74.17	80.37	61.75	32.64	92.09	73.95	40.4
		(2.94)	(9.27)	(3.8)	(18.7)	(7.44)	(13.84)	(17.78)
	300	75.29	78.57	63.06	34.76	90.4	64.98	39.38
		(3.15)	(4.85)	(4.57)	(24.46)	(9.61)	(26.96)	(22.95)
	400	75.43	77.77	63.12	45.7	84.81	65.31	49.33
		(1.82)	(7.21)	(3.89)	(18.08)	(11.31)	(14.59)	(11.35)
	500	76.93		65.83	45.97	84.96	69.28	46.27
		(2.51)	80.94 (4)	(3.68)	(26.37)	(10.88)	(18.32)	(23.29)
	1000	79.02	80.66	69.39	60.13	80.52	60.77	59.58
		(1.06)	(2.34)	(1.03)	(8.07)	(7.71)	(7.26)	(2.06)
	2000	80.92	83.32	72.37	64.38	80.29	60.74	62.43
		(0.48)	(1.55)	(0.53)	(2.54)	(2.72)	(2.26)	(0.68)
	5000	81.52	83.35	73.89	63.55	82.63	63.39	63.4
		(0.32)	(0.86)	(0.25)	(2.29)	(2.26)	(2.24)	(0.52)
	10000	82.23	84.32	75.08	63.82	83.59	64.89	64.25
		(0.29)	(0.13)	(0.36)	(2.75)	(2.81)	(2.86)	(0.69)
<i>RETAIN</i>	100	51.91	63.7	35.79	49.97	52.57	33.03	39.64
		(4.01)	(21.91)	(3.3)	(8.07)	(5.29)	(2.55)	(4.24)
	200	51.36	55.23	35.52	41.52	62.78	34.55	36.14
		(6.91)	(6.35)	(4.81)	(18.56)	(14.49)	(4.94)	(8.55)
	300	55.5	61.78	40.48	45.49	63.99	38.63	40.33
		(5.12)	(13.5)	(5.85)	(13.47)	(14.95)	(5.81)	(5.72)
	400	55.62	66.65	40.74	27.77	83.06	44.35	32.23
		(7.58)	(9.5)	(6.7)	(15.33)	(10.29)	(8.79)	(10.35)
	500	57.58	64.82	42.22	39.1	73.58	42.42	39.1
		(9.69)	(11.11)	(8.1)	(15.49)	(13.64)	(9.82)	(10.57)
	1000	66.03	71.89	53.8	38.09	85.79	56.44	44.68
		(4.87)	(4.48)	(6.15)	(9.63)	(5.56)	(5.83)	(7.02)
	2000	71.24	73.45	61.61	41.4	88.32	62.67	49.6
		(3.78)	(3.28)	(4.36)	(5.86)	(2.45)	(3.48)	(4.81)
	5000	76.5	78.73	68.59	49.28	88.03	66.25	56.38
		(0.48)	(1.6)	(0.69)	(3.17)	(2.4)	(3.07)	(1.37)
	10000	78.43	79.87	70.77	50.76	89.05	68.7	58.31
		(0.41)	(0.69)	(0.39)	(2.28)	(1.49)	(2.12)	(0.99)

Model	Training Set size	TEST AUC	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1-score
	Full Cohort (19250)	79.68 (0.32)	80.99 (0.32)	72.02 (0.3)	52.78 (1.76)	88.35 (0.79)	68.1 (1.03)	59.45 (1.04)
<i>RETAIN+Med-BERT</i>		58.49 (9.49)	74.02 (12.18)	39.89 (9.18)	42.82 (42.12)	62.22 (39.31)	26.74 (15.61)	28.03 (22.95)
	100	59.58 (8.6)	71.51 (12.05)	40.49 (7.87)	4.13 (8.62)	97.75 (4.93)	30.31 (35.37)	5.77 (11.6)
	200	60.61 (8.26)	71.82 (12.36)	42.78 (9.97)	10.33 (19.24)	92.61 (14.3)	44.82 (40.39)	10.52 (16.85)
	300	69.3 (6.45)	78.19 (8.32)	53.11 (8.22)	6.09 (17.35)	98.25 (5.43)	33.19 (44.03)	6.84 (17.95)
	400	74.01 (2.6)	79.56 (6.57)	59.56 (3.64)	14.1 (19.42)	96.61 (6.85)	57.46 (41.8)	18.38 (20.93)
	500	76.35 (1.21)	79.61 (3.59)	63.92 (2.28)	20.49 (17.97)	96.13 (7.29)	85.52 (14.21)	28.16 (18.37)
	1000	78.12 (1.18)	80.54 (1.99)	65.74 (2.77)	33.69 (14.33)	93.59 (7.41)	77.44 (11.28)	44.05 (11.49)
	2000	79.7 (0.62)	82.46 (1.29)	68.25 (1.27)	35.12 (8.17)	95.41 (2.43)	79.69 (5.43)	47.92 (6.91)
	5000	80.36 (0.32)	83.08 (0.99)	69.43 (1.46)	41.82 (5.7)	93.96 (2.14)	77.24 (4.36)	53.84 (4.2)
	10000	81.3 (0.55)	83.34 (0.13)	71.78 (3.1)	59.24 (9.26)	84.94 (6.43)	66.36 (6.51)	61.71 (2.43)
	Full Cohort (19250)							
<i>Logistic Regression (LR)*</i>		64.93 (3.83)	76.39 (16.16)	53.37 (3.28)	35.43 (7.86)	87.62 (7.8)	60.78 (10.26)	43.41 (4.3)
	100	70.68 (2.2)	71.35 (10.96)	59.69 (2.01)	45.3 (6.91)	87.18 (3.47)	62.94 (3.45)	52.24 (3.94)
	200	70.59 (1.56)	71.78 (8.16)	60.71 (1.59)	44.89 (4.4)	87.68 (3.29)	63.71 (4.35)	52.39 (2.29)
	300	71.55 (1.43)	71.02 (5.44)	61.85 (1.46)	46.48 (3.65)	87.99 (3.01)	65.04 (4.02)	53.99 (1.33)
	400	71.47 (1.69)	69.66 (5.23)	61.85 (1.51)	46.59 (3.44)	87.62 (3.3)	64.44 (4.48)	53.85 (1.47)
	500	73.7 (0.81)	75.23 (3.78)	65.02 (0.73)	51.32 (3.04)	87.01 (2.26)	65.26 (2.74)	57.33 (1.01)
	1000	75.04 (1.16)	75.51 (2.76)	67.33 (1.21)	52.66 (1.31)	87.66 (1.12)		58.88 (1.19)
	2000	76.91 (0.6)	77.72 (2.59)	69.99 (0.45)	54.62 (1.29)	88.06 (0.84)	66.81 (2) (1.15)	60.69 (0.69)
	5000	78.63 (0.3)	78.72 (1.01)	72.18 (0.26)	55.49 (0.94)	89.21 (0.59)	70.78 (0.91)	62.2 (0.48)
	10000							
	Full Cohort (19250)	79.94 (0)	79.45 (0)	73.59 (0)	56.82 (0)	89.49 (0)	71.79 (0)	63.43 (0)

Supplementary 4 Table 2C. Experiment 3 - Additional Metrics for PaCa-Truven prediction evaluation using smaller training set size

Threshold used for Sensitivity, Specificity, Precision and F1-score is 0.5

Model	Training Set size	TEST AUC	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1-score
GRU	100	55.23	73.49	38.39	26.27	79.69	39.79	31.39
		(4.18)	(15.84)	(4.07)	(5.88)	(4.48)	(4.44)	(5.19)
	200	59.04	69.82	42.00	33.52	78.09	42.99	36.65
		(3.76)	(6.92)	(3.78)	(12.19)	(6.81)	(3.74)	(10.55)
	300	61.33	71.51	44.16	34.47	79.93	46.11	38.12
		(3.79)	(10.05)	(4.17)	(13.24)	(7.37)	(3.07)	(11.19)
	400	62.43	71.22	45.80	37.04	79.33	47.88	41.33
		(3.26)	(10.08)	(3.54)	(7.83)	(4.89)	(3.99)	(5.52)
	500	65.24	68.51	50.05	40.91	79.30	50.7	45.00
		(1.64)	(7.23)	(2.37)	(3.97)	(4.90)	(3.67)	(1.32)
	1000	67.32	69.98	52.65	43.43	79.73	52.72	47.15
		(1.06)	(3.55)	(1.83)	(7.27)	(6.00)	(2.88)	(3.19)
	2000	70.45	69.56	57.12	47.44	79.76	55.11	50.38
		(1.07)	(3.38)	(1.52)	(7.99)	(5.98)	(3.79)	(3.61)
	5000	73.15	72.81	62.16	46.61	84.17	61.04	52.2
		(1.50)	(2.87)	(1.61)	(7.15)	(5.49)	(5.31)	(3.45)
GRU+Med-BERT	100	75.95	74.41	65.53	45.58	88.08	66.29	53.87
		(0.54)	(2.12)	(0.74)	(3.67)	(2.04)	(2.39)	(2.13)
	Full Cohort (23609)	78.17	77.31	68.93	49.54	88.06	68.2	57.2
		(0.21)	(0.35)	(0.19)	(4.02)	(2.67)	(2.92)	(1.61)
	100	53.72	61.24	36.85	49.76	50.57	24.41	25.63
		(5.74)	(21.96)	(4.00)	(52.06)	(51.8)	(17.23)	(26.23)
	200	59.88	71.19	42.5	20.89	79.77	20.79	11.80
		(6.13)	(13.57)	(6.08)	(41.57)	(41.76)	(24.2)	(20.95)
	300	63.49	73.48	46.18	10.9	89.42	27.41	6.86
		(7.62)	(11.09)	(7.97)	(30.78)	(31.12)	(35.66)	(15.76)
	400	66.2	73.58	49.53	17.24	93.11	42.30	20.00
		(6.15)	(6.13)	(8.56)	(22.9)	(10.54)	(33.99)	(23.31)
	500	69.06	75.85	55.47	16.47	94.29	54.48	19.84
		(8.53)	(10.5)	(9.56)	(22.38)	(10.29)	(33.29)	(21.38)
	1000	73.81	73.21	61.35	21.37	94.09	54.53	26.81
		(0.92)	(4.71)	(1.83)	(21.54)	(8.77)	(32.94)	(22.6)
	2000	75.75	75.51	64.32	20.15	96.53	83.46	28.71
		(0.72)	(3.04)	(1.39)	(15.68)	(4.03)	(11.69)	(19.02)
	5000	77.9	76.7	67.75	40.41	91.48	73.15	50.31
		(0.35)	(2.39)	(0.64)	(11.99)	(4.56)	(7.47)	(9.72)

Model	Training Set size	TEST AUC	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1-score
	10000	79.1 (0.24)	77.65 (1.68)	69.43 (0.49)	52.73 (4.68)	87.09 (2.77)	67.86 (2.82)	59.13 (2.02)
	Full Cohort (23609)	80.37 (0.12)	79.33 (0.17)	71.21 (0.28)	56.81 (2.00)	86.07 (1.32)	67.56 (1.31)	61.68 (0.66)
<i>Bi-GRU</i>	100	50.45 (1.94)	59.66 (17.95)	34.23 (1.32)	37.54 (21.46)	62.66 (23.22)	34.65 (1.8)	32.97 (9.44)
	200	50.43 (1.59)	51.92 (15.51)	34.19 (1.12)	28.1 (21.16)	72.21 (21.78)	34.74 (1.91)	27.44 (10.94)
	300	52.6 (2.02)	55.37 (11.29)	35.78 (1.64)	22.88 (12.28)	79.02 (12.34)	36.72 (2.82)	26.2 (8.54)
	400	55.05 (4.58)	58.98 (10.97)	37.94 (3.90)	19.22 (13.51)	83.99 (14.00)	40.2 (6.02)	23.33 (9.85)
	500	53.46 (5.97)	58.58 (8.37)	36.83 (5.14)	37.72 (22.29)	64.99 (25.26)	38.61 (7.72)	33.66 (10.3)
	1000	62.15 (7.57)	64.33 (9.19)	46.09 (7.94)	34.61 (16.91)	78.88 (19.14)	48.57 (9.66)	37.78 (10.80)
	2000	70.18 (0.99)	69.87 (3.19)	55.6 (1.39)	41.12 (4.45)	84.21 (3.05)	57.32 (2.64)	47.65 (2.64)
	5000	73.36 (0.50)	73.12 (2.56)	60.95 (0.76)	41.75 (1.89)	87.38 (1.17)	62.82 (1.41)	50.12 (1.16)
	10000	74.74 (0.50)	74.29 (1.69)	63.64 (0.70)	44.96 (2.12)	87.2 (0.86)	64.16 (0.87)	52.84 (1.40)
	Full Cohort (23609)	76.79 (0.29)	76.66 (0.21)	66.87 (0.32)	46.65 (0.81)	88.61 (0.63)	67.61 (1.05)	55.2 (0.58)
<i>Bi-GRU+Med-BERT</i>	100	54.57 (7.05)	56.85 (18.97)	37.58 (5.96)	20.9 (34.83)	78.47 (35.71)	22.74 (20.78)	14.5 (20.11)
	200	59.54 (4.48)	63.22 (14.21)	41.62 (3.61)	0.88 (2.78)	99.54 (1.47)	4.91 (15.54)	1.49 (4.71)
	300	67.78 (4.43)	75.3 (8.9)	51.37 (5.42)	13.02 (20.18)	95.33 (7.93)	29.88 (32.71)	15.49 (22.68)
	400	67.54 (5.4)	73.05 (8.54)	51.02 (7.00)	12.24 (18.47)	95.78 (7.31)	43.37 (39.43)	15.33 (20.86)
	500	70.7 (4.49)	75.78 (8.35)	56.25 (6.32)	26.65 (22.73)	90.8 (9.85)	63.59 (26.98)	31.18 (22.98)
	1000	74.03 (1.22)	74.16 (4.73)	60.89 (2.26)	12.1 (10.1)	98.27 (1.78)	65.38 (35.03)	19.38 (15.2)
	2000	76.07 (0.47)	75.9 (3.42)	64.04 (1.09)	33.77 (15.03)	92.2 (5.29)	72.59 (8.26)	43.19 (14.5)
	5000	77.94 (0.25)	76.92 (2.26)	67.23 (0.62)	47.16 (5.6)	88.4 (2.89)	67.92 (3.29)	55.33 (3.17)
	10000	79.25 (0.22)	77.88 (1.72)	69.42 (0.81)	56.75 (3.36)	84.56 (2.17)	65.33 (1.9)	60.64 (1.19)
	Full Cohort (23609)	80.57 (0.21)	79.45 (0.22)	71.54 (0.45)	56.8 (1.5)	86.02 (0.96)	67.45 (0.99)	61.65 (0.54)

Model	Training Set size	TEST AUC	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1-score
<i>RETAIN</i>	100	51.73 (3.01)	57.5 (19.24)	34.78 (2.08)	45.47 (11.36)	57.19 (10.29)	35.15 (2.35)	39.13 (4.73)
	200	54.57 (4.3)	61.01 (15)	37.61 (3.37)	17.43 (11.76)	85.71 (14.16)	42.05 (6.33)	22 (9.60)
	300	54.14 (3.47)	57.1 (10.46)	37.58 (3.22)	17.82 (15.40)	85.3 (16.33)	43.92 (7.97)	21.22 (11.3)
	400	53.88 (6.08)	59.6 (9.87)	37.07 (5.46)	25.99 (13.39)	76.94 (14.96)	39.78 (9.64)	28.3 (11.29)
	500	57.65 (4.69)	67.11 (7.61)	40.42 (4.27)	17.1 (17.65)	87.22 (16.85)	47.75 (9.84)	20.12 (12.92)
	1000	57.91 (6.54)	63.56 (7.1)	41.6 (6.88)	16.67 (13.04)	89.49 (12.34)	50.36 (12.54)	21.85 (12.95)
	2000	69 (2.72)	69.37 (4.18)	54.04 (3.31)	27.32 (10.28)	91.76 (4.01)	63.68 (3.31)	36.88 (11.02)
	5000	73.59 (0.68)	74.32 (1.63)	61.07 (1.38)	38.62 (2.20)	89.94 (0.68)	66.19 (0.82)	48.75 (1.8)
	10000	75.75 (0.46)	75.77 (0.87)	64.96 (0.6)	42.3 (0.92)	89.84 (0.68)	67.99 (1.35)	52.14 (0.77)
	Full Cohort (23609)	78.02 (0.19)	77.8 (0.20)	68.93 (0.35)	45.74 (0.63)	90.24 (0.54)	70.5 (0.96)	55.48 (0.40)
<i>RETAIN+Med-BERT</i>	100	55.72 (3.99)	68.33 (8.82)	38.61 (2.99)	26.96 (43.41)	73.78 (42.37)	18.61 (19.8)	15.24 (23.62)
	200	59.63 (7.48)	71.06 (11.55)	42.84 (7.60)	15.59 (33.72)	84.51 (33.86)	17.42 (24.06)	9.91 (19.05)
	300	61.36 (7.58)	69.63 (9.67)	46.38 (8.38)	6.55 (20.24)	93.18 (21.34)	22.4 (33.70)	4.58 (13.59)
	400	67.47 (6.21)	72.17 (6.93)	52.37 (8.11)	3.59 (11.24)	96.93 (9.67)	27.05 (41.06)	3.71 (11.48)
	500	67.39 (6.23)	75.13 (3.97)	52.05 (7.60)	4.9 (8.76)	98.69 (2.71)	36.52 (40.45)	7.67 (12.74)
	1000	71.26 (4.87)	71.55 (5.03)	57.69 (6.84)	7.71 (9.43)	98.07 (3.78)	48.48 (44.4)	12.18 (14.36)
	2000	75.67 (0.73)	75.23 (3.59)	63.46 (1.51)	0.49 (1.07)	100 (0)	30 (48.30)	0.96 (2.07)
	5000	77.91 (0.25)	77.3 (2.11)	67.27 (0.42)	9.23 (9.25)	99.38 (1.27)	94.66 (6.4)	15.46 (13.67)
	10000	78.85 (0.24)	77.45 (1.28)	68.59 (0.48)	17.2 (7.57)	98.71 (1.18)	89.82 (6.02)	27.98 (10.24)
	Full Cohort (23609)	79.98 (0.17)	79.2 (0.16)	69.39 (0.97)	33.43 (8.33)	95.49 (2.83)	80.56 (5.16)	46.4 (7.02)
<i>Logistic</i>	100	67.44 (2.77)	67.28 (13.72)	51.77 (2.96)	25.99 (9.85)	90.43 (5.27)	59.31 (4.53)	34.9 (8.58)
	200	69.74 (1.15)	73.98 (13.98)	54.83 (1.55)	33.08 (9.3)	88.59 (4.48)	60.63 (3.92)	41.75 (8.16)

Model	Training Set size	TEST AUC	Validation AUC	AUPRC	Sensitivity	Specificity	Precision (PPV)	F1-score
	300	70.98	70.91	56.66	36.56	87.9	60.85	45.52
		(1.16)	(6.95)	(1.61)	(3.57)	(2.22)	(2.74)	(2.5)
	400	70.56	72.32	56.08	39.36	85.99	59.26	47
		(1.42)	(6.97)	(1.36)	(5.46)	(3.37)	(2.99)	(3.10)
	500	71.28	74.45	57.4	40.94	86.19	60.34	48.64
		(1.36)	(3.31)	(1.77)	(3.88)	(2.17)	(2.48)	(2.5)
	1000	71.98	72.79	58.84	43.25	85.59	60.53	50.37
		(0.86)	(3.97)	(1.20)	(2.92)	(1.62)	(1.48)	(1.73)
	2000	73.38	72.55	61.36	46.67	85.21	61.67	53.1
		(0.56)	(3.1)	(0.61)	(2.09)	(0.97)	(0.66)	(1.24)
	5000	74.58	74.26	63.69	46.02	87.04	64.4	53.68
		(0.51)	(1.71)	(0.51)	(0.62)	(0.58)	(1.10)	(0.65)
	10000	75.72	75.5	65.33	46.23	87.81	65.9	54.33
		(0.43)	(1.13)	(0.46)	(1.12)	(0.58)	(0.90)	(0.81)
	Full Cohort (23609)	77.28 (0)	77.11 (0)	67.33 (0)	45.52 (0)	89.17 (0)	68.16 (0)	54.58 (0)

4.8. References

1. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology* 2017;**2**(4):230-43
2. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nature biomedical engineering* 2018;**2**(10):719-31
3. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. *Ieee Access* 2017;**5**:8869-79
4. Wang H, Cui Z, Chen Y, Avidan M, Abdallah AB, Kronzer A. Predicting hospital readmission via cost-sensitive deep learning. *IEEE/ACM transactions on computational biology and bioinformatics* 2018;**15**(6):1968-78
5. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future healthcare journal* 2019;**6**(2):94
6. Lysaght T, Lim HY, Xafis V, Ngiam KY. AI-Assisted Decision-making in Healthcare. *Asian Bioethics Review* 2019;**11**(3):299-314
7. Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database* 2020;**2020**
8. Manogaran G, Lopez D. Health data analytics using scalable logistic regression with stochastic gradient descent. *International Journal of Advanced Intelligence Paradigms* 2018;**10**(1-2):118-32
9. Keerthika T, Premalatha K. An effective feature selection for heart disease prediction with aid of hybrid kernel SVM. *International Journal of Business Intelligence and Data Mining* 2019;**15**(3):306-26

10. Sadek RM, Mohammed SA, Abunbehan ARK, et al. Parkinson's Disease Prediction Using Artificial Neural Network. 2019
11. Payan A, Montana G. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. arXiv preprint arXiv:1502.02506 2015
12. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. Advances in Neural Information Processing Systems; 2016.
13. Doctor ai: Predicting clinical events via recurrent neural networks. Machine Learning for Healthcare Conference; 2016.
14. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine 2018;**1**(1):18
15. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;**542**(7639):115-18 doi: 10.1038/nature21056[published Online First: Epub Date]].
16. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering 2018;**2**(3):158
17. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nature medicine 2018;**24**(10):1559-67
18. Chung SW, Han SS, Lee JW, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta orthopaedica 2018;**89**(4):468-73
19. Shen J, Zhang CJ, Jiang B, et al. Artificial intelligence versus clinicians in disease diagnosis: Systematic review. JMIR medical informatics 2019;**7**(3):e10010
20. Revisiting unreasonable effectiveness of data in deep learning era. Proceedings of the IEEE international conference on computer vision; 2017.

21. Cho J, Lee K, Shin E, Choy G, Do S. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? arXiv preprint arXiv:1511.06348 2015
22. Gentil M-L, Cuggia M, Fiquet L, et al. Factors influencing the development of primary care data collection projects from electronic health records: a systematic review of the literature. BMC medical informatics and decision making 2017;**17**(1):139
23. Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering 2009;**22**(10):1345-59
24. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems; 2013.
25. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014.
26. Peters ME, Neumann M, Iyyer M, et al. Deep contextualized word representations. arXiv preprint arXiv:1802.05365 2018
27. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. URL [https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf](https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language%20understanding%20paper.pdf) 2018
28. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog 2019;**1**(8):9
29. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 2018
30. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems; 2019.
31. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 2020

32. Videobert: A joint model for video and language representation learning. Proceedings of the IEEE International Conference on Computer Vision; 2019.
33. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**(4):1234-40 doi: 10.1093/bioinformatics/btz682[published Online First: Epub Date]].
34. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323 2019
35. Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129 2019
36. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 2019
37. Adhikari A, Ram A, Tang R, Lin J. Docbert: Bert for document classification. arXiv preprint arXiv:1904.08398 2019
38. Pires T, Schlinger E, Garrette D. How multilingual is Multilingual BERT? arXiv preprint arXiv:1906.01502 2019
39. SciBERT: A pretrained language model for scientific text. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019.
40. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 2018
41. Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342 2019
42. Jha AK, DesRoches CM, Campbell EG, et al. Use of electronic health records in US hospitals. *New England Journal of Medicine* 2009;**360**(16):1628-38

43. Blumenthal D, Tavenner M. The “meaningful use” regulation for electronic health records. *New England Journal of Medicine* 2010;**363**(6):501-04
44. Gupta P, Malhotra P, Narwariya J, Vig L, Shroff G. Transfer learning for clinical time series analysis using deep neural networks. *Journal of Healthcare Informatics Research* 2020;**4**(2):112-37
45. Beam AL, Kompa B, Fried I, et al. Clinical concept embeddings learned from massive sources of multimodal medical data. *arXiv preprint arXiv:1804.01486* 2018
46. Xiang Y, Xu J, Si Y, et al. Time-sensitive clinical concept embeddings learned from large electronic health records. *BMC Med Inform Decis Mak* 2019;**19**(Suppl 2):58 doi: 10.1186/s12911-019-0766-3[published Online First: Epub Date].
47. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*; 2014.
48. Howard J, Ruder S. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* 2018
49. Peters ME, Neumann M, Iyyer M, et al. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* 2018
50. Li Y, Rao S, Solares JRA, et al. BeHRt: transformer for electronic Health Records. *Scientific Reports* 2020;**10**(1):1-12
51. Shang J, Ma T, Xiao C, Sun J. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346* 2019
52. Xiang Y, Xu J, Si Y, et al. Time-sensitive clinical concept embeddings learned from large electronic health records. *BMC Med Inform Decis Mak* 2019;**19**(Suppl 2):58 doi: 10.1186/s12911-019-0766-3[published Online First: Epub Date].
53. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems*; 2016.

54. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining; 2017.
55. Xiao C, Ma T, Dieng AB, Blei DM, Wang F. Readmission prediction via deep contextual embedding of clinical concepts. PloS one 2018;**13**(4):e0195024
56. Xiang Y, Ji H, Zhou Y, et al. Asthma Exacerbation Prediction and Risk Factor Analysis Based on a Time-Sensitive, Attentive Neural Network: Retrospective Cohort Study. Journal of medical Internet research 2020;**22**(7):e16981
57. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 2018
58. Baytas IM, Xiao C, Zhang X, Wang F, Jain AK, Zhou J. Patient Subtyping via Time-Aware LSTM Networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017:65-74.
59. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 2014
60. Zhao R, Wang D, Yan R, Mao K, Shen F, Wang J. Machine health monitoring using local feature-based gated recurrent unit networks. IEEE Transactions on Industrial Electronics 2017;**65**(2):1539-48
61. Xiang Y, Xu J, Si Y, et al. Time-sensitive clinical concept embeddings learned from large electronic health records. BMC Med Inform Decis Mak 2019;**19**(Suppl 2):58 doi: 10.1186/s12911-019-0766-3[published Online First: Epub Date].
62. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 2017;**5**:135-46

63. Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics 2015;3:211-25
64. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014.
65. Wolf T, Debut L, Sanh V, et al. Huggingface’s transformers: State-of-the-art natural language processing. ArXiv, abs/1910.03771 2019
66. Vig J. A multiscale visualization of attention in the transformer model. arXiv preprint arXiv:1906.05714 2019
67. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. URL [https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf](https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language%20understanding%20paper.pdf) 2018
68. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165 2020

Chapter 5: Conclusions, Discussions, and Recommendations

5.1. Use of secondary EHR data for evidence-based medicine

Secondary EHR and claims data are rich sources of patients' clinical data. As biomedical researchers, we can extract knowledge and train useful machine learning models to predict different types of clinical events, taking into consideration the common quality issues associated with such data. A comprehensive understanding of the flow of the data from the point at which it was created at the healthcare site until our access to it is a key success factor in the majority of studies that use secondary clinical data. Such knowledge can help us to make informed decisions on the relevance of the data for answering the proposed research question and can guide the data extraction and preparation efforts. The information that we need to know includes:

- How the data were entered by the clinical staff into the EHR.
- When the data are recorded in the system versus the dates inserted manually.
- How the billing department reviews and assigns billing codes, and the frequency of reviewing and transmitting claims.
- How data providers collect the data from different sites and how they combine, clean, and de-identify the data.

A clear understanding of the research question and the prediction task as well as the nature of the data are needed for a good study design. As presented in Chapter 2, the main purpose of the study was to compare the three most commonly used vasopressors,

namely dopamine, norepinephrine, and phenylephrine, and their associations with the in-hospital mortality of nontraumatic SAH patients. Out of the originally identified 39,000 SAH patients, our study considered the data of only 4,810 SAH patients after applying conservative exclusion criteria to ensure that the study cohort data are of sufficient quality based on the nature of the clinical problem. In addition, we used mainly causal inference analysis and propensity score models to account for potential confounding factors. We also followed a similar strategy in a later study to understand the effect of Tocilizumab on the in-hospital mortality or need for the intubation of COVID-19 patients [87]. Given that the data were not originally collected for these research purposes (they were collected mainly for billing purposes), the findings from such retrospective cohort observational studies, using secondary EHR data, need to be further validated through clinical trials.

In Chapters 3 and 4, we focused on training predictive models for clinical events, such as predicting a patient's risk of being diagnosed with pancreatic cancer (PaCa) or a diabetic patient's risk of developing heart failure (DHF), based on all historical clinical data available at the time of prediction. We compared machine learning and DL methods to train such predictive models converting all input data to a categorical format in a way that can reduce the risk of common data issues, such as missing or implausible data[88]. In Chapter 4, we applied a more conservative phenotyping algorithm for the definition of diabetic and heart failure patients. We found that, even as the eligible cohort size becomes reduced by about 40%, the performance of the trained model was not severely affected. For example, the logistic regression model AUC using raw diagnosis data as

shown in chapter 3 table 2 was 80.6% while in chapter 4 table 4 the logistic regression model AUC for the DHF task is 81.0%. Notably, as stated in both chapters, DL-based algorithms were commonly associated with higher prediction accuracy as compared to traditional machine learning models.

Although we can somewhat open the black-box and visualize the correlation between different clinical codes at each layer in the model architecture and the final prediction, as we demonstrated in Chapter 4, it is not wise at this stage in the research process to use such correlations as clinical evidence or a method to identify clinical risk factors. The evaluation of the explanations of DL-based models is challenging because the currently available methods are providing sample (patient) level explanations rather than population-level explanations. Future work to improve the explainability of DL-based models and to describe coherent and feasible methods to clinically evaluate such explanations is warranted.

5.2. Sequential Deep Learning Modeling using EHR data

In Chapters 3 and 4, we used two different baseline DL methods for modeling sequential EHR data, RNN and Med-BERT.

5.2.1. Recurrent neural network-based models

RNN is known for its ability to learn from sequential information; it maps perfectly the structure of the patient health record, which consists of a sequence of visits, with each visit is rich in data[89,90]. For RNN models, we first embed our input information into lower dimension vectors for computational efficiency and better data representation learning.

That embedding layer will be fed into the RNN layer(s), which will build the model learning from the patients' sequential visits information. We then feed the final layer of RNN to a fully connected layer with a sigmoid function to predict the probability of the clinical event.

RNN models are known to have multiple hyperparameters that can have an impact on the training quality of the model. In earlier research[90,91], we found that a single-layer RNN-based model, using a gated recurrent unit cell structure and careful selection of hyperparameters, can achieve equivalent or even sometimes better results as compared to a more complicated architecture. Therefore, we used the Bayesian optimization search[92] to determine the best hyperparameters, given our data source and prediction task, and used those hyperparameters in the majority of the experiments, as explained in Chapter 3. A key advantage of this simple model architecture is the small number of parameters that can contribute to better efficiency as we focus on the implementability of the proposed model.

5.2.2. Transformers

Transformers proved to be a valid alternative for sequential modeling. Since the end of 2018, after Google released its first pre-trained BERT model, based on the transformer structure, transformer-based models have continued to evolve. In Med-BERT, as explained in Chapter 4, we trained a BERT-like- model on diagnosis data of more than 20 million patients, and the very first version was able to improve the prediction accuracy when fine-tuned to predict patient risk to develop a specific disease, such as heart failure or pancreatic cancer. In addition to the limitations described in Chapter 4, a drawback of Med-BERT is

the high number of parameters, which makes the model size reach over 60 MB, while the RNN-based model does not exceed 10 MB when trained on the same data. We are currently evaluating newer transformers architectures, such as Distill-BERT and ALBERT, and linear transformers, such as Luna, which provide better prediction accuracy in the NLP domain while providing better computational efficiency with a smaller number of parameters, lower memory consumption, and much shorter running time. We are also training a more comprehensive version of Med-BERT that can consider additional clinical events, such as medications and procedures. The initial results show a slight improvement in the performance and efficiency of ALBERT-based Med-BERT; however, this study is still in progress.

5.2.3. Model explainability

There are two popular mechanisms to interpret or explain the sequential DL architectures predictions, namely attention and attribution mechanisms. In Chapter 4, we used code-level attentions to visualize the relations between different clinical codes at each layer. In an earlier study[86], we evaluated the generalizability of an RNN-based model architecture, RETAIN[89], which uses the attention mechanism to calculate the contribution score for each clinical code at each visit. We found, however, that adding more trainable model parameters could have an impact on the efficiency of the model. Therefore, in our latest work[93], we used an attribution mechanism, such as integrated gradients, to provide a contribution score for each code at each visit for the final prediction. Attribution-based explanations can be independently calculated upon request, without the need to add more trainable parameters to the prediction model. As we discussed, a major limitation of both

mechanisms is the difficulty to clinically evaluate the calculated contribution of each medical code toward personalized predictions, which should be addressed in future research. First, we will need to compare the contribution scores for each code-visit level, using the different attribution techniques available in the CAPTUM [94] package, such as DeepLift [71,95] or SHAP[96], or an attention mechanism, such as RETAIN [70,96,97]. Second, we will need to evaluate how meaningful the explanations are to clinicians.

5.3. Contribution to Science

In this dissertation, we presented three studies in which we leveraged one of the earliest and largest secondary EHR data source, namely Cerner HealthFacts®. In the first study presented in Chapter 2, we conducted a retrospective observational study on approximately 3,000 nontraumatic SAH patients, which was the largest cohort used to compare the choice of the initial vasopressor associations with the outcomes of nontraumatic SAH patients, at the study time. Using a novel statistical method, we found evidence that the selection of phenylephrine as the primary vasopressor to induce hypertension for the management of nontraumatic SAH is associated with better outcomes as compared to selecting norepinephrine or dopamine. In the second study presented in Chapter 3, we used two large case-control cohorts to train generalizable models for the risk prediction of specific clinical events, such as heart failure in diabetes patients or pancreatic cancer. The first cohort consisted of approximately 30,000 patients for the PaCa prediction task, including more than 10,000 PaCa patients, and the second cohort consisted of 120,000 diabetic patients with 50% of them developed heart failure at

least 30 days after their first DMII diagnosis. Besides the phenotyping algorithms and the terminology mapping methods which we shared in the supplementary material (section 3.9), we were the first to show that training machine learning and deep learning models on diagnosis information as originally recorded or mapped to expressive terminologies, which preserve a high level of granularity, were associated with better prediction performance as compared with models trained on diagnosis information mapped to lower dimension terminologies that have only a few hundreds of codes. Additionally, we proved that RNN based models were providing better prediction performance as compared to baseline machine learning methods, such as logistic regression, especially when trained on large cohorts. Finally, in our third study described in Chapter 3, we presented Med-BERT, the first foundation model trained on structured diagnosis data for more than 20 million patients that are coded in the standard ICD-9 and ICD-10 format. Med-BERT was found to improve the prediction performance of downstream tasks that have a small sample size, which otherwise would limit the ability of the model to learn good representation. In conclusion, we found that we can extract useful information and train helpful deep learning-based predictive models using a rich secondary EHR data source, however, the findings need clinical validation.

5.4. In-progress Work and Future Directions

Shiffman et al.[98,99] defined the implementability for clinical guidelines, as the set of characteristics that predict ease of use or determine the key obstacles for guideline implementation. Similarly, we define implementability in the context of artificial

intelligence and predictive modeling of clinical events, as the feasibility assessment of the developed predictive model to demonstrate whether the model can be further evaluated in clinical settings. Therefore, along with subject matter experts, we are working on establishing a framework that defines the factors associated with the implementability of DL-based predictive models (Table 5.1) and that describes how we, as biomedical data science researchers, can consider the implementability evaluation during our research projects design.

Table 5.1. Deep learning based models Implementability evaluation factors.

Implementability Factors	Definition	Evaluation Criteria
Performance	How far predictions deviate from actual observation on a testing dataset	Discriminative accuracy, model calibration, and meaningful metrics: AUROC, AUPRC, sensitivity, specificity, PPV,...etc
Transparency	How a given technology reached a certain decision	input features contribution scores to facilitate clinical validation, TRIPOD
Generalizability	The ability of the model, after being trained to digest new data and make accurate predictions regardless of the setting or the population	External validation, reproducibility, scalability, fairness, adaptability
Data Mechanics	How data can flow between systems and computational infrastructures	Data extraction, standardization, and preprocessing steps
Efficiency	The amount of time and computational resources required by the model to work properly	model size, running time, and associated cost of running

Data Privacy	Models should avoid using any protected health information (PHI)	Amount of PHI data required by the model
--------------	--	--

While a comprehensive evaluation plan needs to be agreed on during the early phase of prediction task definition, we propose six factors that need to be considered in the evaluation of the implementability of a predictive model, based on our proposed framework. Those factors are: prediction performance, transparency, generalizability, data mechanics, efficiency, and data privacy. In the following sections, I briefly describe each factor and demonstrate by example, either from Chapters 3 and 4 or from our latest published work (CovRNN)[93], how to consider such factors during the model’s early development phase. Prior to that, I describe our Pytorch_EHR framework, which we established based on our research. Our CovRNN paper, in which we adopted the latest version of Pytorch_EHR, is the first to consider all six factors. A full evaluation of our approach, however, is still needed.

5.4.1. Pytorch_ehr framework to train and evaluate an implementable deep learning predictive model

Based on our previously described work, we established a framework to train and evaluate DL-based models to predict different types of clinical events, as seen in Figure 5.1. The framework involves the data preparation flow (highlighted in green), which includes the definitions of the prediction task and evaluation plan, which guide the definition of the cohort characteristics and labeling, later dividing into training and evaluation test sets. Once the cohort is defined and our eligible patients are identified, we start the process of raw data extraction, applying terminology normalization if needed,

and, later, we preprocess the data to become in the input format that can be efficiently consumed by the model. The first layer of the model is the embedding layer, which can be randomly initialized or initialized from a pre-trained static or a contextualized embedding, such as Med-BERT. The feature representations trained/fine-tuned in the embedding layer, along with time information or other continuous data variables, are then fed into the core model architecture, which can be as simple as a linear layer or a more complicated RNN model, to achieve a binary classification or a survival prediction. In addition to the basic predictive model training module, we included three ancillary modules. First is the explainability module, which currently uses the integrated gradient technique. Second is a multimetric performance evaluation module, which includes functions to find the recommended threshold for classification when using unbalanced datasets as well as functions to calculate various evaluation metrics, such as specificity at 95% sensitivity, sensitivity at the best/recommended threshold value, AUROC and AUPRC for binary predictions, and the concordance index (c-index) for survival, of interest to clinicians. This module also includes several informative plots, such as the calibration plot and stratified Kaplan-Meier (KM) curves, based on the predicted survival probabilities. Third is a subgroup analysis module that can be used to calculate the model performance for different subgroups based on their demographics, location, or common comorbidities.

For reproducibility and further evaluation by researchers, we share our codebase as an open-source repository at https://github.com/ZhiGroup/pytorch_ehr. In addition, an end-to-end tutorial, using an MIMIC IV database and Google collaboration workbooks, is

available through the https://github.com/ZhiGroup/pytorch_ehr/tree/ACM_BCB-Tutorial, which was used during our tutorial presentation at the 2021 ACM BCB conference.

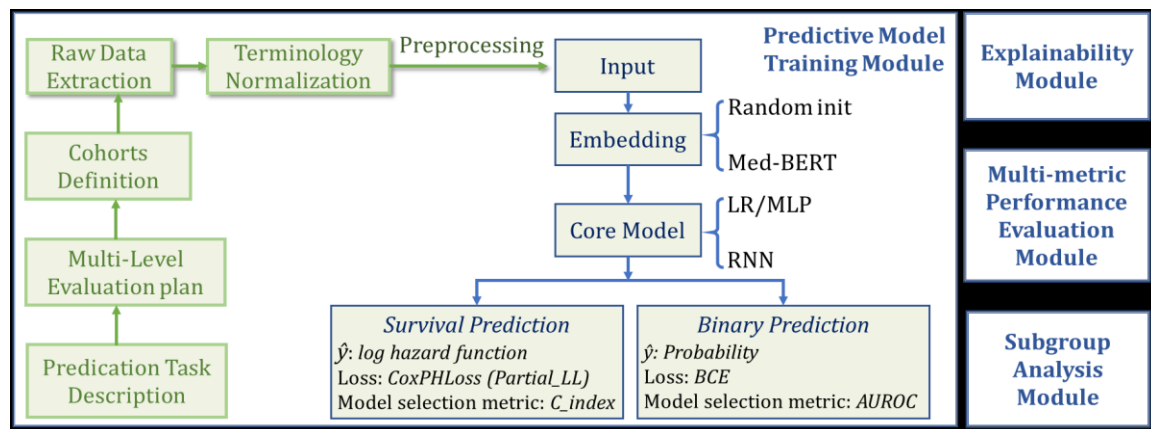


Figure 5.1. Pytorch_EHR Framework

5.4.2. Implementability evaluation factors

5.4.2.1. Prediction Performance

We define prediction performance as how far predictions deviate from observation in a testing dataset. Prediction performance is the most commonly reported evaluation result, and the AUROC, which represents the discriminative accuracy of the proposed models, especially those based on classification methods, is the most commonly reported metric. In our proposed framework, there is a multi-metric performance evaluation module to facilitate the calculation of clinically relevant metrics based on the prediction task. As an example, in Chapter 4 and in our latest study[93], we utilized this module to report the model's AUROC, AUPRC, and specificity at 95% sensitivity, as well as the sensitivity, specificity, and F1 score at the recommended thresholds. These recommended thresholds

were defined as the thresholds that achieve the best balance between the models' sensitivity and specificity using the validation set. We also plotted the calibrations curves as well as the stratified KM curves for low, medium, and high-risk groups, using the survival models.

5.4.2.2. Transparency

We define transparency as the detailed explanation of how a given technology reaches a certain decision. Transparency is the most common principle seen in all responsible/ethical AI frameworks. To evaluate the transparency of the proposed models, we need to explain the model predictions and to transparently report our study design, including the cohort definition and labeling criteria as well as our results. As seen in the section with the definition of cohorts in Chapters 2, 3, and 4 and the referenced supplementary material, we explain the details of our phenotyping algorithms and share our inclusion and exclusion criteria to enable reproduction and further evaluation of our work. Allowing that using the TRIPOD assessment checklist can help researchers to transparently report the study design and findings, our framework offers an explainability module to explain the DL model prediction, using an attribution mechanism known as integrated gradients. As an example, in our latest study[93], we included a sample patient-visit-level explanation and our reported TRIPOD assessment in the supplementary material.

5.4.2.3. Generalizability

We define generalizability as the ability of the model, after being trained to digest new data and make accurate predictions, regardless of the setting or population. To evaluate

the generalizability of the model, we present three recommendations. The first is to predefine a multi-level evaluation plan and, accordingly, define the evaluation test sets in an early phase of the study design. Such test sets need to be from different sources and of different sizes and complexity. The second is the use of different data sources to evaluate the transferability of the trained models and, accordingly, decide on the terminology normalization strategy, if needed. The third is the reporting of any subgroup analysis to acknowledge any difference in the proposed model performance among different age groups, gender, race, or ethnicity. Therefore, in our proposed framework, we have a specific subgroup analysis module that can be used to calculate the model performance of different subgroups based on their demographics, location, or common comorbidities. As an example, in our CovRNN study[93], we defined four different test sets. The first is a large multi-hospital test set with more than 48,000 patients. Then, there are two test sets, with each representing full hospital data from different regions and of different sizes. Finally, there is a test set from a completely different data source (Optum®). As we found that the training data, except for medications, are mainly in the common standard terminologies in use in different EHRs, and given our finding that the raw data format is commonly associated with one of the best model prediction performances, we decided to train the CovRNN models on the standard terminologies available for diagnosis, procedures, assessments, and laboratory results. We normalized medication information into Multum identifiers and categories, mapping NDC codes to their corresponding Multum codes, using standard software. Finally, we utilized the subgroup analysis module to compare the performance of the model for different age groups, regions, races,

and comorbidities and acknowledged that the model's discriminative accuracy slightly decreased for older age groups.

5.4.2.4. Data Mechanics, Efficiency, and Privacy

We define data mechanics as how the data flow between systems and computational infrastructures. These mechanics are essential to understand when designing the model's integration into the clinical workflow. In addition, as we understand the steps of the data flow until it reaches the model and how the model will process the data to calculate the prediction risk score and corresponding explanations, we can make a preliminary judgment of the proposed model's efficiency.

As per MLOps standards, we define efficiency as the amount of time and computational resources required by the model to work properly. Efficiency evaluation factors include the projected running time and computational resource, e.g., memory and storage utilization, consumption, which ultimately becomes a cost. Figure 5.2 provides an example of the flow of data from EHR standard tables to the CovRNN model to calculate the patient risk score as well as to provide prediction explanations.

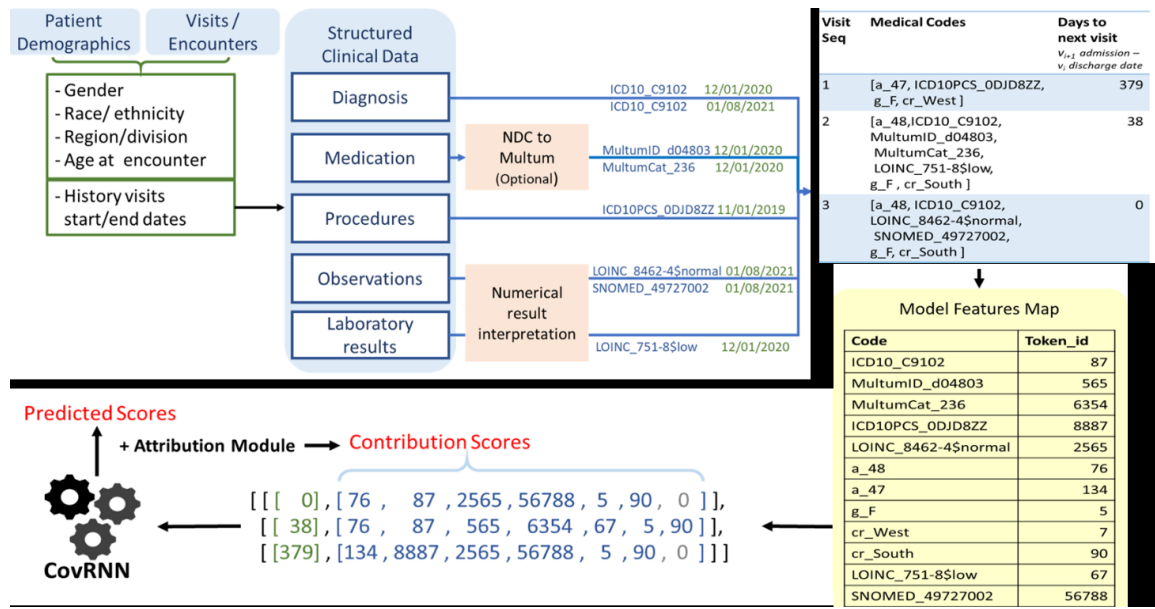


Figure 5.2. CovRNN Data Mechanics

Ensuring the security of the data flow, from the EHR to the model and the prediction back to the EHR, is more manageable during the implementation phase. During the model development, researchers need to avoid the utilization of any PHI data as much as possible. In our work, as explained in chapters 2,3, and 4, as well as in other publications [40 47], we never use PHI data to feed our models. In the majority of our work, we train our models on de-identified data, which helps to ensure that our model is HIPAA compliant. For example, when we use patient age, any patient who is 90 years or older is assigned the age of 90. Although that might explain the slight decrease in the prediction accuracy in the elderly age group, it is essential to ensure patient data privacy.

5.5. Final Thoughts

In this dissertation, we demonstrated how the utilization of large secondary EHR databases can help to create knowledge and train generalizable predictive models, using

innovative tools and methods. Future work should consider the improvement of performance and the usefulness of the work from the early model development phase. Despite the advancements in AI and DL-based methods for predictive modeling, there is a lack of evaluation of such methods. Researchers need to improve the explainability of clinical DL-based predictive models and, more importantly, to study how to efficiently evaluate the explanation of model predictions in a more clinically relevant way. It is our hope that researchers also will consider evaluating the efficiency and costs associated with their proposed methods, following industry standards, such as MLOps.

References

- 1 Evans RS. Electronic Health Records: Then, Now, and in the Future. Published Online First: 2016. doi:10.15265/IYS-2016-s006
- 2 RF G. From papyrus to the electronic tablet: a brief history of the clinical medical record with lessons for the digital age. *Am J Med* 2013;**126**:853–7. doi:10.1016/J.AMJMED.2013.03.024
- 3 CMS AND ONC FINAL REGULATIONS DEFINE MEANINGFUL USE AND SET STANDARDS FOR ELECTRONIC HEALTH RECORD INCENTIVE PROGRAM | CMS. <https://www.cms.gov/newsroom/fact-sheets/cms-and-onc-final-regulations-define-meaningful-use-and-set-standards-electronic-health-record> (accessed 13 Oct 2021).
- 4 Henry J, Pylypchuk Y, Searcy T, *et al.* Data Brief 35: Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015.
- 5 Huser V, Cimino JJ. Desiderata for Healthcare Integrated Data Repositories Based on Architectural Comparison of Three Public Repositories. *AMIA Annu Symp Proc* 2013;**2013**:648.
- 6 i2b2: Informatics for Integrating Biology & the Bedside. <https://www.i2b2.org/about/> (accessed 20 Dec 2019).
- 7 Klann JG, Phillips LC, Herrick C, *et al.* Web services for data warehouses: OMOP

- and PCORnet on i2b2. *J Am Med Informatics Assoc* 2018;**25**:1331–8.
doi:10.1093/JAMIA/OCY093
- 8 Klann JG, Joss MAH, Embree K, *et al.* Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. *PLoS One* 2019;**14**:e0212463. doi:10.1371/journal.pone.0212463
 - 9 McMurry AJ, Murphy SN, MacFadden D, *et al.* SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies. *PLoS One* 2013;**8**.
doi:10.1371/JOURNAL.PONE.0055811
 - 10 Brat GA, Weber GM, Gehlenborg N, *et al.* International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *npj Digit Med* 2020 31 2020;**3**:1–9. doi:10.1038/s41746-020-00308-0
 - 11 Weber GM, Zhang HG, L’Yi S, *et al.* International Changes in COVID-19 Clinical Trajectories Across 315 Hospitals and 6 Countries: Retrospective Cohort Study. *J Med Internet Res* 2021;**23**(10)e31400 <https://www.jmir.org/2021/10/e31400> 2021;**23**:e31400. doi:10.2196/31400
 - 12 The “All of Us” Research Program. *N Engl J Med* 2019;**381**:668–76.
doi:10.1056/NEJMSR1809937
 - 13 Data Methodology – All of Us Research Hub.
<https://www.researchallofus.org/data-tools/methods/> (accessed 8 Dec 2021).
 - 14 National COVID Cohort Collaborative (N3C) | National Center for Advancing Translational Sciences. <https://ncats.nih.gov/n3c> (accessed 8 Dec 2021).
 - 15 Bennett TD, Moffitt RA, Hajagos JG, *et al.* Clinical Characterization and

- Prediction of Clinical Severity of SARS-CoV-2 Infection Among US Adults Using Data From the US National COVID Cohort Collaborative. *JAMA Netw Open* 2021;**4**:e2116901–e2116901. doi:10.1001/JAMANETWORKOPEN.2021.16901
- 16 OMOP CDM v5.4. <http://ohdsi.github.io/CommonDataModel/cdm54.html> (accessed 8 Dec 2021).
 - 17 *Chapter 4 The Common Data Model | The Book of OHDSI.*
 - 18 Electronic Medical Records and Genomics (eMERGE) Network. <https://www.genome.gov/Funded-Programs-Projects/Electronic-Medical-Records-and-Genomics-Network-eMERGE> (accessed 8 Dec 2021).
 - 19 Gottesman O, Kuivaniemi H, Tromp G, *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med* 2013 *15*10 2013;**15**:761–71. doi:10.1038/gim.2013.72
 - 20 Collins FS, Hudson KL, Briggs JP, *et al.* PCORnet: turning a dream into reality. *J Am Med Inform Assoc* 2014;**21**:576–7. doi:10.1136/amiajnl-2014-002864
 - 21 Fleurence RL, Curtis LH, Califf RM, *et al.* Brief communication: Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc* 2014;**21**:578. doi:10.1136/AMIAJNL-2014-002747
 - 22 CBER Biologics Effectiveness and Safety (BEST) System | FDA. <https://www.fda.gov/vaccines-blood-biologics/safety-availability-biologics/cber-biologics-effectiveness-and-safety-best-system> (accessed 8 Dec 2021).
 - 23 BEST Initiative. <https://www.bestinitiative.org/> (accessed 8 Dec 2021).
 - 24 Klann JG, Abend A, Raghavan VA, *et al.* Data interchange using i2b2. *J Am Med*

- Informatics Assoc* 2016;**23**:909–15. doi:10.1093/jamia/ocv188
- 25 Real-World Data solution | Cerner. <https://www.cerner.com/solutions/real-world-data> (accessed 9 Dec 2021).
- 26 Epic Health Research Network. <https://ehrn.org/> (accessed 9 Dec 2021).
- 27 Software | Epic. <https://www.epic.com/software#Cosmos> (accessed 13 Dec 2021).
- 28 Milinovich A, Kattan MW. Extracting and utilizing electronic health data from Epic for research. *Ann Transl Med* 2018;**6**:42–42. doi:10.21037/ATM.2018.01.13
- 29 BCBS Axis | Blue Cross Blue Shield. <https://www.bcbs.com/about-us/capabilities-initiatives/bcbs-axis> (accessed 9 Dec 2021).
- 30 Zhu G, Ly VK, Gonzalez M, *et al*. EHR Databases and Data Management: Data Query and Extraction. *Stat Mach Learn Methods EHR Data* 2020;:53–77. doi:10.1201/9781003030003-3
- 31 MarketScan Research Databases | IBM. <https://www.ibm.com/products/marketscan-research-databases> (accessed 9 Dec 2021).
- 32 Data & Analytics Services For Government Health Agencies. <https://www.optum.com/business/solutions/government/federal/data-analytics-federal.html> (accessed 9 Dec 2021).
- 33 Real World Data Sets - IQVIA. <https://www.iqvia.com/solutions/real-world-evidence/real-world-data-and-insights> (accessed 9 Dec 2021).
- 34 Franklin JM, Gopalakrishnan C, Krumme AA, *et al*. The relative benefits of claims and electronic health record data for predicting medication adherence trajectory.

- Am Heart J* 2018;**197**:153–62. doi:10.1016/J.AHJ.2017.09.019
- 35 Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Med Res Methodol* 2021; **21**:1–10. doi:10.1186/S12874-021-01416-5
 - 36 Dedman D, Cabecinha M, Williams R, *et al.* Approaches for combining primary care electronic health record data from multiple sources: a systematic review of observational studies. *BMJ Open* 2020;**10**:e037405. doi:10.1136/BMJOPEN-2020-037405
 - 37 Feng C, Le D, McCoy AB. Using Electronic Health Records to Identify Adverse Drug Events in Ambulatory Care: A Systematic Review. *Appl Clin Inform* 2019;**10**:123–8. doi:10.1055/S-0039-1677738
 - 38 Callahan A, Shah NH, Chen JH. Research and Reporting Considerations for Observational Studies Using Electronic Health Record Data. <https://doi.org/107326/M19-0873> 2020;**172**:S79–84. doi:10.7326/M19-0873
 - 39 Kohane IS, Aronow BJ, Avillach P, *et al.* What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. *J Med Internet Res* 2021;**23**(3)e22219 <https://www.jmir.org/2021/3/e22219> 2021;**23**:e22219. doi:10.2196/22219
 - 40 Botsis T, Hartvigsen G, Chen F, *et al.* Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit on Translat Bioinforma* 2010;**2010**:1.
 - 41 Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Med Res Methodol*

- 2021;**21**:1–10. doi:10.1186/S12874-021-01416-5/TABLES/1
- 42 Diaz-Garelli F, Johnson TR, Rahbar MH, *et al.* Exploring the Hazards of Scaling Up Clinical Data Analyses: A Drug Side Effect Discovery Case Report. *AMIA Summits Transl Sci Proc* 2021;**2021**:180.
- 43 Williams G, Maroufy V, Rasmy L, *et al.* Vasopressor treatment and mortality following nontraumatic subarachnoid hemorrhage: a nationwide electronic health record analysis. *Neurosurg Focus* 2020;**48**:E4. doi:10.3171/2020.2.FOCUS191002
- 44 Graham S, Depp C, Lee EE, *et al.* Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. *Curr Psychiatry Rep* 2019;**21**. doi:10.1007/S11920-019-1094-0
- 45 Kilic A. Artificial Intelligence and Machine Learning in Cardiovascular Health Care. *Ann Thorac Surg* 2020;**109**:1323–9. doi:10.1016/J.ATHORACSUR.2019.09.042
- 46 Li R, Chen Y, Ritchie MD, *et al.* Electronic health records and polygenic risk scores for predicting disease risk. *Nat Rev Genet* 2020 218 2020;**21**:493–502. doi:10.1038/s41576-020-0224-1
- 47 Hossain ME, Khan A, Moni MA, *et al.* Use of Electronic Health Data for Disease Prediction: A Comprehensive Literature Review. *IEEE/ACM Trans Comput Biol Bioinforma* 2021;**18**:745–58. doi:10.1109/TCBB.2019.2937862
- 48 Fu LH, Schwartz J, Moy A, *et al.* Development and Validation of Early Warning Score System: A Systematic Literature Review. *J Biomed Inform* 2020;**105**:103410. doi:10.1016/J.JBI.2020.103410

- 49 Goldstein BA, Navar AM, Pencina MJ, *et al.* Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017;**24**:198. doi:10.1093/JAMIA/OCW042
- 50 Xiao C, Choi E, Medical JS-J of the A, *et al.* Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *academic.oup.com*
- 51 Moullin JC, Sabater-Hernández D, Fernandez-Llimos F, *et al.* A systematic review of implementation frameworks of innovations in healthcare and resulting generic implementation framework. *Heal Res Policy Syst* 2015;**13**:16. doi:10.1186/s12961-015-0005-z
- 52 Shickel B, Tighe P, *et al.* Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *ieeexplore.ieee.org*
- 53 An Y, Huang N, Chen X, *et al.* High-risk Prediction of Cardiovascular Diseases via Attention-based Deep Neural Networks. *ieeexplore.ieee.org*
- 54 Ashfaq A, Sant'Anna A, Lingman M, *et al.* Readmission prediction using deep learning on electronic health records. *J Biomed Inform* 2019;**97**:103256. doi:10.1016/j.jbi.2019.103256
- 55 Avati A, Jung K, Harman S, *et al.* Improving palliative care with deep learning. *BMC Med Inform Decis Mak* 2018;**18**:122. doi:10.1186/s12911-018-0677-8
- 56 Mayampurath A, Sanchez-Pinto LN, Carey KA, *et al.* Combining patient visual timelines with deep learning to predict mortality. *PLoS One* 2019;**14**:e0220640. doi:10.1371/journal.pone.0220640

- 57 Ayala Solares JR, Diletta Raimondi FE, Zhu Y, *et al.* Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J. Biomed. Inform.* 2020;**101**:103337. doi:10.1016/j.jbi.2019.103337
- 58 Kwak GH-J, Hui P. DeepHealth: Deep Learning for Health Informatics. 2019.
- 59 Bakator M, Radosav D. Deep Learning and Medical Diagnosis: A Review of Literature. *Multimodal Technol Interact* 2018;**2**:47. doi:10.3390/mti2030047
- 60 Sheikhalishahi S, Miotto R, Dudley JT, *et al.* Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Med informatics* 2019;**7**:e12239. doi:10.2196/12239
- 61 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 2019;**25**:44–56. doi:10.1038/s41591-018-0300-7
- 62 Shickel B, Tighe PJ, Bihorac A, *et al.* Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J Biomed Heal informatics* 2018;**22**:1589. doi:10.1109/JBHI.2017.2767063
- 63 Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc* 2018;**25**:1419. doi:10.1093/JAMIA/OCY068
- 64 Miotto R, Wang F, Wang S, *et al.* Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* 2018;**19**:1236–46. doi:10.1093/BIB/BBX044
- 65 Holzinger A, Langs G, Denk H, *et al.* Causability and explainability of artificial intelligence in medicine. *WIREs Data Min Knowl Discov* 2019;**9**.

doi:10.1002/widm.1312

- 66 Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 2018;**6**:52138–60.
doi:10.1109/ACCESS.2018.2870052
- 67 Gilpin LH, Bau D, Yuan BZ, *et al.* Explaining Explanations: An Overview of Interpretability of Machine Learning. 2018.
- 68 Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. 2019;**14**.
- 69 Choi E, Bahadori MT, Sun J, *et al.* RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. 2016;:3504–12.
- 70 Kwon BC, Choi M-J, Kim JT, *et al.* RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. Published Online First: 27 May 2018. doi:10.1109/TVCG.2018.2865027
- 71 Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating Activation Differences. *34th Int Conf Mach Learn ICML 2017* 2017;**7**:4844–66.
- 72 Mallya S, Overhage M, Bodapati S, *et al.* SAVEHR: Self Attention Vector Representations for EHR based Personalized Chronic Disease Onset Prediction and Interpretability. 2019.
- 73 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267-70.
doi:10.1093/nar/gkh061

- 74 documentation:cdm [Observational Health Data Sciences and Informatics].
<https://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm> (accessed 20 Dec 2019).
- 75 Data-Driven | The National Patient-Centered Clinical Research Network.
<https://pcornet.org/data-driven-common-model/> (accessed 20 Dec 2019).
- 76 Rasmy L, Tiriyaki F, Zhou Y, *et al.* Representation of EHR data for predictive modeling: a comparison between UMLS and other terminologies. *J Am Med Inform Assoc* 2020;**27**. doi:10.1093/jamia/ocaa180
- 77 UMLS Knowledge Sources: File Downloads.
- 78 2018-ICD-10-CM-and-GEMs. 2017.
- 79 PheWAS - Phenome Wide Association Studies.
<https://phewascatalog.org/phecodes> (accessed 13 Mar 2019).
- 80 Beta Clinical Classifications Software (CCS) for ICD-10-CM/PCS.
<https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp> (accessed 13 Mar 2019).
- 81 Elixhauser A, Steiner C, Palmer L. Clinical Classifications Software (CCS). 2015.
- 82 Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses.
https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp (accessed 30 Mar 2020).
- 83 Bommasani R, Hudson DA, Adeli E, *et al.* On the Opportunities and Risks of Foundation Models. 2021.
- 84 Rasmy L, Xiang Y, Xie Z, *et al.* Med-BERT: pretrained contextualized

- embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med* 2021;**4**. doi:10.1038/S41746-021-00455-Y
- 85 Brown DW, DeSantis SM, Greene TJ, *et al.* A novel approach for propensity score matching and stratification for multiple treatments: Application to an electronic health record–derived study. *Stat Med* 2020;**39**:2308–23. doi:10.1002/SIM.8540
- 86 Rasmy L, Wu Y, Wang N, *et al.* A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. *J Biomed Inform* 2018;**84**:11–6. doi:10.1016/J.JBI.2018.06.011
- 87 Nigo M, Rasmy L, May SB, *et al.* Real World Long-term Assessment of The Efficacy of Tocilizumab in Patients with COVID19: Results From A Large De-identified Multicenter Electronic Health Record Dataset in the United States. *Int J Infect Dis* Published Online First: 29 September 2021. doi:10.1016/J.IJID.2021.09.067
- 88 Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry DRAFT GUIDANCE.
- 89 Choi E, Bahadori MT, Kulas JA, *et al.* RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. *Adv Neural Inf Process Syst* 2016;;3504–12.
- 90 Rasmy, L., Zhu, J., Li, Z., Tran, HTN., Wu, Y., Zhou, Y., Tiryaki, F., Xiang, Y., Xu, H. and Zhi, D. 2018. Medinfo 2019 (podium abstract submitted Nov 2018).

- Simple Recurrent Neural Networks is all we need for clinical events predictions using EHR data. In: *MedInfo 2019*. 2019.
- 91 Rasmy L, Zhu J, Li Z, *et al*. Simple Recurrent Neural Networks is all we need for clinical events predictions using EHR data. Published Online First: 3 October 2021. doi:10.13140/rg.2.2.13199.51368
 - 92 Shahriari B, Swersky K, Wang Z, *et al*. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc IEEE* 2016;**104**:148–75. doi:10.1109/JPROC.2015.2494218
 - 93 Rasmy L, Nigo M, Kannadath BS, *et al*. CovRNN—A recurrent neural network model for predicting outcomes of COVID-19 patients: model development and validation using EHR data. *medRxiv* 2021;:2021.09.27.21264121. doi:10.1101/2021.09.27.21264121
 - 94 Introduction · Captum. <https://captum.ai/docs/introduction> (accessed 20 Jul 2020).
 - 95 Poerner N, Roth B, Schütze H. Evaluating neural network explanation methods using hybrid documents and morphological agreement. 2018.
 - 96 Lundberg SM, Allen PG, Lee S-I. A Unified Approach to Interpreting Model Predictions. 2017.
 - 97 Wang C, Wang X, Ma K-L. Visual Summary of Value-level Feature Attribution in Prediction Classes with Recurrent Neural Networks. 2020.
 - 98 Kastner M, Estey E, Hayden L, *et al*. The development of a guideline implementability tool (GUIDE-IT): a qualitative study of family physician perspectives. *BMC Fam Pract* 2014;**15**:19. doi:10.1186/1471-2296-15-19

- 99 Shiffman RN, Dixon J, Brandt C, *et al.* The GuideLine Implementability Appraisal (GLIA): development of an instrument to identify obstacles to guideline implementation. *BMC Med Inform Decis Mak* 2005;**5**:23. doi:10.1186/1472-6947-5-23

