QATAR UNIVERSITY

COLLEGE OF ENGINEERING

INTERPRETABLE DEEP LEARNING MODELS FOR PREDICTION OF CLINICAL OUTCOMES FROM ELECTRONIC HEALTH RECORDS

BY

RAWAN TAYSEER ALSAAD

A Dissertation Submitted to

the College of Engineering

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Engineering

June 2022

COMMITTEE PAGE

The members of the Committee approve the Dissertation of

Rawan AlSaad defended on 29/11/2021.

Prof. Qutaibah Malluhi
Dissertation Supervisor

Prof. Serkan Kiranyaz
Committee Member

Dr. Tamer Elsayed
Committee Member

Dr. Halima Bensmail
Committee Member

Approved:

Khalid Kamal Naji, Dean, College of Engineering

ABSTRACT

ALSAAD, RAWAN, T., Doctorate : June: 2022, Doctor of Philosophy in Engineering

Title: Interpretable Deep Learning Models for Prediction of Clinical Outcomes from Electronic Health Records

Supervisor of Dissertation: Prof. Qutaibah Malluhi.

The rapid adoption of electronic health records (EHRs) has generated tremendous amounts of valuable clinical data on complex diseases and health trajectories. Yet, achieving successful secondary use of this EHR data for expanding our knowledge about diseases, expediting scientific discoveries in medicine, and facilitating clinical decision-making has remained challenging, owing to the complexity and data quality issues of these EHR data. Artificial intelligence, specifically deep learning, presents a promising approach for analyzing this rich EHR data, represented as a series of time-stamped multivariate data packed in irregular intervals. Deep learning-based predictive modeling with longitudinal EHR data offers a great promise for accelerating personalized medicine, enabling disease prevention, better informing clinical decision-making, and reducing healthcare costs. However, employing deep learning on EHR data for personalized prediction of clinical outcomes requires coping with numerous issues simultaneously.

In this thesis, we focus on addressing three important challenges: data heterogeneity, data irregularity, and model interpretability. We utilize state of the art deep learning techniques and modern machine learning methods to develop accurate and interpretable predictive models using EHR data. Specifically, we demonstrate how temporal clinical

data contained in EHRs can be harnessed for providing patient-specific predictions and interpretations for several clinical outcomes. We focus on two aspects: 1) code-level and visit-level interpretations for predicted outcomes using recurrent neural networks (RNNs), attention mechanism, and contextual decomposition interpretation method, and 2) leveraging the non-stationarity characteristics in EHR data into the predictive models using self-attention mechanism and kernels approximation technique.

Our proposed EHR-based deep learning models demonstrate improved performance in terms of predictive accuracy and interpretability on multiple clinical prediction tasks, compared to existing work in this area. These tasks include preterm birth prediction, school-age asthma prediction, and predicting the set of diagnosis codes in the next visit. Such models have a great potential to assist healthcare professionals in making decisions, which are not only dependent on the clinician's clinical knowledge and expertise, but also based on personalized and precise insights about future patient outcomes.

DEDICATION

To my parents, my eternal source of inspiration and strength.

ACKNOWLEDGMENTS

First, I would like to express my sincere gratitude to my research advisor, Prof. Qutaibah Malluhi, for his generous support, guidance, patience, and encouragement throughout my PhD journey. He taught me a lot from critical thinking, motivating new problems, scientific evaluation, presentation organization to seeing the big picture. His irreplaceable encouragement and supervision are the main reasons of the successful outcomes of my research.

I would like to express my special appreciation and thanks to Dr. Sabri Boughorbel, for encouraging my research and for allowing me to grow as a research scientist. His advice on both research as well as on my career have been invaluable.

I am grateful to Prof. Ibrahim Janahi, for the fruitful clinical discussions and for providing insightful comments and helpful feedback.

I am very appreciative of the assistance provided by Sidra Medicine team in facilitating access to the data and infrastructure used in this dissertation.

Finally, I am deeply indebted to my family. My mother and father, whose unconditional love and support made me who I am today. My siblings for always being there for me. My husband, whose endless support made the completion of my dissertation possible. My lovely daughters, Leen, Dana, and Raya who are the pride and joy of my life. You have made me stronger, better and more fulfilled than I could have ever imagined.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The accelerated adoption of electronic health records (EHRs) in modern healthcare systems generates massive amounts of clinical data in a readily computable form. This, in turn, offers great potential for making meaningful secondary use of clinical data by coupling it with modern machine learning approaches for data-driven prediction of future clinical outcomes. Building accurate and interpretable machine learning methods for the prediction of patients' future clinical outcomes has a great promise to assist healthcare providers and patients in the navigation of ever-increasingly complex treatment decisions. Decisions, which are not only dependent on the clinician's clinical knowledge and experience, but also based on personalized and precise insights about future patient outcomes [1]. Therefore, naturally, modern machine learning techniques, combined with massive EHR data generated from healthcare organizations have potential to bring dramatic changes to the clinical practice by improving healthcare quality, reducing healthcare costs, and advancing medical research [2]–[4].

1.1. Motivation

Conventional predictive models of clinical decision support are heavily dependent on human involvement in the identification and extraction of medical variables and optimization of prediction models. This results in task-specific models and classifier-dependent features, limiting its generalizability across tasks and datasets [5]. Recently, deep learning techniques have emerged as an alternative to other machine-learning algorithms and conventional statistics in the medical field. Data-driven deep learning methods for healthcare applications demonstrated promising performance for various clinical prediction tasks, such as mortality prediction, admission/readmission prediction,

length-of-stay prediction, and phenotyping classification [6]–[12]. Deep learning has attracted considerable attention in healthcare for two reasons. First, deep learning models have impressively outperformed traditional machine learning methods in many tasks, with less manual, repetitive, and lengthy feature engineering [8], [12]. This has changed the paradigm of healthcare data analytics modeling, from expert-driven feature engineering to data-driven feature construction. Second, the digitization of healthcare systems resulted in large and complex datasets (e.g. longitudinal multivariate event sequences), which enable efficient training of complex deep learning models [12]. Deep learning is facilitated by neural networks which mimic human intelligence in increasingly sophisticated and independent ways. Recurrent neural networks (RNNs) are one class of deep learning, where connections between the neurons form a directed graph along a temporal sequence, enabling it to consolidate the entire longitudinal EHR to generate accurate predictions for several clinical tasks [13]–[19].

However, using deep learning algorithms on EHR data has a number of challenges. First, majority of these models aggregates medical codes into visit representation without accounting for codes heterogeneity, resulting in inadequate visit representation. For example, the same diagnosis might indicate different conditions when combined with different medications or procedures. Second, the irregular temporal structure and dependencies among visits is not properly modeled, as visits are fed into the deep learning models, such as RNNs, in a sequential order ignoring potentially valuable temporal relationships among visits. Third, clinicians often consider deep learning models as highly expressive black-boxes which are difficult to understand and trust their predictions, limiting its adoption in real-world healthcare applications [20], [21]. This is crucial in the medical practice, where interpretable deep learning models can empower healthcare

experts to make informed, evidence-based recommendations for adequate interventions. Therefore, there is a pressing and growing need for efficient deep learning predictive techniques which can appropriately model EHR data and balance the trade-offs between model accuracy and interpretability.

1.2. Problem Statement

Our problem of predicting clinical outcomes using EHR data can be formulated as a multivariate time series forecasting problem, which involves the design and development of predictive models on data that comprises ordered relationships between its observations. The models use past structured observations to predict one or more possible future observations. This presents a classification problem, where the input variables are endogenous, i.e., it is influenced by other variables in the system (including themselves) and the output variable depends on them. In addition, it includes discontiguous time series, where observations are not uniform over time. The lack of uniformity of the observations is a feature of our problem, where data is collected in an unscheduled fashion (i.e., because the physician orders care, or the patient seeks care), at varying time intervals. Finally, the predictions produced by our models need to be accompanied with interpretation mechanisms capable of explaining the predicted results and highlighting evidence from patient's EHR timeline. The high-dimensional longitudinal data encompassed in EHRs jointly provide a rich representation of patient trajectories in the healthcare process. Yet, when EHR data is used for predictive modeling, researchers have to deal with various significant challenges arising from the representation of EHR temporal data. We elaborate on these challenges in this section.

1.2.1. Model Interpretability in Temporal EHR

Despite the remarkable performance of deep learning algorithms, the implications of using deep learning-based predictive models in healthcare have been limited owing to the lack of interpretability of these models. An interpretable deep learning model is a model which can provide explanations concerning why particular predictions are made. In healthcare, it is often insufficient to merely provide traditional machine learning metrics like area under the curve (AUC), sensitivity, or specificity. Instead, those metrics must be complemented with explanations about predicted outcomes. For example, an algorithm that diagnoses pneumonia but cannot explain why a patient is diagnosed with this disease, is less likely to be valued and trusted by clinicians and patients than an algorithm that can explain its reasoning for a particular predicted outcome. Currently, majority of predictive machine learning systems in healthcare generate predictions without explanations. However, in practice, medical practitioners frequently experience use cases which require adequate reasoning convincing them to use feedback from such models. In order to be truly considered, predictive models should be accompanied with interpretability mechanisms which are (1) able to explain evidence in patient's EHR timeline, (2) applicable with respect to the targeted use case, and (3) comprehensible to the potential user (healthcare provider) from a domain perspective. Thus, there is an important need for incorporating accurate interpretation mechanisms into prediction models in healthcare, to provide explanations about why it is making a certain prediction or giving a recommendation.

1.2.2. Data Heterogeneity in Temporal EHR

Electronic health records capture rich heterogeneous data about patients' history, including demographics, vital signs, diagnosis, medications, procedures, immunizations, laboratory and radiology test results, and treatment plans. Every time a patient visits a hospital or healthcare facility, this information is recorded. As a result, multivariate learning on EHR data presents a challenging task due to the high degree of dimensionality of the data, which is collected from multiple sources. For example, a standard coding system of diagnosis, the International Classification of Diseases (ICD-10-CM), contains 68,000 codes, and a standard coding system of procedures (ICD-10-PCS) contains 87,000 codes [22]. Efficient predictive models need to incorporate appropriate mechanisms capable of identifying which modalities are most important in an individual patient for a particular outcome at a specific time. Assigning the appropriate weight for each element in the data modalities has a significant role in boosting the performance of the predictive models. Therefore, developing predictive models which integrate data from all available EHR sources and focus on learning rich representations of the patients to predict specific clinical outcomes remains a challenging research problem.

1.2.3. Data Irregularity in Temporal EHR

In a typical multivariate time-series classification problem, an object is described by multiple time series of similar length that are measured at the same time point and at equal time intervals (multivariate time series). However, this simplified assumption does not always hold in the healthcare domain. Electronic health records include clinical events which are not evenly distributed over time and data are collected in an irregular

sampling rate. This poses a challenge for conventional machine learning algorithms, which often fail to capture the complex temporal dependencies in the EHR data and lead to the loss of potentially valuable sequential information. Although RNNs have achieved cutting edge performances in many prediction tasks in healthcare, like most other sequence models, they do not account for the time span between events in a patient's EHR. Thus, they mainly capture sequential signals rather than temporal patterns, which hinders their ability to fully leverage the complex dependencies across multiple time series in EHRs. These challenges demand for accurate temporal modeling methods which can account for temporal dependencies in the EHR multivariate time series data while training the predictive models. Therefore, how to better consider the temporal dimension of the clinical EHR data remains an important research question.

1.3. Research Objectives

In this thesis, we aim to utilize state of the art deep learning techniques and modern machine learning methods to develop efficient and interpretable predictive models. In addition, we demonstrate how longitudinal clinical data contained in electronic health records can be harnessed for providing patient-specific predictions and interpretations for multiple clinical prediction tasks. These tasks include preterm birth prediction, school-age asthma prediction, and predicting the set of diagnosis codes in the next visit.

1.4. Thesis Contributions

This thesis contributions are described from a machine learning and clinical informatics perspectives in the following subsections. Figure 1.1 provides an overview of the thesis contributions, presented in Chapter 3, Chapter 4, and Chapter 5.

1.4.1. Contributions in Machine Learning

The machine learning contributions of the thesis are:

- We introduce a code-level attention-based recurrent neural network (RNN) model, which models variables (medical codes) stored in EHRs to accurately predict and interpret the risk of clinical outcomes. The proposed model is a simplified version of the RETAIN (Reversed Time Attention Model) architecture [23]. Our method employs RNNs to model the longitudinal patient's EHR visits and exploits a single code-level attention mechanism to improve the predictive performance, while providing temporal code-level and visit-level explanations for the predicted outcomes. Experimental evaluation demonstrates the effectiveness of our proposed approach which outperformed the original RETAIN model.

- We develop an interpretation method based on the contextual decomposition of bidirectional long short-term memory networks (BiLSTMs), without any changes to the underlying model, to identify the contributions of individual medical visits as well as combinations of visits to the predicted clinical outcome. This development beyond individual visit importance is key for understanding a model as complex and highly non-linear as BiLSTM.

- We describe a new technique for modelling the non-stationary temporal relationships in EHRs, for prediction tasks. Our approach is based on extending the self-attention mechanism to handle the irregular time intervals between consecutive visits in a patient's EHR. This work demonstrates the feasibility of jointly learning the temporal non-stationary structure of EHR while performing supervised prediction tasks on EHR data.

- We demonstrate that integrating attention mechanisms into standard RNN models can improve the models prediction performance, mitigate the interpretability problems, and enhance the modelling of temporal trajectories of EHR time-series data.

1.4.2. Contributions in Clinical Informatics

The clinical informatics contributions of the thesis can be summarized as follows:

- We introduce the PredictPTB model to predict the risk of preterm birth at 1,3,6, and 9 months prior to its occurrence, using a real EHR dataset of more than 222,000 pregnancies. The model is complimented by an attention-based interpretation mechanism, providing interpretation at code-level and visit-level. The data modalities include diagnoses, medications, procedures, and lab orders.

- We present a BiLSTM-based model to predict school-age asthma for pre-school children with respiratory system-related complications, without interacting with the patients or collecting information beyond patients' EHRs. The predictive model incorporates a contextual decomposition-based interpretation method, which provides explanations of the predicted outcome via importance scores of individual visits as well as combinations of visits. This method is evaluated on a real EHR dataset of more than 11,000 children with respiratory system-related symptoms.

- We propose a model to predict the next visit information. Specifically, given the historical visit records of a patient, the model predicts the set of diagnosis codes present in the patient's next visit. The model takes into consideration the temporal dimension of the data, which increases its accuracy in predicting the

next diagnosis codes, compared to other non-temporal models. We demonstrate the advantages of this method on the task of predicting the next diagnosis code, using a real EHR dataset of 11451 patients.

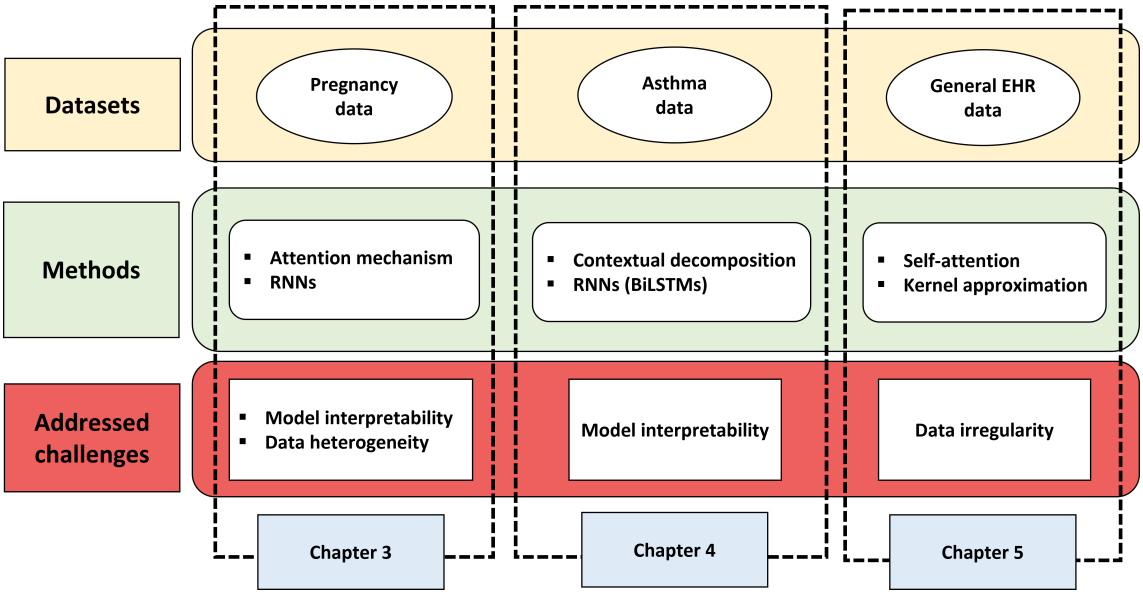


Figure 1.1. Overview of the research conducted in this thesis.

1.5. Thesis Organization

This thesis is organized as follows. Chapter 2 presents background and related work covering main relevant areas and introduce some basic terminologies used in this thesis. Chapter 3 introduces our code-level attention-based recurrent neural network predictive model. Chapter 4 presents a new interpretation method based on contextual decomposition of bidirectional long short-term memory networks (BiLSTMs) to characterize the contributions of individual medical visits as well as combinations of visits to the predicted clinical outcomes. Chapter 5 describes a method for modeling the temporal trajectories of EHRs using self-attention and non-stationary kernel approximation. Chapter 6 discusses the main findings of the work proposed in this thesis. Chapter 7

highlights a summary of this dissertation as well as discusses directions for future work.

1.6. Publications Resulting from Thesis

In the process of completing this work, the following publications have been submitted or published:

- AlSaad, R., Malluhi, Q., Boughorbel, S. PredictPTB: an interpretable preterm birth prediction model using attention-based recurrent neural networks. BioData Mining 15, 6 (2022). [24]

- AlSaad, R., Malluhi, Q., Janahi, I., Boughorbel, S. Interpreting patient-Specific risk prediction using contextual decomposition of BiLSTMs: application to children with asthma. BMC Med Inform Decis Mak 19, 214 (2019). [25]

- AlSaad, R., Malluhi, Q., Janahi, I., Boughorbel, S. Predicting Emergency Department Utilization Among Children with Asthma Using Deep Learning Models (submitted).

CHAPTER 2: BACKGROUND AND RELATED WORK

The thesis explores the application of deep learning-based algorithms for predictive modeling of clinical outcomes using temporal EHR data. To recognize the challenges associated with the use of EHR data for predictive modeling in healthcare and the importance of deep learning techniques in addressing these challenges, we present background and related work covering main relevant areas and introduce some basic terminologies used in this thesis.

2.1. Artificial Intelligence in Healthcare

Over the past decade, artificial intelligence (AI), including machine learning (ML) and more specifically deep learning (DL), has accomplished significant developments in the analysis of healthcare data [26], [27]. Artificial intelligence is an umbrella term that was first proposed in the 1950s. Essentially, it refers to a broad field within computer science dealing with problems related to machines that mimic human intelligence through sensing, reasoning, acting, and adapting [28]. Machine learning is a subset of AI at the intersection of statistics and data mining, which uses probability to make decisions or predictions about data with a reasonable degree of certainty [29]. Deep learning is a class of machine learning which extends the machine learning capabilities across multilayered neural networks where the level of abstraction increases gradually by non-linear transformations of massive amounts of input data [27].

Traditional ML techniques (e.g., logistic regression, random forest, support vector machines, and decision trees) may be used for problems which include hundreds or thousands of variables. However, extensive manual pre-processing effort is often essential to prepare the input data for the model, requiring human intervention of domain experts

to determine the hierarchy of features to recognize the differences between data inputs. Therefore, these techniques can be insufficiently powerful for healthcare risk prediction tasks, which usually involves large amounts of data and complex non-linear relationships. In contrast, deep learning models are fundamentally leveraged for more complex use cases in healthcare which require processing of huge volumes of clinical, genomic, and imaging data, such as imaging analytics and diagnostics [30]–[33], drug discovery [34], electroencephalograms (EEGs) and electrocardiograms (ECGs) [35], [36], clinical natural language processing [37], [38], precision medicine [39], and clinical decision support and predictive analytics [2], [8], [23], [25], [40]–[44].

In light of the above advantages for using deep learning models for healthcare applications, deep learning techniques hold a great potential for research in artificial intelligence for disease risk prediction and clinical decision support using electronic health records (EHRs). However, currently, many deep learning tools still struggle with the task of recognizing significant clinical features, learning meaningful correlations between them, and translating those relationships into actionable information which can assist healthcare providers in making informed decisions [27].

2.2. Challenges in Secondary Usage of Electronic Health Record Data

The large-scale adoption of electronic health records (EHRs) generates abundant amount of clinical data world wide and its volume is undergoing a rapid growth. An electronic health record (EHR) is a clinical data repository which stores patient-level clinical and administrative data in a structured format (e.g. diagnoses, medications, laboratory results, and vital signs) as well unstructured formats (e.g. imaging data, clinical notes, radiology reports, and discharge summary) [45].

Data contained in EHR is vital for many purposes related to patient care. The primary use of the health records is linked directly with providing individual patient's care services. The secondary use concerns the reuse of clinical data for a different purpose than the one for which it was originally collected, such as clinical and translational research, public health monitoring, automated disease surveillance, health system planning, education, regulation, and quality control [46]–[48].

Existing literature addressing the secondary use of EHR data is large and expanding rapidly [49], [50]. Using EHR data for secondary usages involves several major difficulties [46], [47]. First, **data inconsistency** may arise given that the data is collected by numerous individuals, using different technologies, and at different locations. A large number of individuals are involved in entering, reviewing, and maintaining the data. Consequently, the data includes different definitions, standards, terminologies, and units. Inconsistent data may lead to complicated and inaccurate data analysis and incorrect results. Second, since EHR data are not collected specifically for research purposes, they are subject to substantial **data incompleteness**. Therefore, the availability of an electronic record for a given patient does not necessarily imply that the record captures adequate information for a given research task. Data incompleteness can occur for several reasons. For example, because a patient sought care outside of the EHR's health care system, did not seek treatment for the illness, or the healthcare provider did not enter the information [51]. As a result, EHR data includes considerable missing data, which if not addressed properly, could impact the reasonableness of the derived conclusions and result in significant bias [52]. Third, the **security and privacy** issues in healthcare data have long been debated in the context of healthcare technologies, including electronic databases [53], [54]. These issues are primarily linked with the

exchange and sharing of EHR data, which was originally collected for the purpose of patient individual care, among different stakeholders and interconnected networks. The work in [55] provides a comprehensive review of all the relevant concerns and challenges associated with privacy and security features of EHR. Several security features were initiated by the Health Insurance Portability and Accountability (HIPAA) Act [56]. HIPAA compliance is the process that healthcare providers follow to protect and secure protected health information (PHI) [57]. In addition, several privacy-preserving techniques have been proposed for EHRs [58]–[61]. However, such techniques may have an impact on the utility (e.g. predictive power) of EHR data.

These difficulties associated with secondary use of EHR data continue to be the subject of research, and developing solutions is becoming increasingly interdisciplinary [62]–[64].

2.3. Deep Learning for Electronic Health Record

Despite the massive growth in size and diversity of clinical data from EHRs, a recent systematic review of the medical literature [65] found that risk prediction techniques built with EHR data utilize a very limited number of variables (median of 27 variables) and rely on simple and traditional machine learning and statistical techniques such as logistic regression, random forest, and support vector machines (SVM) [66]. While the simplicity and interpretability of such statistical models are desirable for medical applications, their limitations in dealing with high-dimensional data, limited scalability and generalizability, and dependence on manual feature engineering hinder their use for comprehensive analyses of rich EHR data to discover hidden patterns that might characterize a medical condition or a disease progression [67].

In view of the limitations of traditional machine learning approaches in dealing with the challenging characteristics of EHR data, recently, deep learning models are becoming increasingly important in effectively discovering and taking into consideration the complex nonlinear interactions among the high-dimensional, temporal, and multi-modal variables stored in EHRs, for a wide spectrum of clinical informatics tasks [2], [8], [23], [43], [44], [68], [69]. Deep learning techniques can address several technical challenges present in EHR data, including data heterogeneity, irregularity, and sparsity, among others [67]. Deep learning algorithms identify optimal features from the data itself, without the need for human involvement, allowing for the automatic learning of hidden non-linear data correlations that might otherwise be missed or unused [70], [71]. Several deep learning-based methods have been proposed for modeling the predictive patterns in longitudinal EHR data such as recurrent neural networks (RNNs) [72], [73], long short-term memory (LSTM)[74], [75], and gated recurrent unit (GRU) [76]. Such developments in deep learning techniques have been employed in several clinical informatics research problems to address many EHR challenges and unlock the rich information in the EHR [77]–[79].

2.4. Interpretability of Deep Learning Prediction Models in Healthcare

In the medical context, interpretability of deep learning models is essential to allow healthcare experts to make informed, reasonable, and data-driven decisions. Therefore, the difficulty of explaining the deep learning models is currently among the major barriers in the adaptation of such powerful algorithms within the healthcare systems [80]. Although deep learning models offer superior predictive performance, they compromise model explainability and transparency, which is rarely accepted in the healthcare

practice [81]. This is specifically important in the medical settings where physicians are expected to justify their decisions with causal relationships, in accordance with the domain knowledge. Therefore, there is an important need for a human understandable, interpretable, and robust explanations of the resulting outcomes of DL models [82], [83].

The interest in interpretability of deep learning models resulted in several methods [84]–[88] for interpreting results of DL algorithms, sharing a common objective of understanding why a network makes a prediction. Among these methods are Local Interpretable Model-agnostic Explanations (LIME) [89], Shapley Additive Explanations (SHAP) [90], and attention mechanism [91]. These three methodologies share the idea of providing interpretability in the form of feature importance. With LIME and SHAP, feature importance is measured via simulations that modify the data after model training is completed. In contrast, in attention mechanisms, feature importance is computed during the model training, thus improving not only the interpretability of the model but also the performance of the model, by paying more attention to important features.

While none of these interpretability techniques are perfect, they can be used to interpret the results of simple and complex machine learning models [92]. Specifically, we describe here the Shapley Additive explanations (SHAP) technique [90]. The concept of the SHAP method is inspired by the game-theory and is based on computing the contribution score for each feature for individual predictions. A prediction can be explained by assuming that each feature value of the instance is a "player" in a game and the contribution of each player is computed by including and excluding the player from all subsets of the rest of the players. In SHAP, the authors first describe the class of additive feature attribution methods, which unifies six current methods, including LIME

[89], Layer-Wise Relevance Propagation [93] and DeepLIFT [94], which all use the same explanation model. Then, they suggest SHAP values as a unified measure of feature importance that maintains three necessary properties: local accuracy, missingness, and consistency. Finally, they describe several different methods for estimating SHAP values, as well as experiments demonstrating not only the improved performance of these values in terms of distinguishing between different output classes, but also in terms of better aligning with human intuition, when compared to many other existing interpretability techniques.

Attention mechanism has shown to be a promising approach for interpreting DL models [82] and several attention mechanisms have been proposed to address the interpretability of DL models [23], [95]–[98].

2.5. Attention Mechanism

Recently, the attention mechanism has gained popularity in training neural networks [99], [100]. Motivated by the visual attention mechanism in humans, the attention component provides a neural network with the capacity to flexibly focus on a subset of its input (or features) to extract fine-grained informative representations, and thus facilitate global learning [101], [102]. Neural attention networks have also been effectively utilized in several natural language processing (NLP) tasks, such as text summarization [103], [104], machine translation [99], [105], and document classification [106], [107], where attention has been used as a textual level mechanism for modeling interactions within different parts of the text.

From a technical perspective, the concept of "attention" refers to the idea of information selection by weighting. Although attention mechanisms are always customized

for certain types of input data and the corresponding tasks, instead of having standard formulations, they share the same design principles [91]. These principles include (1) extracting the latent representations of the input data via deep neural networks, (2) calculate attention signals based on the latent representations, and then (3) assign attention signals to generate the weighted latent representations of input data.

Moreover, attention explanations are recently leveraged to open a new window for interpreting deep learning models in healthcare, by providing patient-specific attention weights on features to explain the predicted results [23], [98], [108], [109]. For example, in [108], the authors proposed the GRaph-based Attention Model (GRAM) for healthcare representation learning, which utilizes the attention mechanism to infuse information from medical ontologies into deep learning models and explains the attention behavior during prediction by displaying the attention weights of each node in the knowledge graph.

2.5.1. Self-attention Mechanism

Attention and self-attention fundamentally share the same concept and many common mathematical operations. Self-attention is an attention mechanism which relates every position in a sequence to every other position in the sequence, including itself, and reweighs the position embeddings of each position to include contextual relevance. [110]. Self-attention is computed using dot-product attention [91] over a query vector $\boldsymbol{Q}$, a key vector$\boldsymbol{K}$, and a value (representations) of events in a sequence vector $\boldsymbol{V}$, defined as:

$$Att(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}})\boldsymbol{V} \tag{2.1}$$

Self-attention depends on the positional encoding to identify and capture the order of the sequence, where the vector representation for each position is combined (added or concatenated) with the corresponding event embeddings, thus, providing the order context to the non-recurrent architecture of the self-attention mechanism [91]. It accomplishes this through a series of key, query, and value weight matrices via three steps: 1) dot product similarity of query-key pairs to find alignment scores, 2) normalization of the scores to get the weights, and 3) reweighing of the original embeddings using the weights for combining event values.

2.6. Electronic Health Record (EHR)

2.6.1. EHR Datasets

Large, accurate, and comprehensive datasets are necessary for deep learning research. Many EHR-based predictive modeling efforts have used EHR data from single healthcare facility to demonstrate the potential for using EHR-based predictive modeling. For example, the MIMIC-IV (Medical Information Mart for Intensive Care, version IV) dataset [111], is a publicly available single-center critical care database capturing information about patients admitted to ICUs of one hospital. The MIMIC-IV dataset has been used to predict several outcomes related to patients admitted to ICUs [112]–[115]. The Vanderbilt Synthetic Derivative (SD) [50] is another single-center EHR repository, which has been used to study several conditions [116], [117]. However, data from these single-center projects tend to have major limitations such limited number of patients and a potential deficiency of geographic and demographic diversity. A natural solution to these limitations is to integrate EHR data from multiple healthcare facilities with diverse demographics and multiple regions. The Cerner Health Facts Database (currently re-

ferred to as the Cerner Real World Data) maintains a large volume of multi-center EHR clinical data [118]. Health Facts data is de-identified and is HIPAA-compliant to protect the identity/privacy of patients as well as organizations. Health Facts was the main data resource in this thesis, offering a large, accurate, and comprehensive multi-center EHR data resource for training our deep learning models. It stores clinical records with time-stamped and sequenced information on diagnosis, prescriptions, lab orders, surgeries, among others.

2.6.2. EHR Representation

A patient's EHR usually consists of a sequence of visits (encounters) a patient has made to healthcare facilities, and each visit captures the list of medical codes documented by the healthcare practitioners (e.g. diagnoses, medications, lab orders, and procedures). The temporal models developed in this thesis require patient-level time-ordered data that has been collected over time. Therefore, we chose to present our EHR dataset in the form of list of lists of lists. The outermost list corresponds to patients, the intermediate list corresponds to the time-ordered visit sequence each patient made, and the innermost list corresponds to the medical codes that were documented within each visit.

In this thesis, we will use the following notation for representing our EHR dataset. Let $P = \{p_1, \ldots\ldots p_n\}$ be a dataset of n patients. Each patient p_j EHR is comprised of a sequence of T_j patient visits, $p_j = \{x_1, x_2, \ldots, x_{T_j}\}$, ordered by visit date $t \in \{1, T_j\}$, where the last time point T_j denotes the time on which the predicted clinical outcome for patient p_j has occurred. We express medical events in EHR as medical codes (e.g. diagnoses, medications, procedures, and lab orders), denoted as $\{c_1, c_2, \ldots, c_{|C|}\} \in C$, where C represents the entire set of unique medical codes. Each visit x_i can be expressed

as a binary vector $x_i \in \{0,1\}^{|C|}$, where the k-th element is set to 1 if the i-th visit contains the medical code c_k, otherwise it is set to 0.

2.6.3. Clinical Coding in EHR

Clinical coding is the process of translating the terminology used in healthcare into a coded form. It includes standardization of the terms (e.g. diagnoses, procedures, and medications) and the placing of the codes in a structured hierarchy, using a classification system [119]. These codes are used by healthcare providers, software developers, and researchers for a multitude of purposes in clinical care, clinical research, public health, and medical informatics [120]. Several different medical classification systems exist, which can be broadly classified into two main groups: statistical classifications and nomenclatures [121]. A statistical classification groups together similar clinical concepts and classify them into categories. The number of categories is limited so that the classification does not become too big. In contrast, nomenclature assigns a separate listing and code for every clinical concept. This results in a very large number of codes, making nomenclatures impractical for compiling medical data analysis [122]. In this thesis, we use statistical classification system, presented by several types of coding systems specific to health care data. For example, diagnostic codes are used to describe diseases, disorders, symptoms, morbidity, and mortality. We use the International Statistical Classification of Diseases and Related Health Problems (known as ICD) for diagnosis codes, specifically, the ICD-9-CM and ICD-10-CM systems. Procedural codes are numbers or alphanumeric codes used to identify specific health interventions taken by medical professionals. For procedure codes, we use two coding systems: the Current Procedural Terminology (CPT) and ICD-10-PCS (Procedure Classification System).

Pharmaceutical codes are used to identify medications. We use the National Drug Code (NDC) system, which is a unique 10-digit or 11-digit, 3-segment number, used as a universal product identifier for human drugs.

CHAPTER 3: INTERPRETABLE CODE-LEVEL ATTENTION-BASED RECURRENT NEURAL NETWORKS

3.1. Overview

In this chapter, we introduce a clinical code-level attention-based recurrent neural networks (RNNs) prediction model, which combines variables (medical codes) readily accessible through electronic health record to accurately predict the patient's risk for certain diseases. The architecture of our model is based on the previously published RETAIN (REverse Time AttentIoN) model [23], a two-level (visit-level and code-level) neural attention model to predict the risk of heart failure. However, our model exploits a single code-level attention mechanism to improve the predictive performance, while providing temporal code-level and visit-level explanations for the prediction results. We evaluated the performance of our model on the task of predicting preterm birth at 1,3,6, and 9 months prior to delivery, using routinely collected EHR data. We compared the performance of different combinations of prediction time-points, data modalities, and data windows. Our proposed model was able to predict preterm birth with an ROC-AUC of 0.82, 0.79, 0.78, and PR-AUC of 0.40, 0.31, 0.24, at 1, 3, and 6 months prior to delivery, respectively. As for data modalities, results demonstrated that combining all modalities together (diagnosis, medications, procedures, and lab orders) improved the PR-AUC by 10% compared to using the diagnosis modality alone. We also present a case-study of our model's interpretability at visit-level and code-level, which illustrates how clinicians can gain some transparency into the predictions.

3.2. Introduction

Attention mechanisms have been recently advocated to improve the accuracy as well as the interpretability of deep learning models. It was first introduced to improve the performance of the encoder-decoder RNN on machine translation [99]. Recently, attention mechanisms have accomplished considerable success in many prediction tasks in healthcare [23], [98], [108], [109]. Among these efforts, Choi et al. [23] proposed a model known as RETAIN, which uses a two-level neural attention model to predict heart failure using patient's temporal EHR data. The predictive performance or RETAIN is comparable to recurrent neural networks (RNNs), while providing explanations for the visit-level and code-level contributions to the final prediction results.

In parallel, the expeditious growth in size and diversity of clinical data from electronic health records (EHR) has attracted the utilization of this data to predict a wide spectrum of clinical outcomes. However, EHR resources have been largely unexploited in the study of pregnancy. In contrast to other clinical contexts, the clinical surveillance of pregnancy data and its outcomes take place in a well-defined time frame, based on gestational length. Hence, EHRs seem to be very appropriate for modeling pregnancy complications, including preterm birth. To this end, predictive modeling using attention mechanisms with EHR data is anticipated to provide accurate individualized predictions for expecting mothers threatened with preterm birth.

Preterm birth (PTB) is defined as a delivery that occurs before the start of the 37th week of pregnancy, as opposed to full-term birth which occurs anytime from 37 to 42 weeks of gestation [123]. Worldwide, more than 15 million babies, or about $10 - 15\%$ of all alive births, are born preterm every year [124], [125]. PTB accounts for over one

third of infant mortality, and babies born preterm are at increased risk of significant long-term morbidity and disability, such as cerebral palsy, neurological disorders, behavioral problems, developmental delays, and mental health conditions [126]–[130]. Therefore, identifying pregnancies at risk for PTB and accordingly providing adequate interventions can improve both short- and long-term outcomes for babies born preterm.

Majority of existing work on PTB prediction aims to identify risk factors of PTB through a hypothesis-testing methodology, under highly-controlled settings. A number of risk factors have been reported to increase the risk of PTB such as: previous preterm labor, multiple gestation (being pregnant with more than one baby), diabetes, complications with the cervix, uterus, or placenta, tobacco smoking, and infections [131]–[133]. However, women who have preterm delivery often have no known risk factors [134]. In addition, some of the predictors (such as prior PTB) does not apply for first-time mothers. As a result, machine learning is of great interest to better predict PTB and several studies have attempted to predict PTB using machine learning techniques on a set of pre-defined clinical risk factors [135]–[142], or leveraging diverse variables from electronic health record (EHR) data [143]. More recently, few studies have used deep learning techniques to predict PTB using ultrasound and MRI figures [144], [145], and high-dimensional EHR data [146], [147], with promising results.

However, most of these studies reports poor to moderate predictive performance ranging from 59% to 75% ROC-AUC, ignores the sequential or temporal trajectory of events recorded in EHR, mostly used few human-derived features disregarding the huge amount of information embedded in each patient's record, and rarely evaluated the model's performance across multiple time points throughout the pregnancy timeline. To date, we lack effective predictive models for PTB, with two important challenges

identified for deriving a PTB prediction model: (1) designing an accurate and scalable predictive model to handle the sequential high-dimensional EHR data, and being able to automatically select potential predictors from hundreds (if not thousands) of variables, and (2) complimenting these prediction models with reasonable interpretation mechanisms.

In this chapter, we propose the PredictPTB model, an interpretable code-level attention-based recurrent neural network model to predict the risk of preterm delivery using multiple sources of temporal data from EHRs. We use a large dataset of 222,436 deliveries, comprising a total of 27,100 unique clinical concepts, to demonstrate the predictive performance of the proposed model on the preterm birth prediction task compared to the original RETAIN model. We conduct a quantitative analysis to assess the effectiveness of the PredictPTB model among several prediction points, data windows, and data modalities. Finally, we qualitatively examine the interpretability of the PredictPTB model by visualizing the learned attention-based weights and against the attention-based scores learned by the RETAIN model. The contributions of this work can be summarized as follows:

- We propose a simplified version of the RETAIN architecture, where we employ RNNs to model the sequential patient's EHR visits, and exploit a single code-level attention mechanism to improve the predictive performance while providing temporal code-level explanations for the prediction results.

- We compare the performance of our model across different combinations of data modalities, prediction points, and data windows to find an optimal combination for preterm birth prediction. We further compare these combinations of our model

to other baseline models.

- We present a case-study of our model interpretability at visit-level and code-level, which illustrates how clinicians can gain some transparency into the predictions.

3.3. Methods

3.3.1. Problem Formulation

Let $P = \{p_1, \ldots\ldots p_n\}$ be a dataset of n patients. Each patient p_j EHR is comprised of a sequence of T_j patient visits, $p_j = \{x_1, x_2, \ldots, x_{T_j}\}$, ordered by visit date $t \in \{1, T_j\}$, where the last time point T_j denotes the time on which the delivery for patient p_j has occurred. We express medical events in EHR as medical codes (e.g. diagnoses, medications, procedures, and lab orders), denoted as $\{c_1, c_2, \ldots, c_{|C|}\} \in C$, where C represents the entire set of unique medical codes. Each visit x_i can be expressed as a binary vector $x_i \in \{0, 1\}^{|C|}$, where the k-th element is set to 1 if the i-th visit contains the medical code c_k, otherwise it is set to 0. Let $m = \{m_1, \ldots, m_p\}$ be the set of prediction time points.Given the EHR history $p_j = \{x_1, x_2, \ldots, x_{T-m}\}$ of each j^{th} patient up to time point $T - m$, our task is to predict the risk of a PTB at time point T, denoted as $\hat{y}_T^j \in \{0, 1\}$, and accurately interpret why a patient is predicted as PTB, using the patient's temporal EHR history. To address this problem, we introduce a code-level attention-based RNNs to provide an interpretable clinical risk prediction model for preterm birth.

3.3.2. Preliminaries on Attention Mechanism

Attention mechanism has been an important component in RNNs to capture long-term dependencies. It computes the dynamic weights representing the relative importance of the inputs in a sequence for a particular output. Figure 3.1(a) illustrates the architecture of a standard attention model, in which the attention mechanism summarizes the source sequence information in the encoder RNN hidden states (i.e., h_i), computes the dynamic attention scores for each visit v_i as α_i, and then multiplies the weights α_i with the input sequence v_i to weight the sequence. A single context vector c_i for a patient up to the i-th visit can then be calculated using the sum of the weighted vectors as:

$$c_i = \sum_{j=1}^{i} \alpha_j \odot v_j \tag{3.1}$$

3.3.3. Reversed Time Attention Model (RETAIN)

The RETAIN model was first introduced in [23] for the prediction of heart failure using patient's longitudinal EHR data. Given patient records, RETAIN can make accurate predictions, comparable to RNNs, while explaining how each medical code at each visit contributes positively or negatively to the final prediction score. RETAIN is based on a double-attention mechanism, which integrates two single-attention models (the visit-level attention RNN_α and the code-level attention RNN_β) to generate the patient representation, as illustrated in Figure 3.1(b). Using the computed attention weights at visit-level α and code-level β, the context vector c_i for a patient up to the i-th visit is calculated as:

$$c_i = \sum_{j=1}^{i} \alpha_j \beta_j \odot v_j \tag{3.2}$$

3.3.4. Architecture of the PredictPTB Model

The RETAIN architecture seems to have a redundant attention branch for capturing visit-level attentions, which are inherently available in the code-level attentions. Therefore, to construct a more precise contextual representation of each patient, we introduce the PredictPTB model. Our model simplifies the RETAIN architecture into a single code-level attention layer RNN_β. This approach reduces the complexity of the RETAIN architecture while improving the accuracy of the predictions due to 1) directly promoting the code-level information in each step of the model, 2) paying more attention to representative and discriminative features than other features, and 3) limiting the number of model parameters which reduces the risk of over-fitting and possible gradient flow. We use bidirectional RNN, specifically BiLSTM, which enables both future and past information to be accessible by the current state, providing more information about the input. This mimics the practice of a clinician examining a patient's EHR both forward and backward, trying to identify a set of weights representing the relative importance of patient's individual visits or codes within those visits.

The predictions of our proposed PredictPTB model are made using the steps described in Figure 3.2, as follows:

- **Step 1**: The model embeds a patient's visit sequence v_i as:

$$v_i = \sigma(W_{emb}x_i + b_x) \tag{3.3}$$

where $v_i \in \mathbb{R}^m$ is the embedding of $x_i \in \mathbb{R}^C$, $W_{emb} \in \mathbb{R}^{m \times C}$ is the embedding matrix, m is the embedding size across C medical variables , σ is a non-linear

activation function such as rectified linear unit (ReLU) or sigmoid, and b_x is the bias.

- **Step 2**: The embeddings are fed as inputs to a recurrent neural network RNN$_\beta$, which computes the attention-based contribution scores of individual medical variables and generate code-level attention weights β_i. Note that, in contrast to the RETAIN architecture, we eliminate the RNN$_\alpha$ layer and use a single attention layer RNN$_\beta$ to generate the weights, as follows:

$$\begin{aligned} h_i, h_{i-1}, \ldots, h_1 &= RNN_\beta(v_i, v_{i-1}, \ldots, v_1) \\ \beta_j &= tanh(W_\beta h_j + b_\beta) \quad \text{for } j = 1, \ldots, i. \end{aligned} \tag{3.4}$$

where $h_i \in \mathbb{R}^q$ is the hidden layer of RNN$_\beta$ at time step i, q is the hidden size of RNN$_\beta$, β_j is the attention weight for individual variables, $W_\beta \in \mathbb{R}^{m \times q}$ and $b_\beta \in \mathbb{R}^m$ are parameters to learn.

- **Step 3**: The computed attention weights are used to generate the patient representation context vector c_i as:

$$c_i = \sum_{j=1}^{i} \beta_j \odot v_j \tag{3.5}$$

where $c_i \in \mathbb{R}^m$

- **Step 4**: The predictions of our model can then be computed by linearly transforming the context vectors c_i using:

$$\hat{y}_i = Softmax(W c_i + b) \tag{3.6}$$

where $W \in \mathbb{R}^m$ and $b \in \mathbb{R}$ are the parameters to learn.

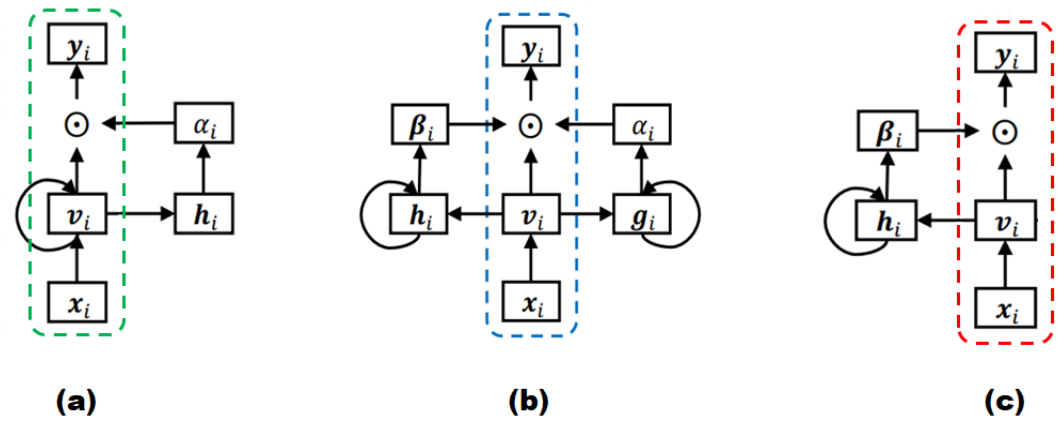


Figure 3.1. (a) Standard attention model, (b) RETAIN model, (c) PredictPTB model.

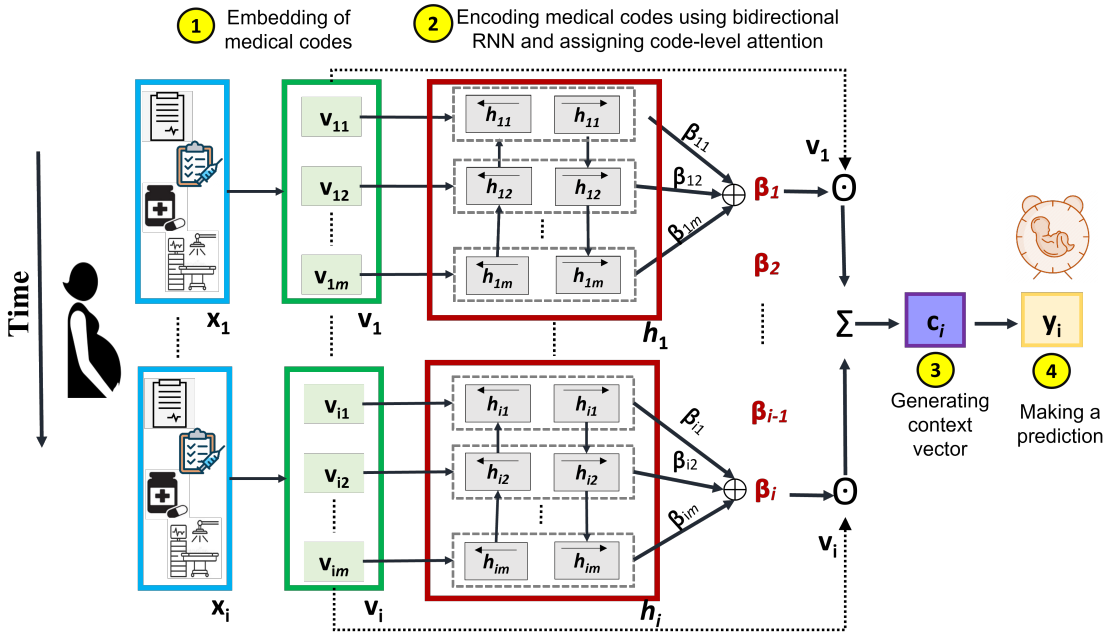


Figure 3.2. An overview architecture of our code-level attention model (PredictPTB).

3.3.5. Model Interpretation

To interpret predictions made by our PredictPTB model, we follow the interpretation approach described in [23]. Given a patient's list of visits $x_1, \ldots, x_i$, the probability of the binary output vector $y_i \in \{0, 1\}$ can be predicted as follows:

$$p(y_i \mid x_1, \ldots, x_i) = p(y_i \mid c_i) = Softmax(W(\sum_{j=1}^{i} \beta_j \odot v_j) + b) \tag{3.7}$$

$$p(y_i \mid x_1, \ldots, x_i) = Softmax(W(\sum_{j=1}^{i} \beta_j \odot \sum_{k=1}^{C} x_{j,k} W_{emb}[:, k]) + b) \tag{3.8}$$

$$= Softmax(\sum_{j=1}^{i} \sum_{k=1}^{r} x_{j,k} W(\beta_j \odot W_{emb}[:, k]) + b) \tag{3.9}$$

where $x_{j,k}$ is the k-th element of the input visit x_j. In order to compute the contribution ω of the k-th code at each visit x_j at time step $j \leq i$ for predicting y_i, we deconstruct Equation 3.9 into 3.10, where we exclude the index i of y_i in the β_j, as follows:

$$\omega(y_i, x_{j,k}) = \underbrace{W(\beta_j \odot W_{emb}[:, k])}_{\text{Contribution coefficient}} \underbrace{x_{j,k}}_{\text{Patient visit}} \tag{3.10}$$

In the real clinical practice, clinicians typically identify different weights on different visits and medical codes, as part of the diagnosis process. In this sense, the above contribution coefficient can be used to highlight important visits and medical codes.

3.3.6. Preterm birth: Data Modalities, Prediction Points, and Data Windows

Data Modalities

For each patient, we extracted data from multiple domains of sources in EHR to enrich the patient representation with multiple data modalities. The modalities include diagnosis, medications, procedures, and lab orders. The data of each modality

was represented as a concept in a set of standardized terminologies, including ICD10 and ICD9 for diagnosis, NDC and brand names for medications, CPT and ICD9 for procedures, and LOINC for lab orders. The numbers of unique medical concepts representing diagnosis, medications, procedures, and lab orders were 14795, 1332, 7640, and 3333, respectively.

Prediction Points

To further elaborate on the advantages of using the PredictPTB model, we design experiments to quantify the performance across different prediction time points during the pregnancy timeline. Using the delivery event as the reference point, we selected the following time points: 1, 3, 6, 9 months, before the delivery event. We refer to these prediction points as P1, P2, P3, and P4, as shown in Figure 3.3. For each prediction point, the patient EHR history up to the prediction point is used for the prediction. For example, if the prediction point is 3 months before delivery, only the EHR data up to the prediction point is used by the model, and the data between the prediction point and the delivery event is discarded. This presents real clinical scenario, where the physician needs to predict the risk of preterm delivery at different time points of the pregnancy timeline.

Data Windows

We used two setups for EHR data windows; long-term (full history) data window and short-term (pregnancy history) data window. In the long-term window setup, all EHR data of the patient up to the prediction point X is used to train the predictive models. In the short-term window setup, only the EHR data between the start of the pregnancy up

to the prediction point X is used, and previous history before pregnancy is discarded. These two data windows will help us assess the influence of the events happening during the pregnancy timeline on the predictive performance of the model, compared to the influence of all events included in the patient EHR history before and after the start of the pregnancy.

Figure 3.3. Prediction points used in the analysis

3.4. Experiments and Results

3.4.1. Dataset

To validate the predictive performance of the proposed model, we conducted experiments on a clinical EHR dataset, obtained from Cerner Health Facts database, with more than 222,000 deliveries in the United States between 2000-2017. We relied on the ICD-9 diagnosis codes to identify preterm and full-term pregnancies due to the lack of information about gestational age in the data. We leveraged the following ICD-9 codes to identify full-term deliveries: 645.xx, 649.8, 650 and 652.5, and used the ICD-9 code of 644.2x to identify preterm deliveries.

The medical record of each patient encompasses the following information: diagnoses, medications, procedures, and lab orders. Tables 3.1 and 3.2 describe the statistics

of the two cohort setups: long-term window and short-term window, respectively. Furthermore, to ensure that the RNN model has sufficient number of visits to train on, only patients who have at least two visits in their EHR were included.

Table 3.1. Cohort: Summary Statistics for Long-term Window Cohort

Cohort: Full History	**Total #pregnancies**	**Counts for each class**	**Mean age**	**Average #visits**
P1 (1 month)	222,436	Fullterm 204,700	28.03	12.19
		Preterm 17,736	28.76	13.37
P2 (3 months)	202,930	Fullterm 187,073	27.81	10.96
		Preterm 15,857	28.42	12.26
P3 (6 months)	177,253	Fullterm 163,641	27.45	10.1
		Preterm 13,612	27.84	11.51
P4 (9 months)	150,904	Fullterm 138,468	26.91	9.92
		Preterm 12,436	27.43	11.24

Table 3.2. Summary Statistics for Short-term Window Cohort

Cohort: Pregnancy History	**Total #pregnancies**	**Counts for each class**	**Mean age**	**Average #visits**
P1 (1 month)	168,932	Fullterm 155,647	28.15	6.96
		Preterm 13,285	29.08	7.07
P2 (3 months)	133,704	Fullterm 123,968	28.04	5.16
		Preterm 9,736	28.87	5.27
P3 (6 months)	77,779	Fullterm 73,477	27.95	3.33
		Preterm 4,302	28.56	3.28

3.4.2. Implementation Details

Models were trained on a DGX-1 server equipped with 8 NVIDIA Tesla V100 GPU accelerators. Models were trained for the task of predicting whether the expecting

mother will deliver a full-term or a preterm baby. We implemented PredictPTB using Tensorflow 2.0+ framework. For all models, patients were randomly split into training (70%), validation (10%), and test (20%) sets. The same proportion of preterm and full-term deliveries was maintained among the training, validation, and test sets. We performed hyper-parameter tuning on the RETAIN model using grid search, and the following parameters provided the best results: embedding size= 256, size of both RNN hidden layers= 256, batch size= 32. The following parameter values were used for the PredictPTB model: embedding size= 200, RNN hidden size= 200, batch size= 32. To help conserve GPU RAM, we set the maximum number of visits for a patient to 200 visits. For patients with more than 200 visits, the most recent 200 visits will be used.

3.4.3. Evaluation Measures

The performance of the models is reported on the test set, and the following evaluation measures were used:

- **ROC-AUC** (Area Under the Receiver Operating Characteristic Curve).The ROC-AUC curve plots the sensitivity (true-positive rate) against 1 – specificity (false-positive rate) for consecutive cut-offs for the predicted risk.

- **PR-AUC** (Area Under the Precision-Recall Curve). The PR-AUC curve is a plot of the precision (y-axis) and the recall (x-axis) for for consecutive cut-offs for the predicted risk.

  - Precision is a metric that quantifies the number of correct positive predictions made. It is calculated as the number of true positives divided by the total number of true positives and false positives.

Precision = True Positives / (True Positives + False Positives)

- Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made. It is calculated as the number of true positives divided by the total number of true positives and false negatives (e.g. it is the true positive rate).

  Recall = True Positives / (True Positives + False Negatives)

The PR-AUC curves are well-suited for imbalanced settings, where the focus of the PR curve on the minority class makes it an effective diagnostic for imbalanced binary classification models [148]. The ROC-AUC and PR-AUC plots are summarized with an area under the curve score which is used to compare the classification models.

- **Sensitivity** refers to a model's ability to designate an individual with preterm delivery as positive. A highly sensitive model means that there are few false negative results, and thus fewer cases of preterm are missed.

- **Specificity** refers to the model's ability to designate an individual who does not have a preterm delivery as negative. A highly specific model means that there are few false positive results.

3.4.4. Baselines

Our PredictPTB model was evaluated against two baselines: the Multi-Layer Perceptron (MLP) and the RETAIN models. for the MLP model, we combine features extracted from all the visits for a patient into a single feature vector. To do this, we use the counts for each medical code in the patient's list of visits. The resulting vector

was used to train the MLP model which has a single hidden layer of size 256 between the input and output, drop-out rate 0.6 on the output of the hidden layer, and 0.0001 L2 regularization coefficient for the hidden layer weight.

3.4.5. Results

We present the results of preterm birth prediction on the EHR dataset by both the baselines and our PredictPTB model. All models were trained at four prediction time points 1 month (P1), 3 months (P2), 6 months (P3), and 9 months (P4) before the delivery, and on two data history setups: long-term and short-term windows.

Prediction Performance Across Prediction Time Points and Data Windows

The objective of this experiment is to compare how models trained using combined data modalities perform in different scenarios, represented by four prediction time points and two data history setups. Figures 3.4 and 3.5 show that our PredictPTB model consistently provided more accurate predictions, as compared to the RETAIN and MLP models, for majority of the prediction points. Models achieved the best performance at 1 month before delivery, using the short-term window history, where our PredictPTB model performed better than the baselines with an PR-AUC and ROC-AUC of 40.4% and 82.2%, compared to 35.5% and 79.5%, 33.5% and 79.0% for RETAIN and MLP, respectively. To further validate the improvement in the predictive performance of our PredictPTB model over the RETAIN model, we conduct a statistical significance test by comparing the difference between the areas under the ROC curve and the areas under the Precision-recall curve for both models, across the four prediction points and two data windows, using the Mann-Whitney-Wilcoxon test, as shown in Table3.3. The results of

this test confirm that the improvement is significant (with p-value < 0.05) among all the prediction points and data windows, using all data modalities.

Table 3.3. Statistical significance test of the difference between the areas under ROC and precision-recall curves for the PredictPTB and RETAIN models.

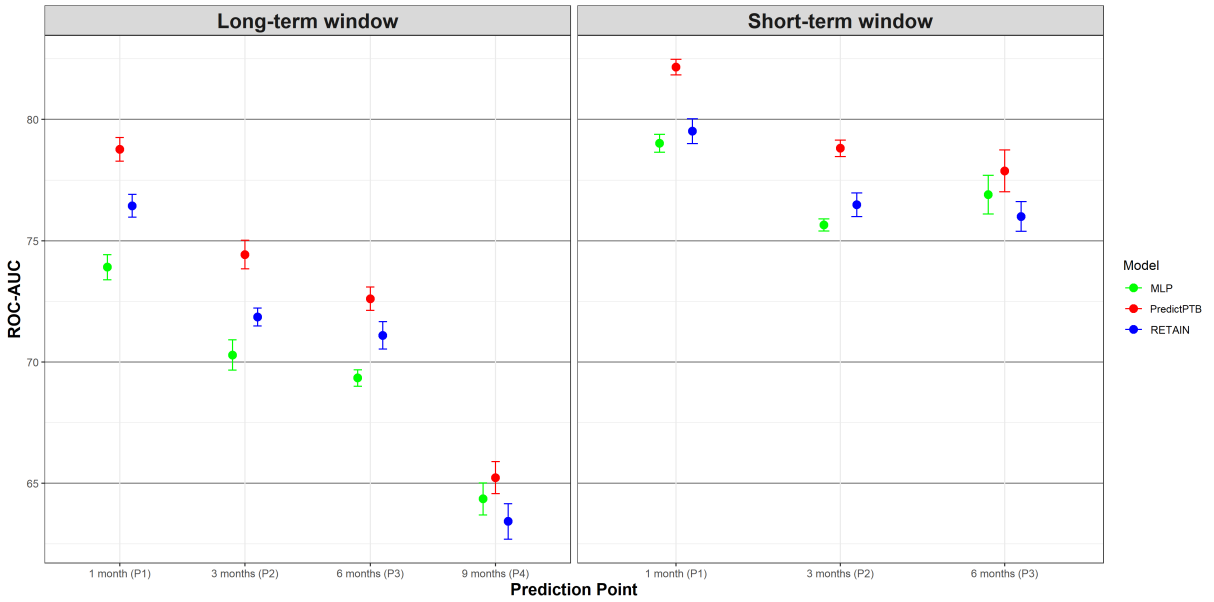
#	Prediction Point	Prediction Window	Curve	P-value
1	P1	long-term	ROC-AUC	0.00295
2	P2	long-term	ROC-AUC	0.00001
3	P3	long-term	ROC-AUC	0.00032
4	P4	long-term	ROC-AUC	0.00018
5	P1	short-term	ROC-AUC	0.00018
6	P2	short-term	ROC-AUC	0.00001
7	P3	short-term	ROC-AUC	0.00013
8	P1	long-term	PR-AUC	0.00010
9	P2	long-term	PR-AUC	0.00004
10	P3	long-term	PR-AUC	0.00001
11	P4	long-term	PR-AUC	0.00021
12	P1	short-term	PR-AUC	0.00151
13	P2	short-term	PR-AUC	0.00001
14	P3	short-term	PR-AUC	0.00003

In addition, results show that models improved in performance as the prediction point gets closer to the delivery date. For example, using the long-term window setup, our PredictPTB model was able to predict the risk of preterm birth at P1 (1 month before delivery) with a PR-AUC and ROC-AUC of 34.7% and 78.8%, compared to 26.1% and 74.4% at P2 (3 months), 21.0% and 72.6% at P3 (6 months), and 17.3% and 65.2% at P4 (9 months). The proposed PredictPTB uses a single attention mechanism to capture both visit and code-level contribution. Therefore the model size is smaller compared to RETAIN and less prune to over-fitting.

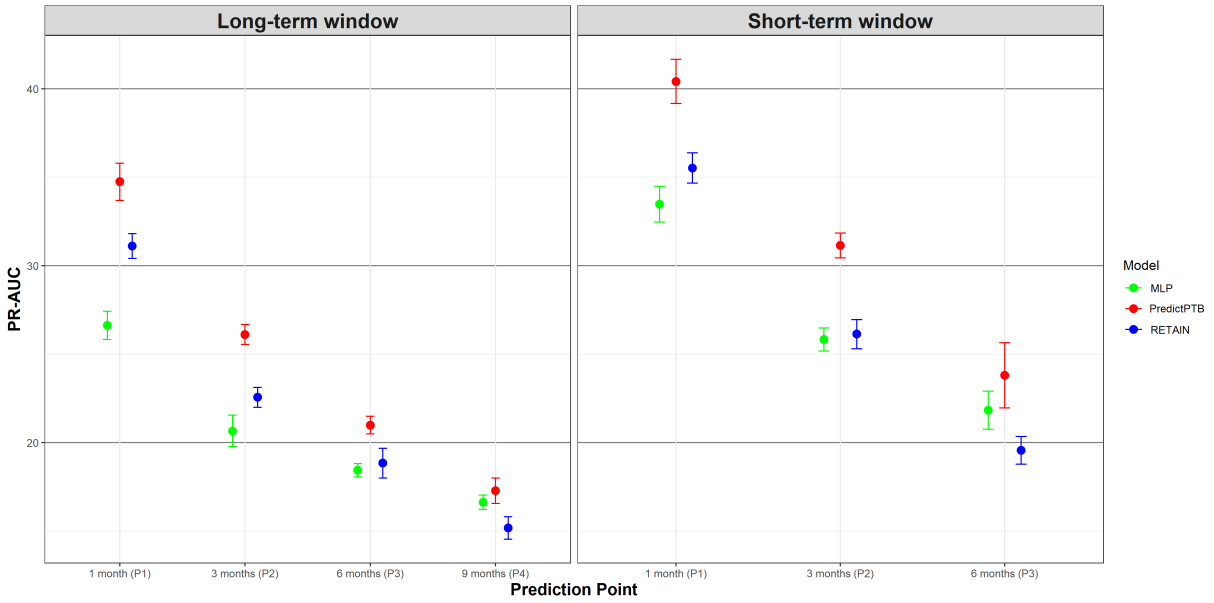
Moreover, our analysis confirmed that EHR features collected after the start of the pregnancy can better identify preterm births, compared to combining features collected before and after the start of the pregnancy. For example, using the short-term data

window, PredictPTB achieved a PR-AUC of 40.4%, compared to 34.7% for long-term window at P1. This finding confirms the general intuition that the short-term condition of a patient is usually the most determinant of health outcome. Chronic conditions might be captured in secondary diagnoses and hence likely to be available in the EHR data for model training.

Finally, the three prediction models showed a better sensitivity and specificity with short-term window setup, compared to long-term window setup, as shown in Figure 3.5. Sensitivity and specificity of the three models using the short-term window range from 68% to 73%. This confirms that using the EHR features collected after the start of the pregnancy can better designate a patient with a preterm delivery as positive, and thus fewer cases of preterm delivery are missed.

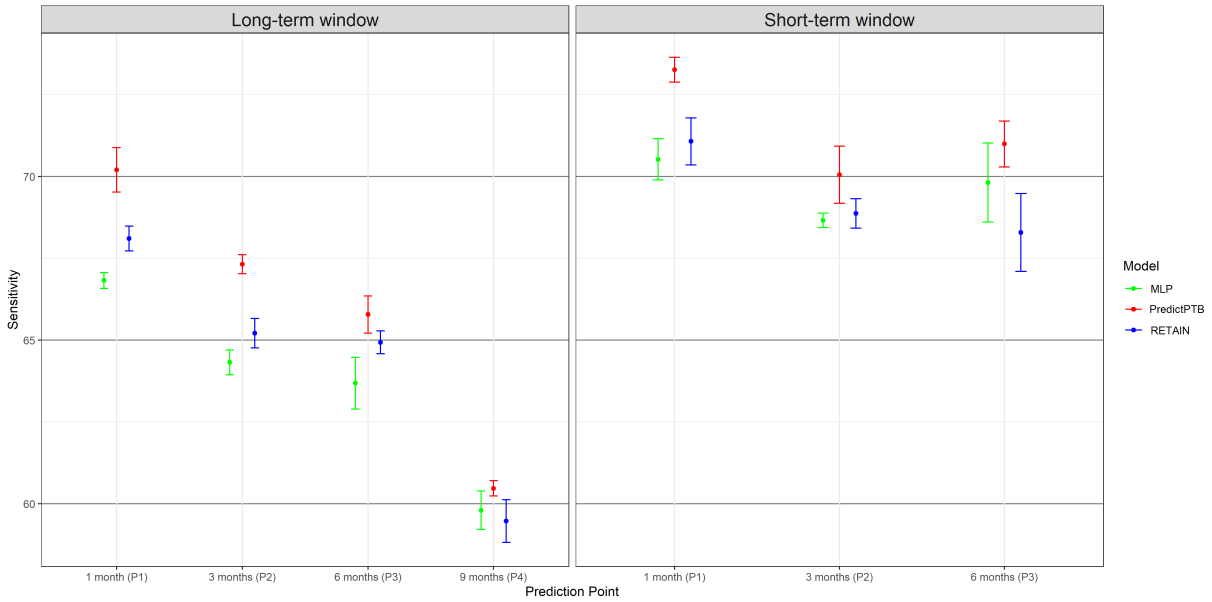


(a) ROC-AUC

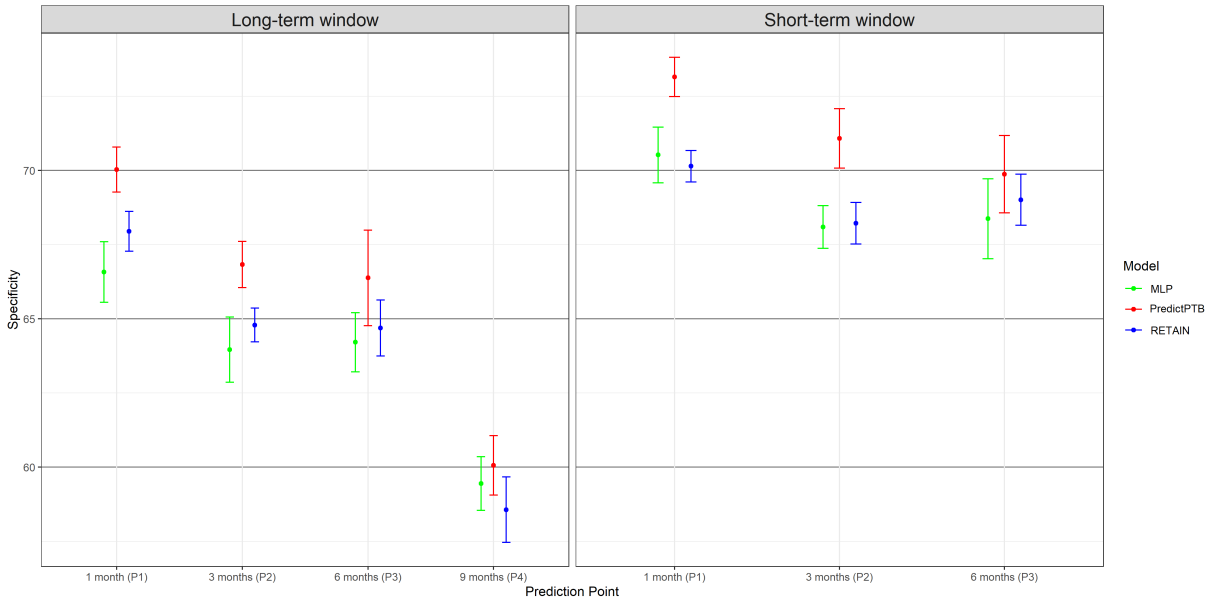


(b) PR-AUC

Figure 3.4. Predictive performance of the implemented models across the four prediction points using all data modalities for long-term and short-term data windows (ROC-AUC and PR-AUC).



(a) Sensitivity



(b) Specificity

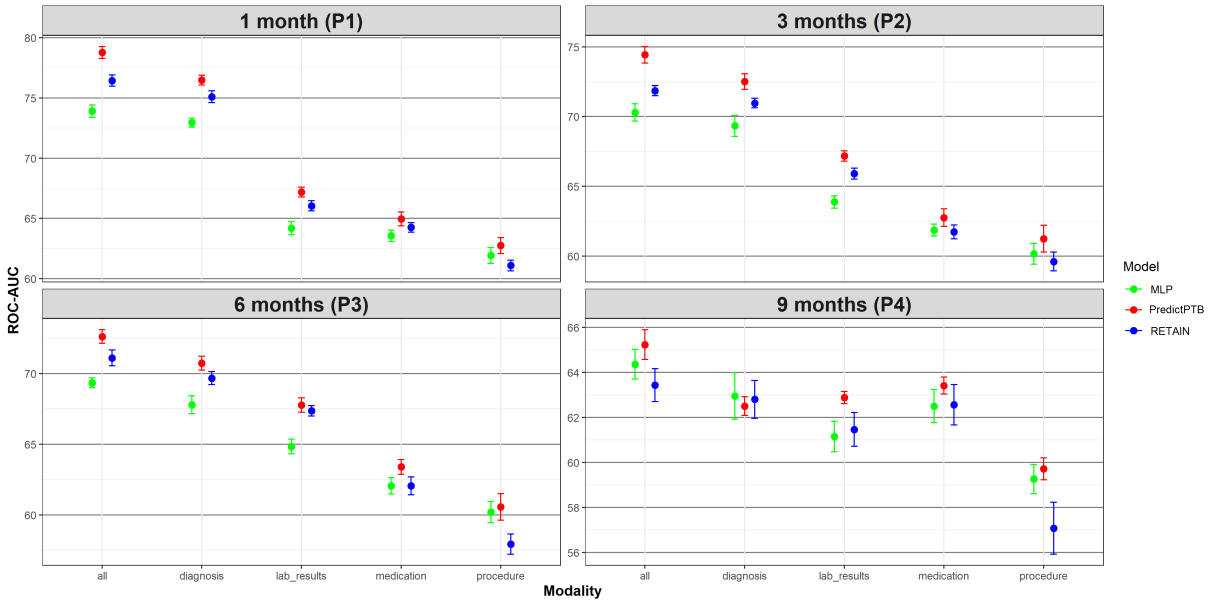
Figure 3.5. Predictive performance of the implemented models across the four prediction points using all data modalities for long-term and short-term data windows (sensitivity and specificity).

Prediction Performance Using Single Data Modality vs. Combined Data Modalities

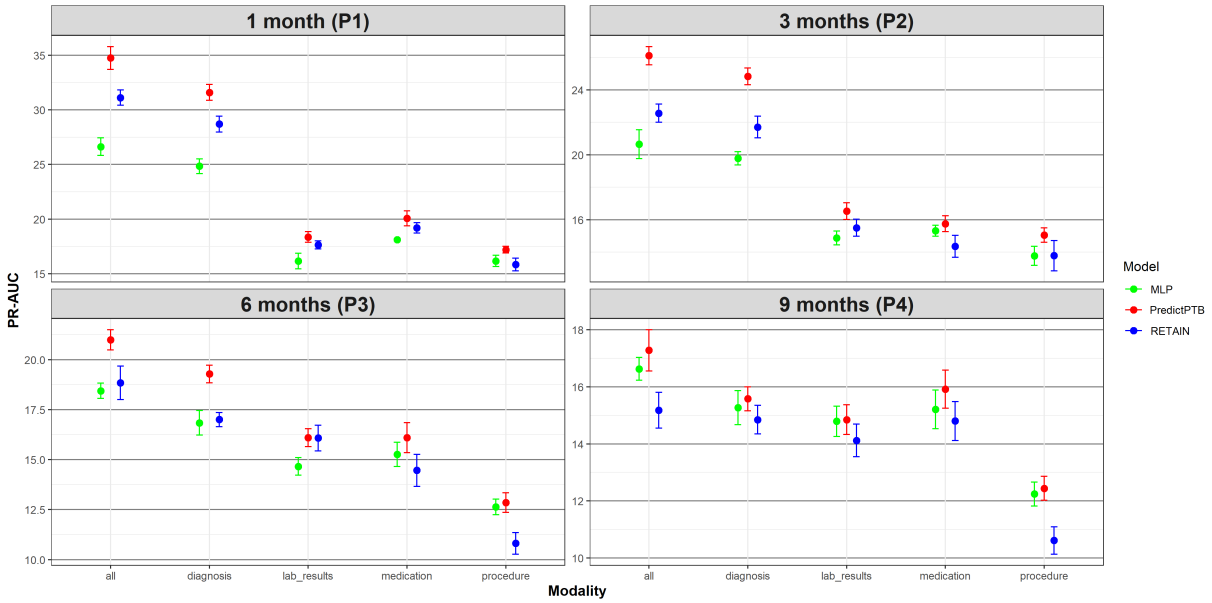
The objective of this experiment is to evaluate the prediction performance of models trained on each individual data modality compared to models trained on all modalities combined together. Results presented in Figures 3.6, 3.7, 3.8, and 3.9 show that combining all modalities together provides the best predictive performance, compared to using individual modalities. As for individual data modalities, diagnosis data achieved the highest performance across all prediction points and data windows. Diagnosis codes represents the most condensed information about patient status and history. While the main purpose of these codes are for billing purposes, it has been widely shown that they are highly predictive of patient health outcome [8]. At P1, diagnosis data was able to predict preterm delivery with PR-AUC and ROC-AUC of 36.7% and 80.1% for short-term window, and 31.6% and 76.5% for long-term window. As for data modalities, results demonstrated that combining all modalities together (diagnosis, medication, procedures, and lab results) improved the PR-AUC by 10% compared to using the diagnosis modality alone. The performance of other modalities varied among different data windows and prediction points. For example, procedures data was the least predictive modality for long-term window across all prediction time points. The error bars in Figures 3.6, 3.7, 3.8, and 3.9 represent the standard deviation of the calculated metrics at each prediction point. Overall, we note that the standard deviation for the combined modalities, diagnosis, and lab orders are smaller than those of medications and procedures across most of the prediction points, data windows, and metrics. Modalities with lower variability resulted in better prediction performance while modalities with higher variability (indicated by long error bars) resulted in lower prediction performance

due to less consistent data included in these modalities. For these results, we note that observational modalities (diagnosis and lab orders) are more predictive than intervention modalities (medication and procedure). Since interventions can be subjective to doctor opinion and understanding of the patient history and condition, it might explain the reason why it is less predictive than the observational data. Another explanation is that the data frequency for interventional data in EHR is lower than observational data. Hence model training is more reliable based on diagnosis and lab order modalities. On the other hand, for short-term window, procedures modality performed better than medications for the three prediction points P1, P2, and P3, and better than lab results for for P1 and P3. This highlights that procedures performed after the start of the pregnancy (short-term window), are better predictors for preterm birth, compared to medications and lab results ordered during pregnancy.

Finally, Figures 3.7 and 3.9 show that predicting preterm birth using diagnosis modality a lone was able to provide the best sensitivity and specificity results across all prediction points and data windows, compared to other data modalities. This suggests that a model trained using diagnosis data can better designate a patient with a preterm delivery as positive, and thus fewer cases of preterm delivery are missed, compared to using other data modalities. Moreover, Figure 3.9 shows that using medications prescribed during the pregnancy timeline for predicting preterm birth, and ignoring those prescribed before the start of the pregnancy, results in the lowest sensitivity and lowest specificity. In the case of preterm prediction, a low specificity is undesirable because the treatment of false positives can be dangerous for the patient and the baby, depriving her of correct treatment, and can also be very costly.

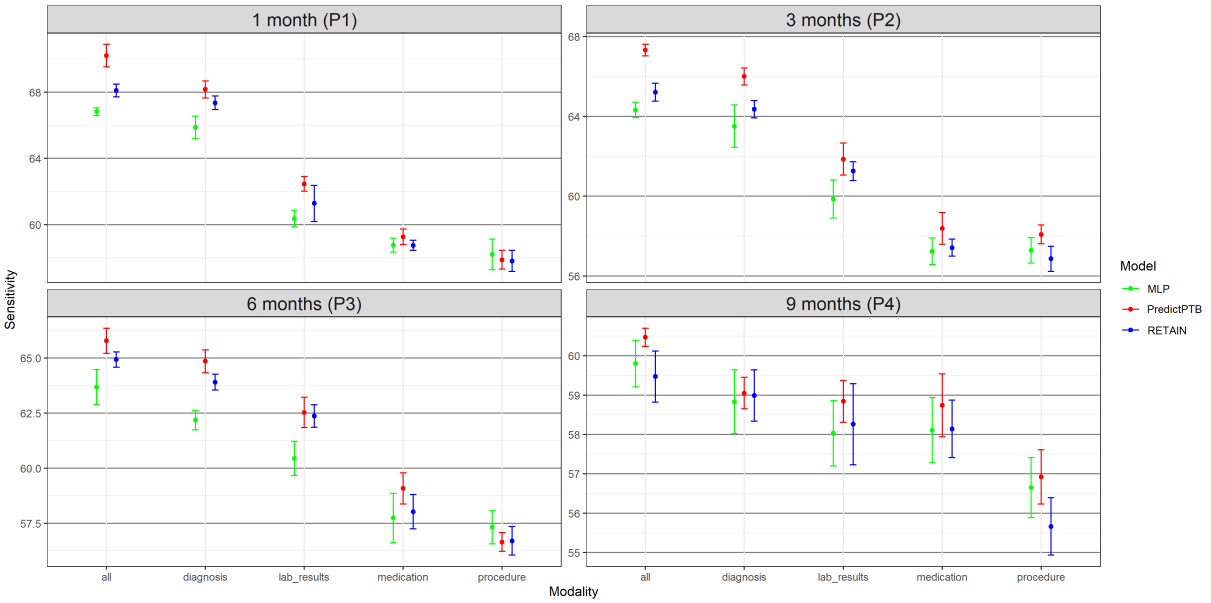


(a) ROC-AUC

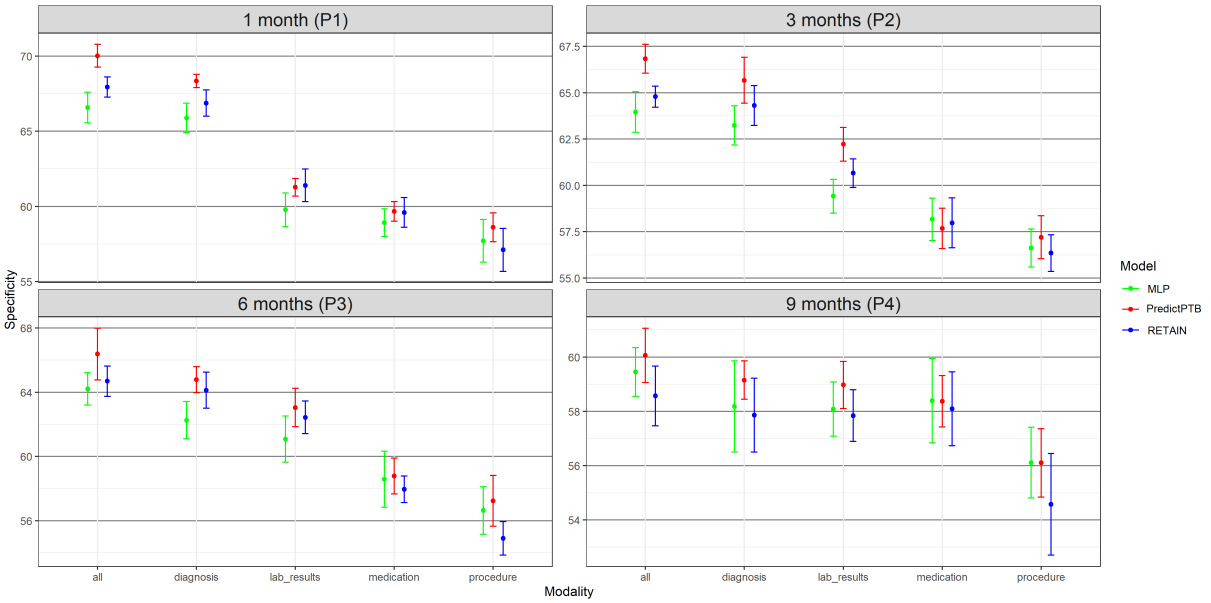


(b) PR-AUC

Figure 3.6. Predictive performance of models trained on single modality and integrated modalities of data, across the four prediction points using long-term data window (ROC-AUC and PR-AUC).

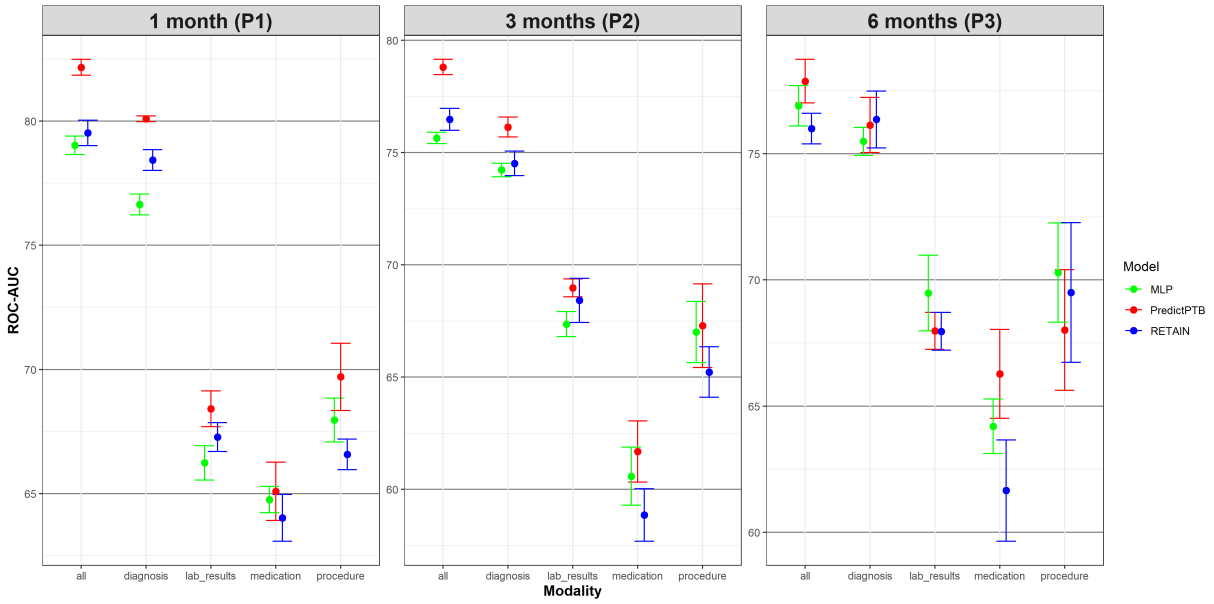


(a) Sensitivity

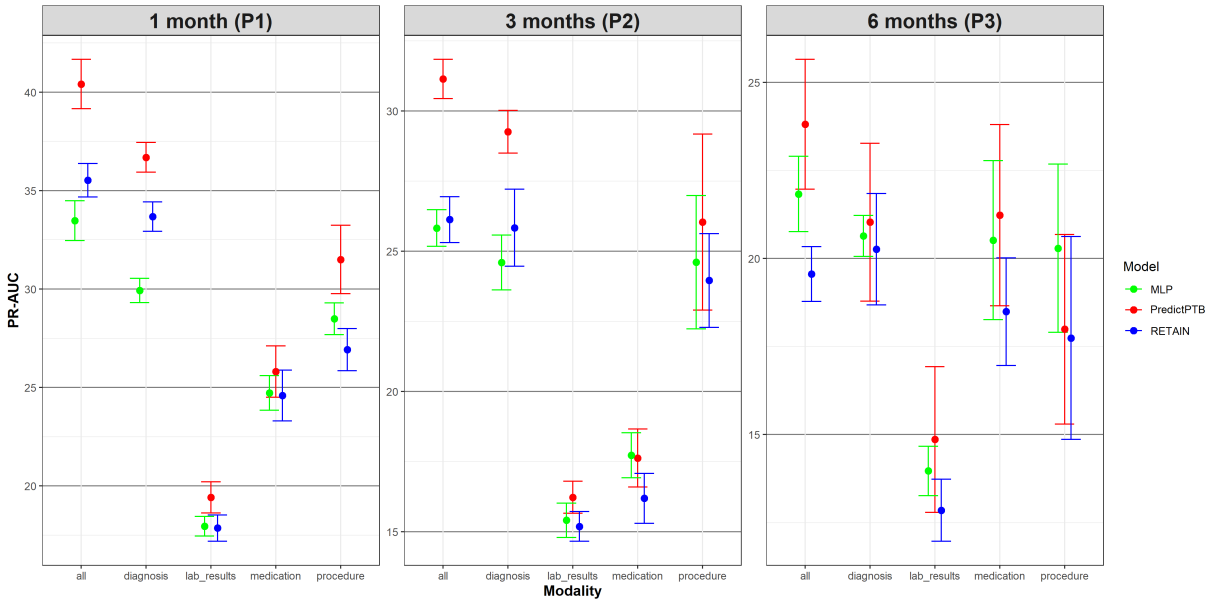


(b) Specificity

Figure 3.7. Predictive performance of models trained on single modality and combined modalities of data, across the four prediction points using long-term data window (sensitivity and specificity).

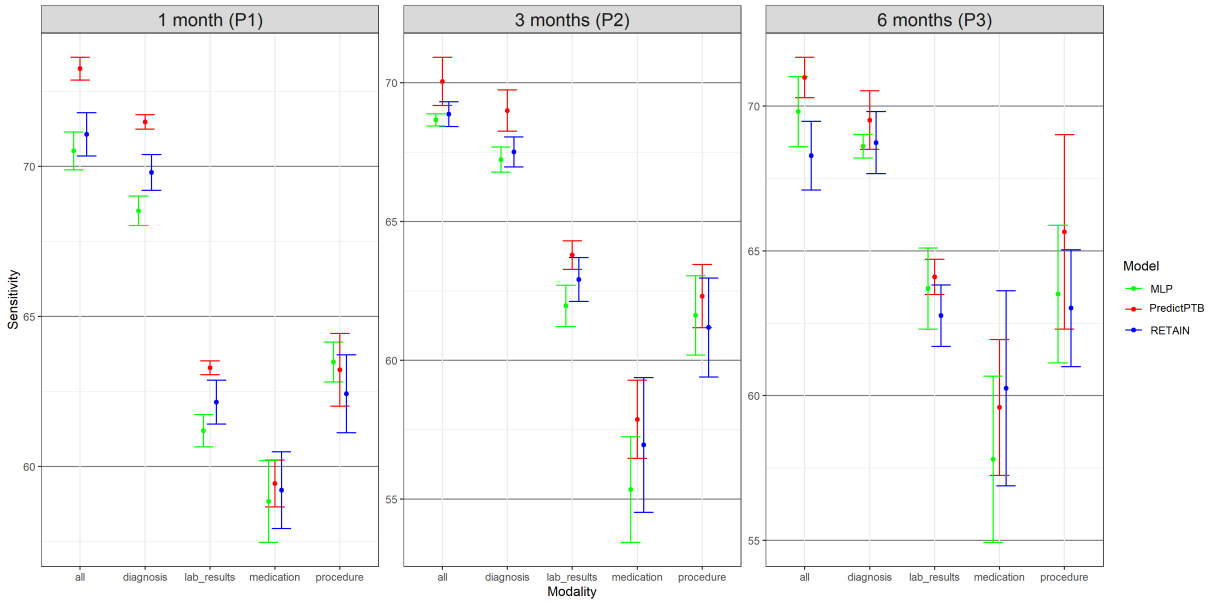


(a) ROC-AUC

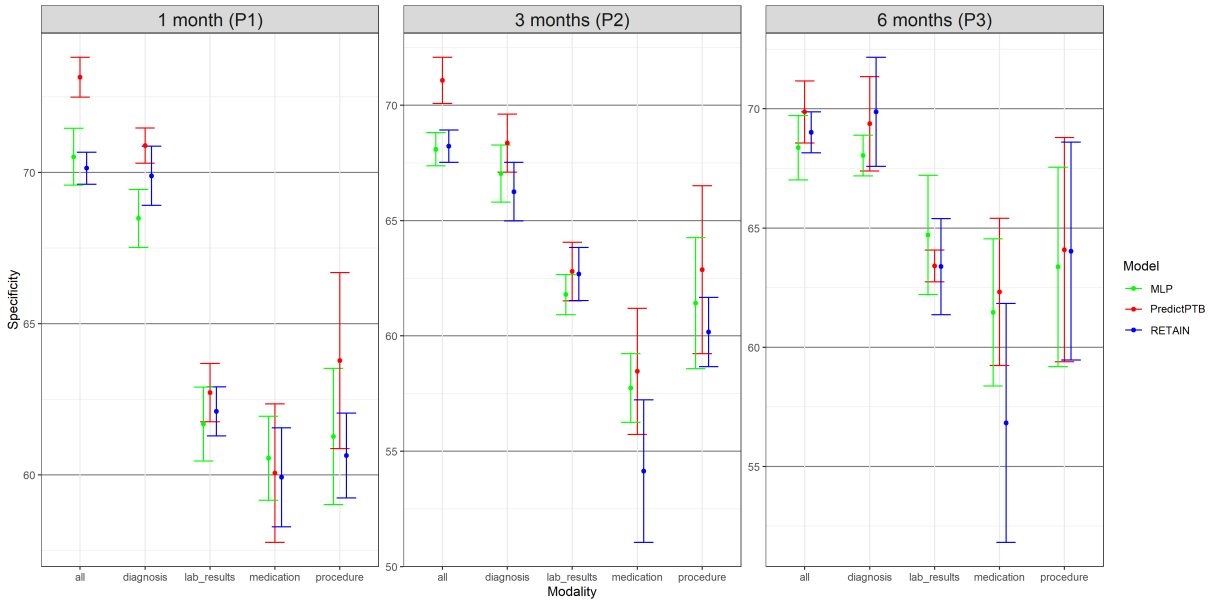


(b) PR-AUC

Figure 3.8. Predictive performance of models trained on single modality and integrated modalities of data, across the three prediction points using short-term data window (ROC-AUC and PR-AUC).



(a) Sensitivity



(b) Specificity

Figure 3.9. Predictive performance of models trained on single modality and integrated modalities of data, across the three prediction points using short-term data window (sensitivity and specificity).

3.4.6. Model Interpretation for Preterm Birth Prediction

We present a use case to demonstrate the ability of our PredictPTB model to explain individual prediction results, at visit-level and code-level, for preterm birth prediction. In addition, we compare the visit-level contributions generated by the PredictPTB and RETAIN models. For PredictPTB, code-level contributions are computed as described in Eq 3.10, and visit-level contributions are computed by aggregating the contributions of individual codes included within a visit.

Code-level and Visit-level Interpretations

In Figure 3.10, we show an example for a patient from the test set, predicted as preterm delivery using the PredictPTB model. The blue square (in Oct 2017) indicates the delivery date. The attention weights show strong contribution of a previous visit, about two years prior to delivery. Looking closer at the medical codes included in this visit (Figure 3.11), we can see that a previous single live birth delivery at 30-34 weeks of gestation (preterm) occurred on that day, due to severe preeclampsia. These findings are in line with published literature reporting that a history of prior preterm birth and preeclampsia in a previous pregnancy are major risk factors for preterm delivery in subsequent pregnancies [149], [150]. In addition, the codes included in this visit indicate a long-term and current use of aspirin during the previous pregnancy, probably to prevent preterm preeclampsia. This observation suggests that there might be potential long-term effect of aspirin given to prevent preterm preeclampsia on subsequent pregnancies, but further research is required to study this association. Moreover, the codes on this visit report that this patient had a previous cesarean delivery, which might also increase the

risk of preterm birth in later pregnancies [151], [152]. This patient is a great example demonstrating the ability of the attention mechanism to go beyond recent events and model long-range dependencies among medical events to the predicted outcome. In addition, Figure 3.12 shows another example for a patient where PredictPTB captures common risk factors and assigns a high importance score to the visit in which these codes are documented. The highlighted visit has two important risk factors: infection of urinary tract in pregnancy and pre-existing diabetes mellitus in pregnancy. Finally, Figure 3.13 presents an example for a patient where PredictPTB was able to learn rare complications as risk factors for preterm birth. This patient was diagnosed with twin-to-twin transfusion syndrome (TTTS), a rare disorder which affects $10 - 15\%$ of monochorionic, diamniotic twin pregnancies [153]. This observation is in line with literature reporting that pregnancies with TTTS complication are at increased risk for PTB [154].

Comparison of Visit-level Attributions of PredictPTB and RETAIN Models

Here, we show an example for a use-case, where PredictPTB was able to produce more clinically-relevant explanations for preterm prediction than RETAIN, using data available up to one month before the delivery event. In Figure 3.14, we can observe that PredictPTB highlights a few visits, a month before the delivery, with high contributions in February 2016, while RETAIN highlights some older visits with high contribution between August 2013 and December 2014. The visits which were highlighted by the PredictPTB model includes the following codes: F41.9: Anxiety disorder, J02.9: Acute pharyngitis, J06.9: Acute upper respiratory infection, and 18481-2: Culture Throat and Group A Beta Strep AG Rapid Screen Qualitative, while visits highlighted by the

RETAIN model includes the these codes: Z32.01: Encounter for pregnancy test result positive, Z30.9: Encounter for contraceptive management, and N76.0: Acute vaginitis. For this patient, PredictPTB visit attributions seem to be more clinically-relevant than RETAIN attributions, since having an acute upper respiratory infection such as acute pharyngitis was found to be positively correlated with preterm birth [155], [156]. In addition, anxiety has been reported in several studies as a risk factor for preterm birth [157], [158]. The visits highlighted by the RETAIN model, are about three years prior to the delivery, which makes them less relevant, especially not reporting potential risk factors for preterm delivery, except for acute vaginitis, which might have only affected the previous pregnancy and not the current one [159]. There is currently no literature reporting that acute vaginitis in a pregnancy increases the risk of preterm birth in subsequent pregnancies.

This use-case demonstrates the ability of the PredictPTB model to utilize the computed contributions of individual medical codes to explain code-level and visit-level contributions to the model's prediction for a particular patient. In addition, PredictPTB may be able to provide more precise interpretations than RETAIN for preterm birth predictions. The use of a single attention layer for both code-level and visit-level attributions seems to provide a more consistent interpretations compared to two separate attention layers as in RETAIN.

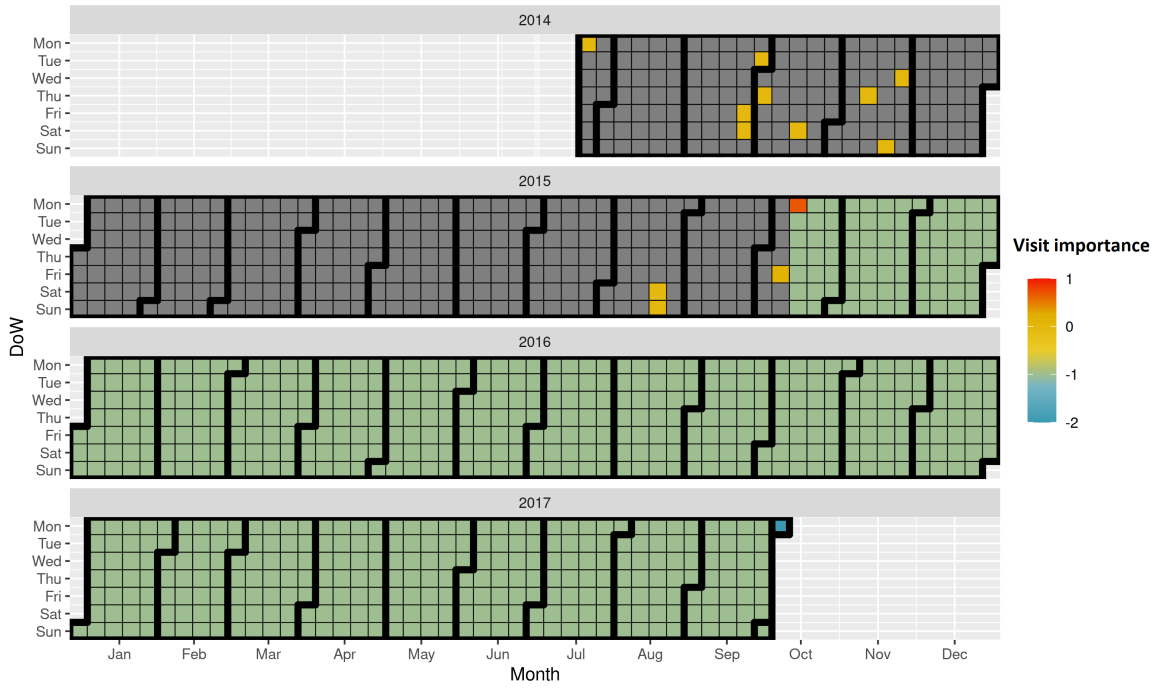


Figure 3.10. Temporal visualization of visit-level contributions over a patient's EHR timeline, using a PredictPTB model trained to predict preterm birth.

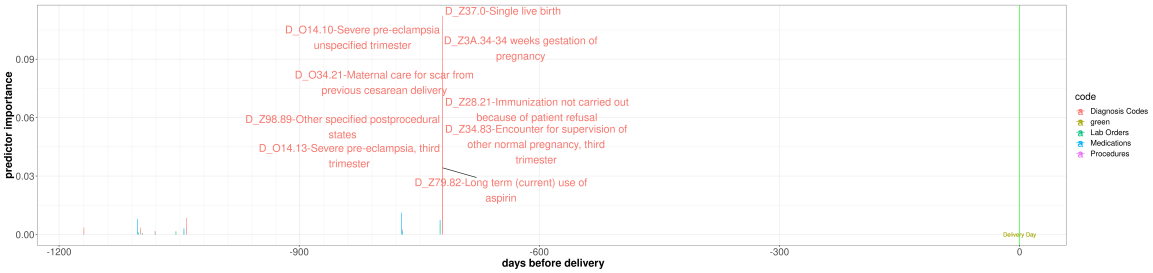


Figure 3.11. Interpretation of prediction results over a patient's EHR timeline. The code-level attribution in each visit is shown along the x-axis (i.e. time) with the y-axis representing the magnitude of individual codes contributions to preterm birth in each visit.

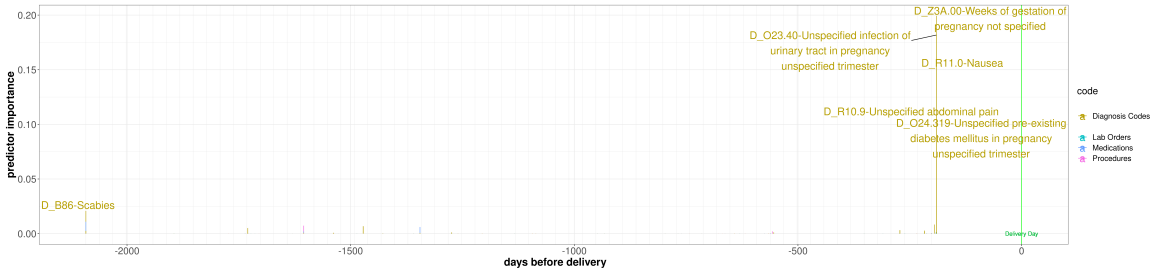


Figure 3.12. Example for a patient where PredictPTB captures common risk factors and assigns a high importance score to the visit in which these codes are documented. The highlighted visit has two important risk factors: infection of urinary tract in pregnancy and pre-existing diabetes mellitus in pregnancy.

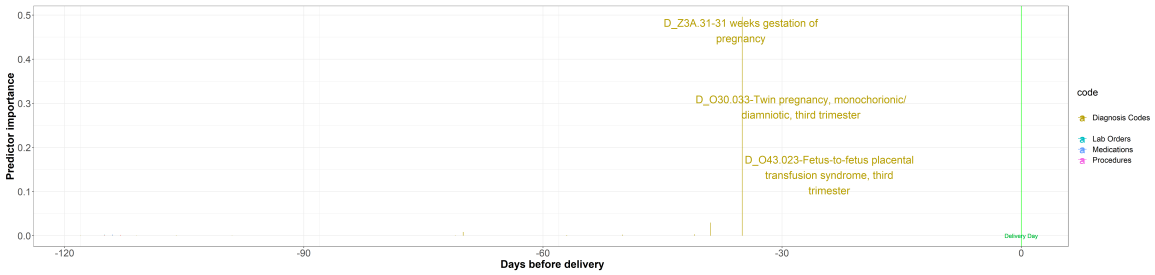


Figure 3.13. Example for a patient where PredictPTB was able to learn rare complications as risk factors for preterm birth. This patient was diagnosed with twin-to-twin transfusion syndrome (TTTS), a rare disorder which affects $10 - 15\%$ of monochorionic, diamniotic twin pregnancies

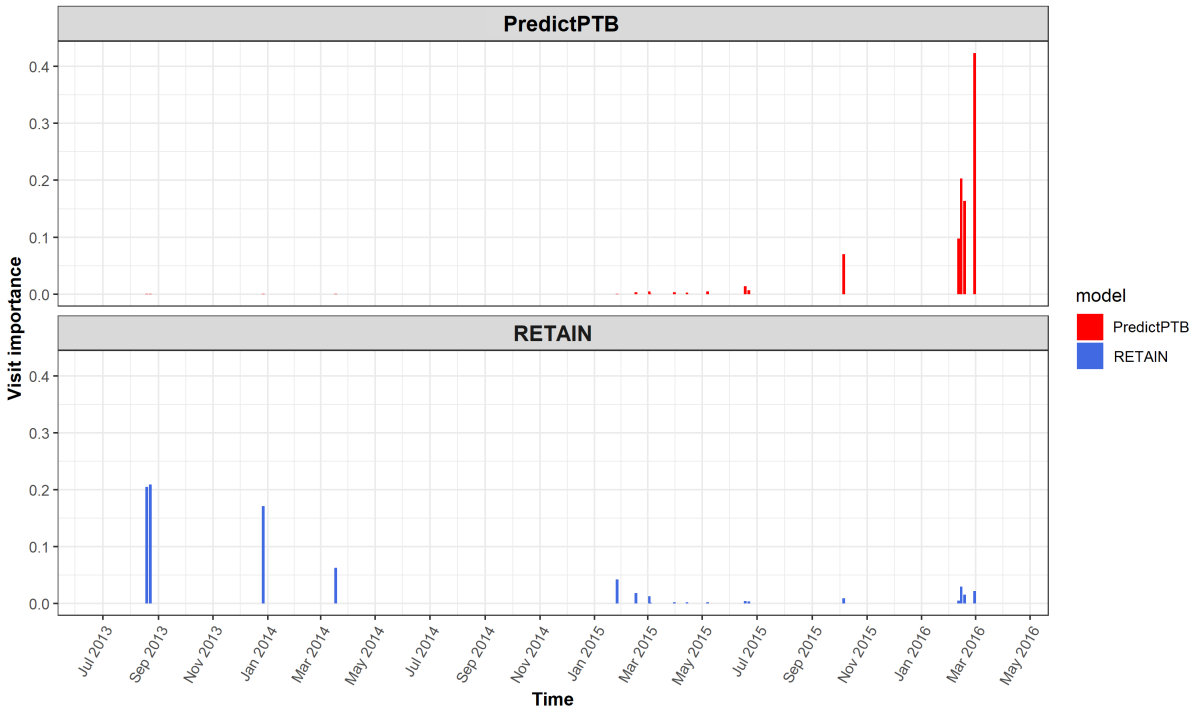


Figure 3.14. Comparison of visit-level attributions between PredictPTB and RETAIN models.

3.5. Discussion

We introduced a code-level attention-based predictive model and demonstrated its usability on the task of predicting the risk of preterm birth, up to 9 months earlier than its occurrence, using data routinely collected in EHR. The core component of our model is a code-level attention-based RNN, which can embed relevant contextual information into each medical code to generate code-level attention weights.

Results showed that our PredictPTB model performed better than the RETAIN model, across all prediction points, data windows, and data modalities. The PredictPTB model achieved the best predictive performance at 30 days before delivery, with an ROC-AUC of 82.2% and PR-AUC of 40.4%, when using combined data modalities and pregnancy history setup. As for data modalities, results demonstrated that combining

all modalities together (diagnosis, medication, procedures, and lab results) improved the PR-AUC by 10% compared to using the diagnosis modality alone and improved the ROC-AUC for about 4%. This suggests that combination of all data modalities have the potential to quantify PTB risk better than using the diagnosis data alone, due to enrichment of individual patient representation with multiple data sources. Moreover, results showed that the use of models trained on data collected after the start of the pregnancy improved the performance by up to 5% on ROC-AUC, compared to models trained on the full patient EHR history, which includes both data before and after the start of the pregnancy. This supports previous findings in [160]–[162], which indicates that the most important risk factors are associated with events happening during the pregnancy timeline.

The PredictPTB model has a number of architectural advantages over previous methods for modeling EHR data for prediction of clinical outcomes. Firstly, the attention-based approach we implemented is more interpretable than other black-box deep learning methods, which often lack the capability of identifying features driving predictions. This property enables clinician to better understand risk factors associated with the predicted clinical outcomes. Second, our flexible architecture enables capturing additional modalities of EHR data (e.g. surgeries, clinical notes, etc.), by simply adding a fifth or sixth (or more) list of concepts to our embedding layer. Third, PredictPTB is based on a relatively simple architecture compared to previous methods such as RETAIN, and GRAM [108]. PredictPTB reduces the complexity of the RETAIN architecture while improving the accuracy of the predictions. In addition, this simplified architecture reduces the risk of over-fitting and possible gradient flow.

Compared to published work on preterm birth prediction, our approach has a number

of advantages. First, to the best of our knowledge, this work is one of the first to consider such a large number of patients (222,436 deliveries), collected from more than 300 healthcare centers in the U.S. This large dataset enabled our model to learn for diverse patient conditions and be able to capture common as well as relatively uncommon risk factors (examples are available in Appendix 6.2). Second, given the unbalanced classification problem where preterm deliveries are much less than full-term deliveries, our model has a high predictive performance with an ROC-AUC of 82.2% and PR-AUC of 40.4% using data available up to one month before delivery. Previous work on predicting preterm reported poor to moderate predictive performance with an ROC-AUC ranging from 59% to 74% [139], [163], [164]. Third, compared to previous work, our model is capable of leveraging the sequential and temporal trajectory of events recorded in EHR, which included large amount of information embedded in each patient's record.

A recent work [147], which predicts preterm birth using gradient boosted decision trees on EHR data, have achieved an a ROC-AUC of 0.75 and PR-AUC of 0.40 at 28 weeks of gestation (which is approximately two months before delivery). Compared to PredictPTB, this model provides predictions starting from two months before delivery and up to 10 days before delivery, and no information was provided about the model's performance at earlier stages of the pregnancy timeline. In addition, this work is limited to diagnosis data and does not consider other data modalities. On the other hand, PredictPTB combines four data modalities (diagnosis, medications, lab orders, and procedures). This combination enables PredictPTB to learn a better representation that can capture a patient's EHR in as much detail as possible.

Furthermore, our interpretability visualization highlighted several known risk factors for preterm birth, which establishes further confidence in our approach. Finally, our

interpretability analysis for some case studies suggested additional potential risk factors for further investigation by domain experts.

Our proposed model supports the idea of personalized clinical decision support, by deriving relative importance of an individual medical code based on the context of the entire EHR history of a patient. A possible clinical application scenario would use our model to scan the medical history of an expecting mother, compute the risk score for preterm birth, and provide a visualization for healthcare providers to help them identify patients at high-risk of preterm delivery and arrange early follow-ups that could prevent complications and additional burden for the healthcare system and patients.

3.6. Summary

The primary contribution of this chapter is the development and evaluation of a code-level attention-based recurrent neural network model for preterm birth prediction. We demonstrate that temporal deep learning models can predict preterm delivery up to nine months earlier than its occurrence, using routinely collected data in electronic health records. Future work may utilize our model to provide patient-specific predictions and interpretations for more pregnancy complications (e.g. hypertension, gestational diabetes, preeclampsia, infections, and iron-deficiency anemia) and pregnancy outcomes (e.g. mode of delivery, stillbirth, miscarriage, and neonatal death).

CHAPTER 4: CONTEXTUAL DECOMPOSITION OF BIDIRECTIONAL LONG SHORT-TERM MEMORY MODEL

4.1. Overview

Predictive modeling with longitudinal electronic health record (EHR) data offers great promise for accelerating personalized medicine and better informs clinical decision-making. Recently, deep learning models have achieved state-of-the-art performance for many healthcare prediction tasks. However, deep models lack interpretability, which is integral to successful decision-making and can lead to better patient care. In this chapter, we build upon the contextual decomposition (CD) method, an algorithm for producing importance scores from long short-term memory networks (LSTMs). We extend the method to bidirectional LSTMs (BiLSTMs) and use it in the context of predicting future clinical outcomes using patients' EHR historical visits. We use a real EHR dataset comprising 11071 patients, to evaluate and compare CD interpretations from LSTM and BiLSTM models. First, we train LSTM and BiLSTM models for the task of predicting which pre-school children with respiratory system-related complications will have asthma at school-age. After that, we conduct quantitative and qualitative analysis to evaluate the CD interpretations produced by the contextual decomposition of the trained models. In addition, we develop an interactive visualization to demonstrate the utility of CD scores in explaining predicted outcomes. Our experimental evaluation demonstrates that whenever a clear visit-level pattern exists, the models learn that pattern and the contextual decomposition can appropriately attribute the prediction to the correct pattern. In addition, the results confirm that the CD scores agree to a large extent with the importance scores generated using logistic regression coefficients. Our

main insight was that rather than interpreting the attribution of individual visits to the predicted outcome, we could instead attribute a model's prediction to a group of visits. We presented a quantitative and qualitative evidence that CD interpretations can explain patient-specific predictions using CD attributions of individual visits or a group of visits.

4.2. Background

The exponential surge in the amount of digital data captured in electronic health record (EHR) offers promising opportunities for predicting the risk of potential diseases and better informs decision-making. Recently, deep learning models have achieved impressive results, compared to traditional machine learning techniques, by effectively learning non-linear interactions between features for several clinical tasks [10]–[12], [27], [29]. Among a variety of deep learning methods, recurrent neural networks (RNNs) could incorporate the entire EHR to produce predictions for a wide range of clinical tasks [13]–[19]. Consequently, there is a growing realization that, in addition to predictions, deep learning models are capable of producing knowledge about domain relationships contained in data; often referred to as interpretations [165], [166].

However, the high-dimensionality and sparsity of medical features captured in the EHR makes it more complex for clinicians to interpret the relative impact of features and patterns which are potentially important in decisions. A patient's EHR usually consists of a sequence of visits a patient has made, and each visit captures the list of diagnosis codes documented by the clinician. Therefore, it is reasonable and important to have interpretable models which can focus on patient visits that have higher impact on the predicted outcome, ignore those visits with little effect on the outcome, and identify and validate the relevant subset of visits driving the predictions.

Interpreting deep models trained on EHR data for healthcare applications is a growing field spanning a range of techniques, which can be broadly categorized into three classes: attention mechanism, knowledge injection via attention, and knowledge distillation [12]. Attention-mechanism-based learning was used in [23], [40], [108], [167]–[170] for explaining what part of historical information weighs more in predicting future clinical events. Knowledge injection via attention often integrates biomedical ontologies, as a major source of biomedical knowledge, into attention models to enhance interpretability, as demonstrated in [108]. Knowledge distillation first trains a complex, slow, but accurate model and then compresses the learned knowledge into a much simpler, faster, and still accurate model, as shown in [42], [171]. However, the majority of previous work has focused on assigning importance scores to individual features. As a result, these techniques only provide limited local interpretations and do not model fine-grained interactions of groups of input features. In addition, most of these techniques require modifications on standard deep learning architectures to make it more interpretable. By contrast, there are relatively few methods that can extract interactions between features that a deep neural network (DNN) learns. In the case of LSTMs, a recent work by Murdoch et al.[172] introduced contextual decomposition (CD), an algorithm for producing phrase-level importance scores from LSTMs without any modifications to the underlying model, and demonstrated it on the task of sentiment analysis.

In this work, we hypothesized that the CD interpretability method translates well to healthcare. Therefore, we build upon the CD technique and extend it to BiLSTMs in the context of predicting future clinical outcomes using EHR data. Particularly, we aimed to produce visit-level CD scores explaining why a BiLSTM model produced a certain prediction using patients' EHR historical visits. Our main insight was that rather

than interpreting the attribution of individual visits to the predicted outcome, we could instead attribute BiLSTM's prediction to a subset of visits. Our main contributions are as follows:

- We introduce a CD-based approach to determine the relative contributions of single visits and a group of visits in explaining the predicted outcome, and subsequently identify the most predictive subset of visits.
- We develop an interactive visualization and demonstrate, using a concrete case study, how CD scores offer an intuitive visit-level interpretation.
- We evaluate and compare CD interpretations from LSTM and BiLSTM models for the task of predicting which pre-school children with respiratory system-related complications will have asthma at school age.
- On a real EHR dataset comprising 11,071 patients having a total of 3,318 different diagnosis codes, we present quantitative and qualitative evidence that CD interpretations can explain patient-specific predictions using CD attributions of individual visits or a group of visits.

4.3. Methods

4.3.1. EHR Data Description

The EHR data consists of patients' longitudinal time-ordered visits. Let P denote the set of all the patients $\{p_1, p_2, \ldots, p_{|P|}\}$, where $|P|$ is the number of unique patients in the EHR. For each patient $p \in P$, there are T_p time-ordered visits $V_1^{(p)}, V_2^{(p)}, \ldots, V_{T_p}^{(p)}$. We denote $D = \{d_1, d_2, \ldots, d_{|D|}\}$ as the set of all the diagnosis codes, and $|D|$ represents the

number of unique diagnosis codes. Each visit $V_t^{(p)}$, where the subscript t indexes the time step, includes a subset of diagnosis codes, which is denoted by a vector $x_t^{(p)} \in \{0,1\}^{|D|}$. The i-th element in $x_t^{(p)}$ is 1 if d_i existed in visit $V_t^{(p)}$ and 0 otherwise. For notational convenience, we will henceforth drop the superscript (p) indexing patients.

4.3.2. Long Short Term Memory Networks

Long short term memory networks (LSTMs) are a special class of recurrent neural networks (RNNs), capable of selectively remembering patterns for long duration of time. They were introduced by Hochreiter and Schmidhuber [173], and were refined and widely used by many people in following work. For predictive modeling using EHR data, LSTMs effectively capture longitudinal observations, encapsulated in a time-stamped sequence of encounters (visits), with varying length and long range dependencies. Given an EHR record of a patient p, denoted by $X = \{x_t\}_{t=1}^{T}$, where T is an integer representing the total number of visits for each patient. The LSTM layer takes X as input and generates an estimate output Y, by iterating through the following equations at each time step t:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{4.1}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{4.2}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{4.3}$$

$$g_t = tanh(W_g x_t + U_g h_{t-1} + b_g) \tag{4.4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{4.5}$$

$$h_t = o_t \odot tanh(c_t) \tag{4.6}$$

where $i, f,$ and o are respectively the input gate, forget gate, and output gate, c_t is the cell vector, and g_t is the candidate for cell state at timestamp t, h_t is the state vector, W_i, W_f, W_o, W_g represent input-to-hidden weights, U_i, U_f, U_o, U_g represent hidden-to-hidden weights, and b_i, b_f, b_o, b_g are the bias vectors. All the gates have sigmoid activations and cells have $\tanh$ activations.

4.3.3. Bidirectional Long Short Term Memory Networks

Bidirectional LSTMs [174] make use of both the past and the future contextual information for every time step in the input sequence X in order to calculate the output. The structure of an unfolded BiLSTM consists of a forward LSTM layer and a backward LSTM layer. The forward layer outputs a hidden state $\overrightarrow{h}$, which is iteratively calculated using inputs in the forward or positive direction from time $t = 1$ to time T. The backward layer, on the other hand, outputs a hidden state $\overleftarrow{h}$, calculated from time $t = T$ to 1, in the backward or negative direction. Both the forward and backward layer outputs are calculated using the standard LSTM updating equations 4.1- 4.6, and the final h_t is calculated as:

$$\overrightarrow{h} = \overrightarrow{LSTM}(x_t) \tag{4.7}$$

$$\overleftarrow{h} = \overleftarrow{LSTM}(x_t) \tag{4.8}$$

$$h_t = [\overrightarrow{h}, \overleftarrow{h}] = BiLSTM(x_t) \tag{4.9}$$

The final layer is a classification layer, which is the same for an LSTM- or BiLSTM-based architecture. The final state h_t is treated as a vector of learned features and used as input to an activation function to return a probability distribution p over C classes. The probability p_j of predicting class j is defined as follows:

$$p_j = \frac{exp(W_j \cdot h_t + b_j)}{\sum_{i=1}^{C} exp(W_i \cdot h_t + b_i)} \tag{4.10}$$

where W represents the hidden-to-output weights matrix and W_i is the i-th column, b is the bias vector of the output layer and b_i is the i-th element.

4.3.4. Contextual Decomposition of BiLSTMs

Murdoch et al.[172] suggested that for LSTM, we can decompose every output value of every neural network component into relevant contributions β and an irrelevant contributions γ as:

$$Y = \beta + \gamma \tag{4.11}$$

We extend the work of Murdoch et al.[172] to BiLSTMs, in the context of patient visit-level decomposition for analyzing patient-specific predictions made by standard BiLSTMs. Given an EHR record of a patient, $X = \{x_t\}_{t=1}^{T}$, we decompose the output of the network for a particular class into two types of contributions: (1) contributions

made solely by an individual visit or group of visits, and (2) contributions resulting from all other visits of the same patient.

Hence, we can decompose h_t in (4.6) as the sum of two contributions β and γ. In practice, we only consider the pre-activation and decompose it for BiLSTM as:

$$W_j \cdot (\overrightarrow{h}, \overleftarrow{h}) + b_j = W_j \cdot [\overrightarrow{\beta}, \overleftarrow{\beta}] + W_j \cdot [\overrightarrow{\gamma}, \overleftarrow{\gamma}] + b_j \tag{4.12}$$

Finally, the contribution of a subset of visits with indexes S to the final score of class j is equal to $W_j \cdot \beta$ for LSTM and $W_j \cdot [\overrightarrow{\beta}, \overleftarrow{\beta}]$ for BiLSTM. We refer to these two scores as the CD attributions for LSTM and BiLSTM throughout the chapter.

4.3.5. Finding Most Predictive Subset of Visits

We introduce a CD-based approach to find the most predictive subset of visits, with respect to a predicted outcome. More specifically, the goal is to find subset of visits $X_S \in X$, where X_S consists of the visits with the highest relevant contribution $W_j \cdot [\overrightarrow{\beta}, \overleftarrow{\beta}]$ presented to the user.

Algorithm 1 describes the exact steps to find the most predictive subset of visits represented by X_S with the highest relative CD attributions. We consider V is the list of all patient visits, W is the list of all window sizes to analyse, and each $w \in W$ is an integer setting the size of the window, s is an integer setting the size of the step between windows, m is the model to be decomposed (LSTM/BiLSTM). In our context, a sliding window is a time window of fixed width w that slides across the list of patient visits V with step size s and returns the list of $CandidateGroups$ (subsets of visits) with the specified w. For each of these $CandidateGroups$, the algorithm takes the subset of visits and apply contextual decomposition on the specified model m to get the relative

contribution scores of this subset of visits against the complete list of patient visits. This procedure is applied iteratively for each window size w. Finally, the $group$ with the highest CD score is assigned to X_S.

This approach, while simple, exhaustively evaluates all possible combinations of subsets of consecutive visits, and then finds the best subset. Obviously, the exhaustive search's computational cost is high. However, since the total number of visits doesn't exceed tens usually, going through all possible combinations of consecutive visits is still computationally feasible.

Algorithm 1 Finding Most Predictive Subset of Visits

Let $V = \{v_1 \dots v_x\}$
Let $W = \{w_1 \dots w_y\}$
Let $s = 1$
Let $model = m$
Let $groupScores = []$
function FINDVISITSUBSET(V, W, m, s)
    for w in W **do**
        $CandidateGroups = slidingWindow(V, w, s)$
        for $group$ in $CandidateGroups$ **do**
            $groupScores[group] = contextualDecomposition(group, V, model)$
        end for
    end for
    $X_s = \text{argmax}(groupScores)$
end function

4.3.6. Dataset and Cohort Construction

The data was extracted from the Cerner Real World Data © EHR database, which consists of patient-level data collected from 561 health care facilities in the United States with 240 million encounters for 43 million unique patients collected between the years 2000-2013 [175]. The data is de-identified and is HIPAA (Health Insurance Portability and Accountability Act)-compliant to protect both patient and organization identity.

For the purpose of our analysis, we identified children with respiratory system-related symptoms by following the International Classification of Diseases (ICD-9) standards. We extracted 323,555 children who had a diagnosis code of 786.x (Symptoms involving respiratory system and other chest symptoms), except 786.3: hemoptysis. This includes the following diagnosis codes: 786.0 dyspnea and respiratory abnormalities (respiratory abnormality, hyperventilation, orthopnea, apnea, cheyne-stokes respiration, shortness of breath, tachypnea, wheezing, other respiratory abnormalities), 786.1 stridor, 786.2 cough, 786.4 abnormal sputum, 786.5 chest pain, 786.6 swelling, mass, or lump in chest, 786.7 abnormal chest sounds, 786.8 hiccough, and 786.9 other symptoms involving respiratory system and chest. After that, we filtered for those patients who had at least one encounter with one of these symptoms and more than two encounters before the age of 5, and were followed-up at least until the age of 8 years. Accordingly, the dataset size reduced significantly to 11,071 patients. The statistics and demographics of the study cohort are described in Table 4.1.

To demonstrate our interpretability approach on this data of pre-school children with respiratory system-related symptoms, we try to predict those children who will have asthma at school-age (cases) and those who will not have asthma at school-age (controls).

Cases were defined as children who had at least one encounter with respiratory system-related symptoms before the age of 5, and at least one encounter with asthma diagnosis ICD 493* after the age of 6. Controls were defined as children who had at least one encounter with respiratory system-related symptoms before the age of 5, and no diagnosis of asthma for at least three years after school-age, which is age 6. This definition splits our data into 6159 cases and 4912 controls. It is worth mentioning here that, for this specific cohort, the proportion of cases is relatively high (56%), compared to other cohorts or diseases, in which the prevalence of the disease is usually less.

The LSTM and BiLSTM models require longitudinal patient-level data that has been collected over time across several clinical encounters. Therefore, we processed the dataset to be in the format of list of lists of lists. The outermost list corresponds to

Table 4.1. Basic Statistics of the Cohort

		Cases	**Controls**
Number of patients		6159	4912
Number of visits		62962	42182
Number of diagnosis		128877	77038
Avg. Number of visits per patient		10.2	8.6
Avg. Number of codes in a visit		2.0	1.8
Gender	Female	2395	2278
	Male	3764	2634
Race	African American	2222	926
	Asian	56	56
	Biracial	83	43
	Caucasian	2361	2805
	Hispanic	602	454
	Native American	22	15
	Pacific Islander	8	2
	Unknown	805	611

patients, the intermediate list corresponds to the time-ordered visit sequence each patient made, and the innermost list corresponds to the diagnosis codes that were documented within each visit. Only the order of the visits was considered and the timestamp was not included.

Furthermore, deep learning libraries assume a vectorized representation of the data for time-series prediction problems. In our case, since the number of visits for each patient is different, we transformed the data such that all patients will have the same sequence length. This is done by padding the sequence of each patient with zeros so that all patients will have the same sequence length, equal to the length of the longest patient sequence. This vectorization allows the implementation to efficiently perform the matrix operations in batch for the deep learning model. This is a standard approach when handling sequential data with different sizes.

4.4. Experiments and Results

In this section, we first describe the experimental setup followed by the results of the models training. After that, we provide quantitative evidence of the benefits of using CD interpretations and explore the extent to which it agrees with baseline interpretations. Finally, we present our qualitative analysis including an interactive visualization and demonstrate its utility for explaining predictive models using individual visit scores and relative contributions of subset of visits.

4.4.1. Experimental Setup

We used the PyTorch implementation of the LSTM and BiLSTM models. We extended the implementation of Murdoch et al.[172] to decompose the BiLSTM models.

As the primary objective of this work is not predictive accuracy, we used standard best practices without much tuning to fit the models used to produce interpretations. All models were optimized using Adam [176] with learning rate of 0.0005 using early stopping on the validation set. The total number of input features (diagnosis codes) was 930 for ICD-9 3-digits format and 3318 for ICD-9 4-digits format. Patients were randomly split into training (55%), validation (15%), and test (30%) sets. The same proportion of cases (56%) and controls (44%) was maintained among the training, validation, and test sets. Model accuracy is reported on the test set, and area under the curve (AUC) is used to measure the prediction accuracy, together with 95% confidence interval (CI) as a measure of variability.

4.4.2. Models Training

To validate the performance of the proposed interpretability approach, we train LSTM and BiLSTM models on the asthma dataset, which has two classes: c=1 for cases, and c=0 for controls. In addition, we compare the prediction performance of these models with a baseline logistic regression model. The average AUC scores for 10 runs, with random seeds, on the full test set are shown in Table 4.2. Overall, the LSTM and BiLSTM models achieve higher AUC scores than baseline models such as logistic regression. Consequently, both models learned useful visit patterns for predicting school-age asthma.

4.4.3. Quantitative Analysis

In this section, we conduct quantitative analysis to (1) validate the contextual decomposition of the trained models, (2) evaluate the interpretations produced by the models,

Table 4.2. Average AUC of Models Trained on Asthma Dataset for the Task of School-age Asthma Prediction

Model	**AUC (95% CI)**
LSTM	0.831 (0.824-0.838)
BiLSTM	0.819 (0.811-0.827)
Logistic Regression	0.702 (0.692-0.712)

and (3) understand the extent to which the learned patterns correlate with other baseline interpretations.

Validation of Contextual Decomposition for BiLSTMs

Objective: To verify that the contextual decomposition of LSTMs and BiLSTMs works correctly with our prediction task, we designed a controlled experiment in which we add the same artificial visit to each patient of certain class, testing whether the contextual decomposition will assign a high attribution score to the artificial visit with respect to that specific class.

Given a patient p and a corresponding binary label c, we add an artificial visit v_{art} with one artificial diagnosis code d_{art} to each patient's visits list V. The d_{art} was chosen to be a synthetic diagnosis code which does not exist in the ICD-9 codes list. On the full dataset P, the artificial visit is added with probability p_{art} to patients with label 1, and with probability $1 - p_{art}$ to patients with label 0. As a result, when p_{art} = 1, all patients of class 1 will have v_{art}, and consequently the model should predict label 1 with a 100% accuracy and contribution of v_{art} should always be the maximum among other visits. Similarly, when p_{art} = 0.5, both classes will equally have patients with v_{art}, and therefore v_{art} does not provide any additional information about the label, and v_{art} should thus have a small contribution.

Experimental settings: We train LSTM and BiLSTM models on the asthma dataset with the artificial visit v_{art} setup. To measure the impact of v_{art}, we first add v_{art} to patients of class c=1, with probability p_{art}, varying p_{art} from 1 to 0.5 with steps of 0.1. After that, we train both models on this modified dataset, and then calculate the contribution of each visit by using the CD algorithm. We run the experiment 5 times with a different random seed and report on the average correct attribution. The attribution is correct if the highest contribution among all visits is assigned to v_{art}.

Results: The results of our evaluation are depicted in Figure 4.1. When p_{art} = 1, the models correctly attribute the prediction to the artificial visit at 100% accuracy. Moreover, as p_{art} becomes smaller, the contribution of the artificial visit goes down, since v_{art} becomes less important. Finally, when p_{art}= 0.5, the contribution of the artificial visit becomes irrelevant and the model attributes the prediction to other visits. Both models LSTM and BiLSTM perform similarly with 100% and 0% attribution accuracy at p_{art}= 1 and p_{art}=0.5, respectively. However, when p_{art} is between 0.8 and 0.6, BiLSTM attributes higher contribution to v_{art} than LSTM. This might be due to BiLSTM specific architecture, which accesses information in both forward and backward direction, allowing it to generate better inference about the visits importance with lower sensitivity to the position of v_{art}, compared to unidirectional LSTM. Overall, we can conclude that whenever there is a clear visit-level pattern, the models learn that pattern and the contextual decomposition can appropriately attribute the prediction to the correct visit.

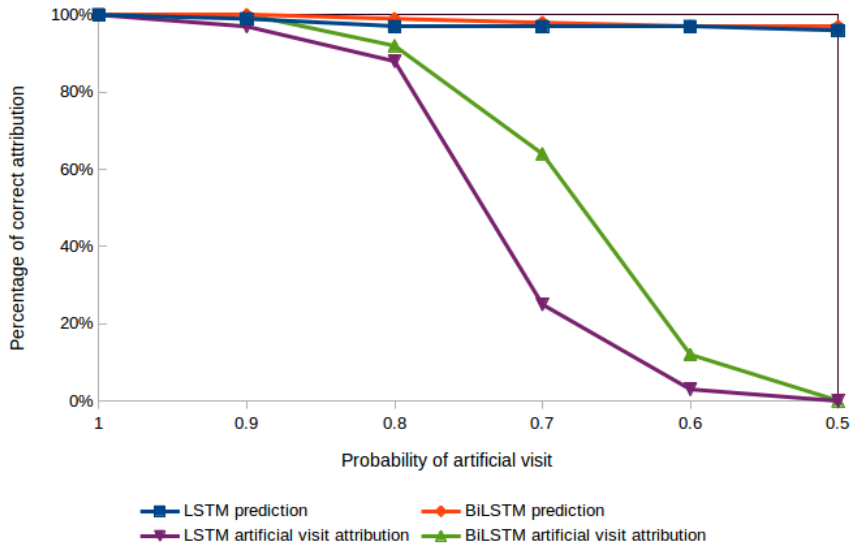


Figure 4.1. Validation of contextual decomposition for LSTM and BiLSTM for the class c=1. The attribution is correct if the highest contribution among all visits is assigned to the artificial visit. The prediction curves indicate the prediction accuracy for class c=1, which also represents the upper bound for the attribution accuracy.

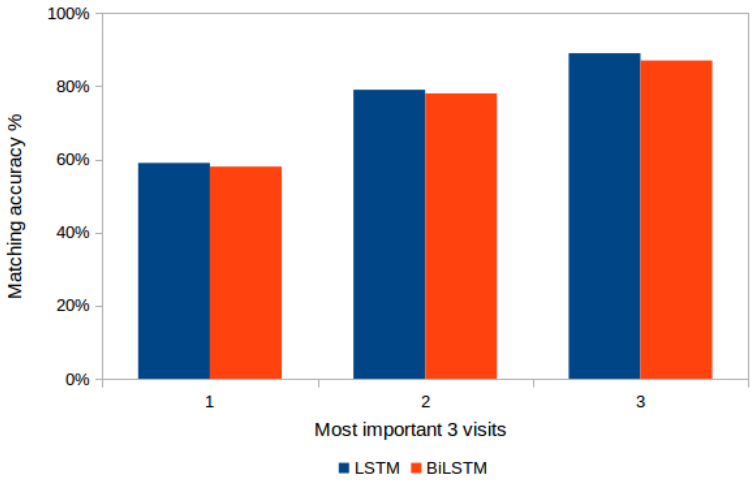


Figure 4.2. Evaluation of the agreement between CD scores and importance scores generated from logistic regression coefficients. The matching is correct if the visit with the highest LSTM/BiLSTM CD attribution matches one of the top three visits, which are generated using logistic regression coefficients.

Evaluation of Interpretations Extracted from BiLSTMs

Before examining the visit-level dynamics produced by the CD algorithm, we first verify that it compares favorably to prior work for the standard use case of producing coefficients for individual visits, using logistic regression. For longitudinal data such as EHR, a logistic regression model summarizes the EHR sequence ensemble to become aggregate features that ignore the temporal relationships among the feature elements. However, when sufficiently accurate in terms of prediction, logistic regression coefficients are generally treated as a gold standard for interpretability. Additionally, when the coefficients are transformed by an exponential function, they can be interpreted as odds ratio [177]. In particular, when applied to clinical outcomes prediction, the ordering of visits given by their coefficient value provides qualitatively sensible measure of importance. Therefore, when validating the interpretations extracted using the CD algorithm we should expect to find a meaningful correlation between the CD scores and the logistic regression coefficients. To that end, we present our evaluation of the interpretations extracted using the CD algorithm with respect to the coefficients produced by logistic regression.

Generating Ground Truth Attribution for Interpretation: Using our trained logistic regression model, we identified the most important three visits for each patient and used it as a baseline to evaluate the correlation between logistic regression coefficients and CD attributions. First, we calculated the importance score for each diagnosis code. After that we used these scores to calculate the importance score for each visit, by summing the importance scores of the diagnosis codes included in each visit. The importance score for each diagnosis code is calculated as follows:

- extract statistically significant diagnosis codes, using p-value criterion $p \leq 0.05$
- for all significant diagnosis codes, calculate coefficients and odds ratios
- filter for diagnosis codes with odds ratio > 1
- sort filtered diagnosis codes in descending order according to their odds ratios
- group the sorted diagnosis codes into 4 groups. Diagnosis codes with similar/closer odds ratios are grouped together
- assign an importance score for each group in descending order, based on the odds ratios of diagnosis codes in each group

Finally, we calculated the importance score for each visit, by summing the importance scores of the diagnosis codes occurred in that visit, and used the visits scores to identify the most important three visits for each patient. We run this analysis on a subset of 5000 patients, who have asthma, and for each patient the ground truth attribution baseline is the most important three visits, ordered according to their importance scores.

Evaluation: For each patient/ground-truth pair, we measured if the ground truth visits match the visit with the highest CD score for the same patient. We ranked the CD scores of visits for each patient and reported on the matching accuracy between the visit with the highest CD contribution and the three ground truth visits for each patient.

Results: The aggregated results for both LSTM and BiLSTM models are presented in Figure 4.2. Overall, we observe that, for the two models, the contextual decomposition attribution overlaps with our generated baseline ground truth attribution for at least 60% of the patient/ground-truth pairs. The matching between the top visit using the CD algorithm and the first top ground truth visit is 60%, the top two ground truth visits

is 80%, the top three ground truth visits is 90%. These results confirm that there is a strong relationship between the importance scores generated using logistic regression coefficients and the CD importance scores based on the patterns an LSTM/BiLSTM model learns.

4.4.4. Qualitative Analysis

After providing quantitative evidence of the benefits of CD to interpret the patient EHR visits importance, we now present our qualitative analysis using three types of experiments. First, we introduce our visualization and demonstrate its utility to interpret patient-specific predictions. Second, we provide examples for using our CD-based algorithm to find the most predictive subset of visits. Finally, we show that the CD algorithm is capable of identifying the top scoring visit patterns and demonstrate this in the context of predicting school-age asthma.

Explaining Predictions Using Individual Visit Scores

In this section, we present our interactive visualization and illustrate it with an example for both LSTM and BiLSTM models. The timeline in Figure 4.3 represents a patient's EHR time-ordered visits and the colors of the visits reflect the CD contributions of each visit to the predicted outcome. Moreover, hovering over the visits with the mouse will display the ICD codes documented by the clinician during the visit. Visualizing the CD contributions of each visit can be used to quickly explain why did the model make a certain prediction. For example, the patient shown in Figure 4.3 was correctly predicted to have asthma at school age. He had 19 data points (visits) before the age of six years and it was all considered by the model. The visualization indicated that visits

15 to 19 have the highest contribution to the prediction for both LSTM and BiLSTM models, and the ICD-9 codes included in these four visits are: 486 (pneumonia), 786 (symptoms involving respiratory system and other chest symptoms), 493 (asthma), and 465 (acute upper respiratory infections of multiple or unspecified sites). Presenting such information to the clinician could be of a great help in the decision making process. For example, this specific patient has been following up at the hospital from age 0 to 5 years, and he had respiratory-related complications throughout the 5 years. Typically, the physician will have to check the full history of a patient to understand the patient condition and make a decision. In contrast, visualizing the CD scores for each visit as shown in Figure 4.3 indicates that, for this specific patient, older visits are not very relevant. The visualization highlights that recent visits are more important to examine. This is probably due to the fact that continuing to have respiratory complications till age 5, just before school-age, is an important indication that this patient will likely continue to have asthma at school age.

Explaining Predictions Using Relative Contributions of Subset of Visits

In this section, we first present our results for the implementation of the algorithm introduced earlier for finding the most predictive subset of visits, and then we qualitatively compare between the relative contributions of the subset of visits produced by LSTM and BiLSTM.

Figure 4.4 shows an example of a patient who was correctly predicted to have asthma at school-age. The patient made 14 visits between age 0 and 5 with different complications. The individual visit scores do not provide clear information about the critical time window which the physician needs to examine. However, using our

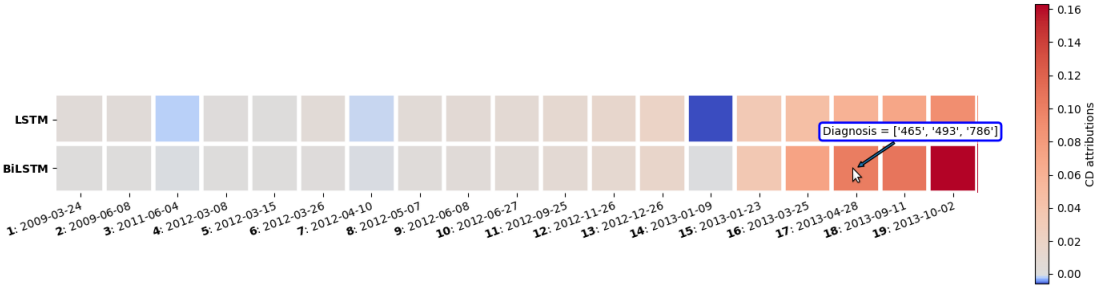


Figure 4.3. CD scores for individual visits produced from LSTM and BiLSTM models trained for the task of predicting school-age asthma. Red is positive, white is neutral and blue is negative. The squares represent patient EHR time-ordered visits, and the label of each square indicates the visit number appended by the date of the visit. The upper row is the LSTM CD attributions and the lower row is the BiLSTM CD attributions

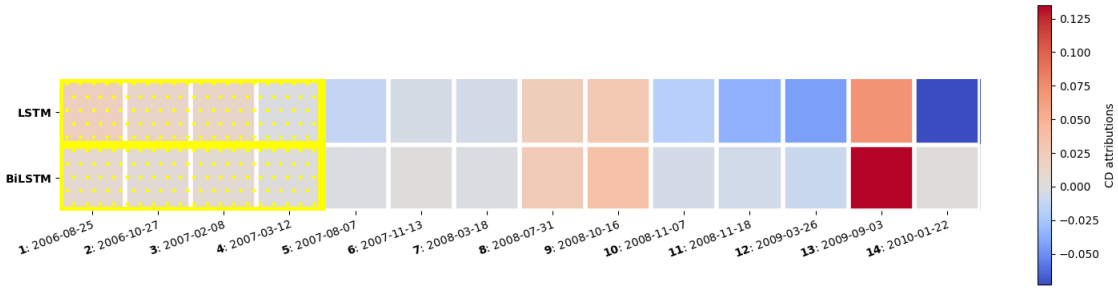


Figure 4.4. Most predictive subset of visits using CD-based scores highlighted in yellow. Example for a patient where relative contributions of subset of visits produced from LSTM and BiLSTM are similar.

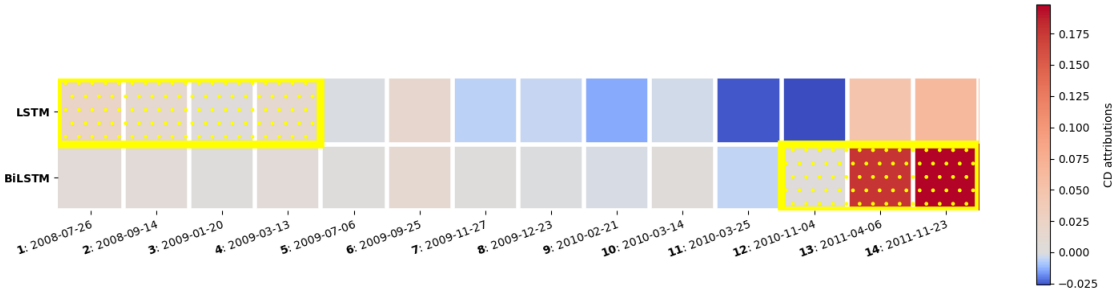


Figure 4.5. Most predictive subset of visits using CD-based scores. Example for a patient where BiLSTM is producing better interpretation than LSTM.

algorithm for finding the most predictive subset of visits, the algorithm identified that grouping visits 1 to 4 together (highlighted in yellow) produced the maximum relative contribution to the predicted outcome, compared to other subset of visits. The ICD codes

Table 4.3. Top scoring patterns of length 1 visit, produced by the contextual decomposition of LSTM and BiLSTM models on the asthma data

	LSTM		BiLSTM	
	ICD Codes	**Frequency%**	**ICD Codes**	**Frequency%**
1	493.9 Asthma Unspecified	40%	493.9 Asthma Unspecified	34%
2	493.9,786.0 Asthma Unspecified, Dyspnea and Respiratory Abnormalities	13%	786.2 Cough	15%
3	786.0 Dyspnea and Respiratory Abnormalities	11%	493.9,786.0 Asthma Unspecified, Dyspnea and Respiratory Abnormalities	21%
4	493.9,786.2 Asthma Unspecified,Cough	10%	786.0 Dyspnea and Respiratory Abnormalities	10%
5	465.9,493.9 Acute Upper Respiratory Infections of Unspecified Site, Asthma Unspecified	9%	493.9,786.2 Asthma Unspecified, Cough	9%
6	493.0 Extrinsic Asthma	4%	465.9,493.9 Acute Upper Respiratory Infections of Unspecified Site,Asthma Unspecified	8%
7	486,493.9 Pneumonia, Asthma Unspecified	4%	465.9,786.2 Acute Upper Respiratory Infections of Unspecified Site,Cough	5%
8	465.9,493.9,786.2 Acute Upper Respiratory Infections of Unspecified Site, Asthma Unspecified, Cough	3%	493.0 Extrinsic Asthma	4%
9	382.9,493.9 Unspecified Otitis Media, Asthma Unspecified	3%	486,493.9 Pneumonia, Asthma Unspecified	3%
10	493.0, 493.9 Extrinsic Asthma,Asthma Unspecified	3%	V67.9 Unspecified Follow-Up Examination	3%

included in these visits indicated that this patient has been diagnosed with congenital anomalies as well as asthma before the age of 1, followed by organic sleep disorders and symptoms involving respiratory system and chest in the following years. Therefore, although the contributions of individual visits were not high, the relative contribution of grouping the visits together provided useful information to explain the prediction.

In general, we found that the relative contributions of subset of visits extracted from BiLSTM and LSTM are often similar. However, for some cases, such as the patient shown in Figure 4.5, we observed that contributions produced from BiLSMT are likely more clinically relevant than LSTM. This is possibly because BiLSTM mimics physician practice by examining the EHR clinical visits not only in forward time order, but also

Table 4.4. Top scoring patterns of length 2 visit, produced by the contextual decomposition of LSTM and BiLSTM models on the asthma data

	LSTM		**BiLSTM**	
	ICD Codes	**Frequency%**	**ICD Codes**	**Frequency%**
1	[493.9], [493.9] [Asthma Unspecified],[Asthma Unspecified]	13%	[493.9], [493.9] [Asthma Unspecified],[Asthma Unspecified]	11%
2	[493.9,786.0],[493.9] [Asthma Unspecified, Dyspnea and Respiratory Abnormalities], [Asthma Unspecified]	2%	[493.9,786.0],[493.9] [Asthma Unspecified, Dyspnea and Respiratory Abnormalities], [Asthma Unspecified]	2%
3	[493.9],[493.9,786.0] [Asthma Unspecified], [Asthma Unspecified, Dyspnea and Respiratory Abnormalities]	2%	[493.9],[493.9,786.0] [Asthma Unspecified], [Asthma Unspecified, Dyspnea and Respiratory Abnormalities]	2%
4	[493.9], [V20.2] [Asthma Unspecified], [Routine Infant or Child Health Check]	2%	[493.9], [V20.2] [Asthma Unspecified], [Routine Infant or Child Health Check]	2%
5	[493.9,786.2], [493.9] [Asthma Unspecified, Cough], [Asthma Unspecified]	2%	[493.9,786.2], [493.9] [Asthma Unspecified, Cough], [Asthma Unspecified]	1%

considers the backward time order so that recent clinical visits are likely to receive higher importance.

Identifying Top Scoring Patterns

We now demonstrate the utility of using the CD attributions to identify the top scoring patterns which was learned by the LSTM and BiLSTM models. To address this, we analysed for each patient for which the class c=1 (having asthma at school age) was correctly predicted, which visit patterns of length one and two visits had the highest positive contribution towards predicting that class. To obtain the important features in the cohort level, we count the frequency of the top-ranked features in all patients for one and two visits. The results of this evaluation are summarized for one visit patterns in Table 4.3 and two visits patterns in Table 4.4. Overall, both models learn similar patterns for both length one and two visits with no significant difference. Moreover, the identified

patterns are inline with the risk factors suggested in the literature for school-age asthma [178]–[180].

4.5. Discussion

In this study, we assessed the potential application of contextual decomposition (CD) method to explain patient-specific risk predictions using quantitative and qualitative evaluation. Our results demonestrated that whenever a clear visit-level pattern exists, the LSTM and BiLSTM models learn that pattern and the contextual decomposition can appropriately attribute the prediction to the correct pattern. In addition, the results confirm that the CD score agrees to a large extent with the importance scores produced using logistic regression coefficients. Our main insight was that rather than interpreting the attribution of individual patient visits to the predicted outcome, we could instead attribute a model's prediction to a group of visits.

Recent work on interpretability of machine learning models in healthcare through visual analytics [165], [181]–[183] shows that such interpretability methods could adequately provide healthcare professionals with insights into the predictions produced by deep learning algorithms. Our contextual decomposition approach provides visualizations that facilitate interaction with algorithms and has several advantages compared to other state-of-the-art approaches. For example, although attention mechanisms offer the possibility to retrieve deriving features at patient-level where individual risk prediction can be paired to individual attention map [40], those interpretations focus on individual risk factors and it's not straightforward to extend it to subsets of risk factors like what we demonstrated using the CD approach. In addition, using the interpretation of activations [184] at the single subject level is not straightforward and more insightful

clinical interpretations of single cases require the knowledge of the complete clinical history of patient. On the other hand, our CD approach can directly highlight evidence for important features deriving the predictions in the patient's EHR history.

A potential limitation of our study is the identification of asthma patients using ICD codes. In particular, although using ICD codes to identify asthma is a popular practice in large-scale epidemiologic research, previous research showed that using ICD-9 codes have a moderate accuracy of identifying children with asthma, compared to criteria-based medical record review [185]. In addition, the contextual decomposition approach was demonstrated on a single cohort of patients. Generalizing the findings and explanations of this study would require assessing multiple datasets representing multiple cohorts, diseases, and age groups.

4.6. Summary

In this chapter, we have proposed using contextual decomposition (CD) to produce importance scores for individual visits and relative importance scores for a group of visits, to explain decisions of risk prediction models. In addition, we developed an interactive visualization tool and demonstrated, using a concrete case study with real EHR data, how CD scores offer an intuitive visit-level interpretation. This movement beyond single visit importance is critical for understanding a model as complex and highly non-linear as BiLSTM. The potential extension of our approach to other sources of big medical data (e.g. genomics and imaging), could generate valuable insights to assist decision-making for improved diagnosis and treatment.

CHAPTER 5: TIME-AWARE PATIENT REPRESENTATION IN EHRS USING SELF-ATTENTION AND NON-STATIONARY KERNEL APPROXIMATION

5.1. Overview

Effective modeling of patient representation from electronic health records (EHRs) is rapidly becoming an important research topic. Yet, modeling the non-stationarity in EHR data has received less attention. Existing studies follow a strong assumption to implicitly assume the stationarity in patient representation from EHRs. However, in reality, the visits of a patient are irregularly distributed over a relatively long period of time and disease progression patterns are non-stationary. Moreover, the time span between patient visits often reflects important domain knowledge and may have significant implications on discovering unknown patterns that might characterize a medical condition. To bridge the gap between modeling stationary and non-stationary event sequences in EHRs, we leverage self-attention mechanism and non-stationary kernel approximation to capture contextual information as well as temporal relationships between patient visits in EHRs. To validate the effectiveness of our proposed approach, we use a real-world EHR dataset for the task of predicting the next diagnosis code for a patient, given the patient's EHR history. We compare our method against self-attention with positional encoding and self-attention with stationary kernels approximation. The experimental results demonstrate that the proposed approach provides a better predictive performance, compared to all baselines. These findings confirm the effectiveness of the proposed method and emphasize the importance of modeling non-stationary time information in healthcare prediction tasks.

5.2. Introduction

With the rapidly increasing volume of clinical information captured in Electronic Health Records (EHRs), various deep learning models have been developed and applied for a wide variety of healthcare prediction tasks [6]–[12]. EHRs include continuous-time events and the time lapse between sequential visits often has hidden patterns (e.g. disease progression or changing variables over time) which could be utilized for risk prediction models. For example, two consecutive diagnosis codes do not indicate they are temporally close, but the correlations between them can be revealed by the time lapse between visits. In addition, medical codes have varying temporal context. For example, flu is a short-lived condition whereas a diagnosis code for a more chronic condition such as diabetes has a longer span and therefore might be present repeatedly over multiple patient's visits.

A patient's record in EHR can be viewed as a sequence of time-ordered visits and each visit consists of a number of unordered medical codes describing the patient conditions at a given time point. A common modeling approach for such structured data is to aggregate medical codes within a visit into a vector. Then the visits are often fed sequentially into deep learning models, such as recurrent neural networks (RNNs). However, given that time gap between consecutive visits in patient records can vary from days to years, the architecture of traditional RNNs fails to handle the irregular time intervals between visits. Several approaches have been proposed to incorporate temporal information into RNNs [8], [23], [96], [168], [186]. Although these techniques have demonstrated that augmenting RNNs with time information can improve their predictive performance, most of these methods are designed to handle data with constant time gaps.

Thus, they are unable to fully exploit the temporal non-stationary dynamics of the EHR data, thus hindering the learning of the prediction algorithms under non-stationarity.

Recently, sequential modeling of EHRs with self-attention [110] has been applied to derive more robust patient representations in many prediction tasks [187]–[189]. However, self-attention uses positional encoding to model the order of the input sequences, limiting the self-attention mechanism to modeling only stationary or discrete-time event sequences. Hence, like many other sequence models, self-attention does not handle the irregular time spans between visits and thus models sequential signals rather than temporal patterns. It is therefore natural to consider alternative approaches to the positional encoding that can incorporate non-stationary temporal patterns into the patient's representation in EHRs.

In this chapter, we propose a method for time-aware attention-based visit representation using self-attention mechanism combined with non-stationary kernel approximation. Specifically, we extended the self-attention architecture by replacing the positional encoding with non-stationary kernel approximation to capture contextual information as well as non-stationary temporal relationships between patient visits in EHR. Using this approach, self-attention embeds time information into visit representations to learn local attention weights for each visit and then a complete representation for each patient. Embedding continuous non-stationary time information using kernel approximation effectively avoids the drawbacks introduced by positional encoding, which does not account for the non-stationarity characteristics of EHR data.

5.3.1. Self-attention Mechanism

Self-attention is an attention mechanism which relates every position in a sequence to every other position in the sequence, and reweighs the position embeddings of each position to include contextual relevance [110]. Self-attention is computed using dot-product attention [91] over a query vector $\boldsymbol{Q}$, a key vector $\boldsymbol{K}$, and a value (representations) of events in a sequence vector $\boldsymbol{V}$, defined as:

$$Att(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}})\boldsymbol{V} \tag{5.1}$$

Self-attention is typically used in conjunction with positional encoding to ensure the order of the sequence is maintained, where the vector representation for each position is combined (added or concatenated) with the corresponding event embedding. Thus, providing the order context to the non-recurrent architecture of the self-attention mechanism [91]. It accomplishes this through a series of key, query, and value weight matrices via three steps: 1) dot product similarity of query-key pairs to find alignment scores, 2) normalization of the scores to get the weights, and 3) reweighing of the original embedding using the weights for combining event values.

In the context of modeling EHRs, a visit-position representation is formed by combining visit embedding with positional encoding. However, this visit-position representation can only capture the sequential patterns and does not handle irregular time intervals between consecutive visits.

5.3.2. The Essence of Stationary Kernels vs Non-Stationary Kernels

Kernel methods are powerful tools in machine learning and provide a framework for non-linear learning for a wide spectrum of machine learning and data mining tasks. However, they often suffer from scalability due to their significant computational cost. Approximating kernels with explicit feature maps provides a practical alternative for using kernels in large-scale applications [190], [191]. However, for many types of kernels, deriving such feature maps remains a challenging problem.

The most general family of kernels is the non-stationary kernels, which include stationary kernels as special cases. Stationary kernels are a class of functions in which correlations of data depend solely on the distance separating the two examples x and y and not on the respective location of the examples. Stationary kernels are of the form:

$$k = k(\|x - y\|) \tag{5.2}$$

In contrast, non-stationary kernel functions depend on both the distance separating the two examples x and y as well as the respective location of the examples. Non-stationary kernels are of the form:

$$k = k(x, y) \tag{5.3}$$

where the kernel function relaxes the restriction so that $k = k(x, y) \neq k(\|x - y\|)$.

Stationary and non-stationary kernels are chosen to be symmetric positive-definite functions. They induce inner products in Hilbert space and therefore impose a metric, which can be interpreted as a similarity measure. Stationary kernel functions are

constant along diagonals, unlike non-stationary kernels. This characteristic of non-stationary kernels translates into potentially highly flexible inner products, and therefore similarity measures, which can encode covariances within the input space, the output space and across the two spaces.

The flexibility of the kernel functions and the associated similarity measures motivated us to use them for modeling non-stationary EHRs, as an efficient alternative to overcome the limitations of the positional encoding component in the self-attention model.

5.4. Methods

We propose an extension to self-attention to consider the varying time gaps between two visits in a patient's sequence. We first outline our approach for constructing the time-visit embeddings. After that, we identify the feature maps for stationary and non-stationary kernels which will be used in our analysis. Finally, we show how a time-visit representation can be used for a prediction task.

5.4.1. Time Embedding

Finding a time embedding from an interval $T = [0, t_{max}]$ to $\mathbb{R}^d$, assuming time starts from origin, can be considered as identifying a mapping $\Phi : T \rightarrow \mathbb{R}^d$. Forming a visit embedding $Z \in \mathbb{R}^{d_E}$ involves constructing a vector representation Z_i for visit v_i, $i = 1, \ldots, N$, where N is the total number of visits for a patient. To build a visit-time embedding, time embeddings are concatenated with visit embeddings so that the dot product between two time-dependent visits (v_1, t_1) and (v_2, t_2) can be defined as $[Z_1, \Phi(t_1)]'[Z_2, \Phi(t_2)] = \langle Z_1, Z_2 \rangle + \langle \Phi(t_1), \Phi(t_2) \rangle$. Since contextual information

between visits are captured by $\langle Z_1, Z_2 \rangle$, we expect that temporal relationships are captured by $\langle \Phi(t_1), \Phi(t_2) \rangle$, particularly those associated with the temporal difference $t_1 - t_2$. We formulate those temporal relationships using feature map approximations Φ for kernels $\mathcal{K}$.

Let the kernel be $\mathcal{K} : T \times T \rightarrow \mathbb{R}$, where $\mathcal{K}(t_1, t_2)$ can be approximated with some explicit feature mapping $\langle \Phi(t_1), \Phi(t_2) \rangle$, and $\mathcal{K}(t_1, t_2) = \psi(t_1 - t_2), \forall t_1, t_2 \in T$ for some $\psi : [-t_{max}, t_{max}] \rightarrow \mathbb{R}$. Here, the mapping of original data into a higher dimensional space is captured by Φ. The intuition here is to convert the task of learning temporal embeddings into a kernel learning problem approximated with feature map Φ. We aim to identify candidate Φ with some functional forms that are appropriate with current deep learning techniques which rely on computations via back-propagation. In addition, the interactions between temporal context and visit context can now be captured with some other mappings as $(Z, \Phi(t)) \rightarrow f(Z, \Phi(t))$, as discussed in Section 5.4.3.

5.4.2. Kernel Approximation

In this section, we provide an overview of the kernels used in this work and its corresponding feature map approximations. We do not aim to provide the explanations behind the different numerical representations and mathematical operations in the proposed kernels. We refer the reader to the original papers of each of these kernels for more details.

Approximation of Stationary Kernels

In the context of large-scale kernel machines, [191] introduced a framework for a random feature map (random Fourier), based on the Bochner's theorem [192], that

approximates any stationary kernel (e.g. Gaussian, Laplace, and Cauchy) function via independent sampling of the probability distribution. To obtain the stationary feature maps motivated by Bochner's time encoding, we follow the approach proposed in [193]. In short, the authors propose to realize the Bochner feature maps using the reparametrization trick as well as parametric and nonparametric inverse cumulative distribution function (CDF) transformation. The Bochner feature maps using the reparametrization trick can be written as:

$$\phi(t) = [cos(\omega_i(\mu)t), sin(\omega_i(\mu)t)] \tag{5.4}$$

where μ is the location-scale parameters specified for the reparametrization trick, $\omega_i(\mu)$ converts the $i-$th sample (drawn from auxiliary distribution) to target distribution under location-scale parameter.

In addition, the Bochner feature maps using the parametric inverse CDF transformation can be written as:

$$\phi(t) = [cos(g_\theta(\omega_i)t), sin(g_\theta(\omega_i)t)] \tag{5.5}$$

where θ is the parameters for the inverse CDF $F^{-1} = g_\theta$, and ω_i is the $i-$th sample drawn from auxiliary distribution.

Approximation of Non-stationary Kernels

The polynomial kernel is a non-stationary kernel which is specially important, because any kernel can be written as a sum of polynomial kernels through a Taylor expansion [194]. If we can appropriately use the feature map approximations of the

polynomial kernel, then we will be able to effectively utilize many other types of kernels. Approximating polynomial kernels with explicit nonlinear maps is a challenging problem [195], however considerable progress has been made in this area recently [190], [194], [196]. Here, we follow the approach described in [196], where polynomial feature maps are derived based on Random Maclaurin technique as follows: Let $k(t_1, t_2) = (\langle t_1, t_2 \rangle + c)^d$ for some constant $c \geqslant 0$ and positive integer d, which is the degree of the polynomial kernel function (e.g., d = 2 for quadratic). Then, the feature map of the second degree polynomial kernel can be approximated as:

$$\phi(t) = [c, \sqrt{2}\sqrt{c}t, t^2] \tag{5.6}$$

Higher-order polynomial kernels can be approximated similar to the quadratic, therefore the third degree polynomial kernel can be approximated as:

$$\phi(t) = [c^{\frac{3}{2}}, \sqrt{3}ct, \sqrt{3}\sqrt{c}t^2, t^3] \tag{5.7}$$

and the fourth degree polynomial kernel can be approximated as:

$$\phi(t) = [c^2, 2c^{\frac{3}{2}}t, \sqrt{6}ct^2, 2\sqrt{c}t^3, t^4] \tag{5.8}$$

5.4.3. Time-Visit Embeddings

To expose time interactions and examine if they improve the predictive performance, we need to combine the visit embeddings generated by the attention with the time embeddings generated by the feature maps. After embedding continuous time intervals between visits into finite-dimensional vector spaces, time and visit embeddings can

be used to model the interactions between visits and its associated time intervals. In order to recognize these interactions, the visit and time representations need to be projected onto a common space. For a patient sequence $\{(v_1, t_1), \ldots, (v_q, t_q)\}$, the visit and time embeddings are concatenated as $[\boldsymbol{Z}, \boldsymbol{Z_T}]$ where $\boldsymbol{Z} = [Z_1, \ldots, Z_q]$ and $\boldsymbol{Z_T} = [\Phi(t_1), \ldots, \Phi(t_q)]$. The concatenated representation is then projected into the query, key and value spaces of self-attention. For example, to capture linear relationships between visit and time representations in query space, we can use $\boldsymbol{Q} = [\boldsymbol{Z}, \boldsymbol{Z_T}]\boldsymbol{W_0} + b_0$. Moreover, nonlinear associations between visit and time representations can be modeled hierarchically, e.g. using a multilayer perceptrons (MLP) with activation functions, as $\boldsymbol{Q} = ReLU([\boldsymbol{Z}, \boldsymbol{Z_T}]\boldsymbol{W_0} + b_0)\boldsymbol{W_1} + b_1$

5.5. Experiments and Results

In this section, we introduce our experimental setup and present our empirical results.

5.5.1. Dataset

The dataset consists of electronic health records from Cerner Real World database.The extracted cohort is not disease-specific and patients were selected randomly. We exclude patients with less than three visits. The total number of patients used to train our models is 11451 patients with a total of 3485 unique diagnosis codes.

5.5.2. Prediction Task

Given a sequence of visits for a patient $v_1, \ldots, v_T$, the task is to predict the diagnosis codes occurring at the next visit $v_2, \ldots, v_{T+1}$, for each time step i.

5.5.3. Implementation Details

The patients cohort was split into the training (70%), validation (10%) and test (20%) sets. The maximum number of visits for a patient was set to be 100. The following parameter values were used for stationary and non-stationary models: hidden size= 72, sequence embedding size= 72, time embedding size= 72, batch size= 128, learning rate= 0.001, number of attention blocks= 2, and number of heads in each attention blocks= 1. We performed hyper-parameter tuning on the self-attention with positional encoding model using grid search, and the following parameters provided the best results: hidden size= 100, sequence embedding size= 100, batch size= 128, learning rate= 0.001, number of attention blocks= 1, and number of heads in each attention blocks= 1.

5.5.4. Evaluation Metrics

We used two evaluation metrics: Normalized Discounted Cumulative Gain at N (NDCG@n) and Hit Ratio at N (Hit@n). The NDCG metric considers the position and penalizes highly relevant diagnosis codes appearing lower in the prediction results, as the graded relevance value is reduced logarithmically proportional to the position of the result. The Hit@n metric computes how many "hits" the model achieved in an n-sized list of ranked diagnoses. For our evaluation, we will use $n = 10$.

5.5.5. Results

We examine seven methods for time encoding which are divided into three categories. The first category is the original self-attention implementation with positional encoding. The second category is stationary kernels, which includes three variations of Bochner's feature maps. The third category is three non-stationary kernels, which are polynomials

Table 5.1. Performance of the Proposed Approach and Baseline Models.

Category	Kernel	NDCG@10 (SD)	Hit@10 (SD)
Positional encoding	-	0.6923 (.0023)	0.8901 (.0031)
Stationary	Bochner (location-scale)	0.7030 (.0041)	0.8963 (.0026)
	Bochner (parametric inverse CDF)	0.6980 (.0057)	0.8934 (.0033)
	Bochner (non-parametric inverse CDF)	0.7120 (.0040)	0.9005 (.0021)
Non-stationary	Polynomial, d=2	0.7507 (.0028)	0.9160 (.0047)
	Polynomial, d=3	0.7501 (.0030)	0.9153 (.0035)
	Polynomial, d=4	0.7485 (.0021)	0.9163 (.0045)

with degree 2, 3, and 4. Results computed over ten runs are reported in Table 5.1.

Results show that modeling time information using non-stationary kernels improves the performance by 8% for NDCG and 3% for Hit metrics. Statistical significance test was conducted to compare the performance of the non-stationary polynomial kernels with positional encoding and stationary kernels over the NDCG@10 and Hit@10 metrics. In order to confirm that the improvement on predictive performance using our proposed model is statistically significant compared to baselines, we conducted a statistical significance test using the Mann-Whitney-Wilcoxon test. The computed p-values were all lower than 5%, indicating that the improvement introduced by using non-stationary kernels is statistically significant. In addition, results demonstrate that modeling the time information with stationary kernels does not result in significant gain in performance, over positional encoding. This observation confirms that stationary kernels may not be sufficient to model the non-stationarity characteristics of EHR data. Moreover, results show that a second degree polynomial may be sufficient to model the temporal information in our dataset, without the need to use higher order polynomials.

5.6. Discussion

In this chapter we introduce a method for time-aware patient representation in EHRs, using non-stationary time embedding and visit-level self-attention model, to handle the irregular time lapse between visits and better model patient representations in EHRs. Our model first generates embeddings for the discrete medical codes to form a visit representation and models the temporal information using feature maps of non-stationary kernels. To generate the visit-time representation for a patient, we feed visit embeddings and its corresponding time embeddings into a visit-level self-attention layer.

Experimental evaluation demonstrates that in a setting that models non-stationary data distributions, methods such as kernel approximation can boost the predictive power compared to the best base models. In addition, we find that self-attention models can capture informative visit interactions in our dataset, however, the benefit from such models is reduced under non-stationarity characteristics of EHRs. These findings emphasize the importance of accounting for the irregular data distributions when modeling patient representations in EHRs.

As part of transforming raw EHR data into a time-aware representation of a patient, the task boils down to generating time series features. Many studies have been undertaken to address this issue [197]–[200] . Some of them [197] build on temporal abstraction [201] to define patterns that can describe temporal relationships among multiple time series. Typically, such methods yield temporal patterns, such as "the occurrence of clinical event X precedes the drop of clinical event Y", which are then analyzed to determine the most informative pattern for the classification task. Other approaches [198], [202] transform time-stamped data points into symbolic time intervals,

then identifies time-interval-related patterns that can be used to build a classification model. Another group of methods [199] uses a graph-based framework, where the temporal relationships are represented as temporal graphs Furthermore, some techniques [200] learn temporal weights for the clinical events of interest, which they then apply in various ways to develop features that can be used by classifiers. In contrast to prior studies, our methodology attempts to take into account time series of different lengths and irregularity of distribution within and across features extracted from the richly structured EHR data for prediction.

5.7. Summary

We described a new technique to model the non-stationery time information in EHRs for prediction tasks. Our approach is based on extending the self-attention mechanism to handle the irregular time intervals between consecutive visits in a typical EHR of a patient. We expect that the proposed approach could easily extend beyond polynomial kernels, and the same technique would apply equally well to other non-stationary kernels, given that their feature maps are well derived. Future work may consider exploring other types of kernels and other methods for kernel approximations, which could be more efficient in modeling the non-stationarity patterns in EHRs.

CHAPTER 6: DISCUSSION

Developing deep learning predictive models on EHR data presents several modeling challenges, including data heterogeneity, data irregularity, and model interpretability. To alleviate these issues, we propose several EHR-based interpretable deep learning predictive models using modelling techniques that can discover and consider complex nonlinear interactions among a large number of variables in EHRs. Powered by attention mechanism (including self-attention), contextual decomposition interpretability method, and kernel approximation, our proposed methods can incorporate the entire medical history of each patient, represented as a series of heterogeneous data packed in irregular intervals, to predict several important clinical outcomes (e.g. preterm birth, school-age asthma). Patient-specific interpretations obtained from the proposed methods offer a great potential to support clinicians in providing informed data-driven clinical decisions.

While there is not much agreement on the definition of interpretability in machine learning, there are a few characteristics of interpretable models that researchers have discussed which can be utilized as a reference to derive the requirements of interpretable models [203]–[205]. In this thesis, modeling complex sequences, such as EHRs, necessitated using attention mechanisms, as an established approach to support deep neural networks with essential interpretability [206]. However, we emphasize that the choice of interpretability mechanism highly depends upon the application and use case for which explanations are required. Hence, a critical application like predicting a patient's death may have much more strict requirements for explanation reliability, as compared to just predicting costs for a procedure where getting the prediction right is much more important than providing explanations. There are still many questions that are unaddressed within the area of interpretable models in the clinical domain, and we foresee that it will

be an active area of research for the next few years.

Stationarity is an important concept in longitudinal data analysis. However, detecting stationarity in the data is a complex task. Therefore, researchers might be forced to make strong assumptions about the data, and rather than deciding between two strict options, they may be able to determine, with high probability, that a series is generated by a stationary or non-stationary process. As a result, when we are unsure about the likelihood or presence of non-stationarity, it may be advisable to stick with simpler models. On the other hand, when we are confident that the data distribution is not stable and will continue as such for some time, more complex models, including methods such as kernel approximation, may provide considerable improvements in discriminative power. In this thesis, under the known non-stationarity exhibited in the EHR datasets, the polynomial kernels outperform the baselines of standard self-attention with positional encoding and extended self-attention with stationary kernels, for the task of predicting the set of diagnosis codes in the next visit. This confirms that ignoring non-stationarity can result in sub-optimal model selection. We highlight, however, that we do not endorse a specific kernel approximation method as being the most resistant to non-stationarity. Moreover, it is difficult to conclude that these results generalize to different tasks. Rather, we explored how best to model the temporal non-stationary patterns in EHR data by investigating different methods for modeling the temporal characteristics of EHR data.

Generalizability of predictive models in healthcare is not a binary concept and has no universally agreed upon definition. According to one common hierarchy [207], a set of rules from a machine learning algorithm, clinician, or user may be applicable: internally, applying only in the limited context in which it was developed; temporally, applying prospectively at the center in which it was developed; or externally, applying

both at new centers and in new periods of time. There are other hierarchical systems that construct even deeper levels of generalizability [208]. A natural solution to these limitations is to integrate EHR data from multiple healthcare facilities with diverse demographics and multiple regions. In this thesis, we have utilized the Cerner Real World Data [118] which maintains a large volume of EHR clinical data from more than 500 healthcare facilities, improving the generalizability of our predictive models.

The term "effectiveness" has been used in this thesis to refer to the relative effectiveness compared to some baseline. Yet, what is to be considered as a baseline is not always obvious when working with EHR data, which implies that a standard baseline that is well recognized is often missing. This is unfortunately a known issue when developing EHR-based predictive models, where access to large repositories of clinical data for research purposes is very restricted due to data sensitivity and ethical issues. Consequently, this has resulted in limited research in the exact same domain and a small number of methods to which the proposed methods can be compared. In this thesis, a baseline is sometimes a self-proposed, relatively traditional ML method or the best method observed in a previous study in the same context. Therefore, this thesis proposes a number of models that aimed to outperform their former models.

Furthermore, this thesis demonstrates how large longitudinal EHR repositories, such as the Cerner Real World Data, can be harnessed to develop EHR-based predictive models using modern machine learning algorithms. Such retrospective observational data will speed up both the rate at which these risk prediction models are developed as well as our understanding of the underlying causes of diseases. EHR data allows the predictive models to consider more observations (clinical predictors and outcomes), on more patients, at more time points, and at a fraction of the expenses of prospective cohort

studies [209]. Additionally, patient populations extracted from such EHR repositories may be more representative of the real-world compared to cohort studies that depend on volunteer participation [210].

We are at a very exciting time – the rapid adoption of EHR systems is generating enormous amounts of unexplored longitudinal health data. Therefore, naturally, deep learning learning techniques, combined with massive EHR data have the potential to revolutionize the quality of our healthcare systems through improved personalized care for everyone.

CHAPTER 7: CONCLUSION AND FUTURE WORK

To conclude the thesis, this chapter presents a brief recapitulation of the main contributions of the thesis, followed by suggestions for future research.

7.1. Recapitulation

Electronic health records are an increasingly comprehensive data source for the development of new clinical risk prediction models, presenting both unique analytic opportunities and challenges. Modeling EHR data to infer health trajectories requires coping with numerous issues simultaneously. In this dissertation, we focus on addressing three important challenges: data heterogeneity, data irregularity, and model interpretability. Specifically, we utilize state of the art deep learning techniques and modern machine learning methods to develop efficient and interpretable predictive models demonstrating how longitudinal clinical data contained in EHRs can be harnessed for providing patient-specific predictions and interpretations for several clinical outcomes.

We first introduced a code-level attention-based recurrent neural network model for prediction of future clinical outcomes (Chapter 3). This method was developed to address two challenges of using temporal EHRs, namely **data heterogeneity** and **model interpretability**. The former was addressed by exploiting the attention mechanism to learn the importance weights of the heterogeneous data included in a patient's sequence of visits and embed contextual information into each code accordingly. The attention weights help the model better identify and learn the significant variables, thus, improving its predictive power. The latter was addressed by utilizing the attention mechanism to generate temporal code-level and visit-level explanations for the predicted outcomes, using attention weights. The method was demonstrated on the task of predicting the

risk of preterm birth, up to 9 months earlier than its occurrence, using data routinely collected in EHR. Our results demonstrate that attention components added on top of existing standard recurrent neural network algorithms provide interpretable patterns that incorporate important insights into explaining the predicted outcomes. In addition, it shows that integrating attention mechanisms exhibits a complementary effect that contributes to the improvement of models' predictive power.

To further address the **model interpretability** challenge, we introduced an additional method for interpreting the outcomes of recurrent neural networks, particularly BiLSTMs (Chapter 4). The proposed technique is based on the contextual decomposition interpretation algorithm [172], where we decompose the computations of the standard BiLSTM algorithm to examine individual predictions generated by the BiLSTMs, without any changes to the underlying model. Using this approach, a predicted outcome can be explained in terms of contributions of individual visits as well as contributions of combinations of visits. This method was evaluated for the task of predicting school-age asthma for pre-school children with asthma-like symptoms. This development beyond individual visit importance is key for understanding a model as complex and highly non-linear as BiLSTM. We also show a use case for utilizing our method to obtain the important diagnosis codes at the cohort-level, for the asthma prediction task, as well as an interactive visualization to provide interpretations at patient-level EHR using the learned importance scores.

Finally, to address the **data irregularity** challenge, we described a method for time-aware attention-based visit-level representation using self-attention mechanism combined with non-stationary kernel approximation (Chapter 5). Specifically, we extended the self-attention architecture by replacing the positional encoding with non-stationary

kernel approximation to model non-stationary temporal relationships between patient visits in EHR. This method was evaluated on the task of predicting the next diagnosis codes, given a patient's EHR history. The experimental evaluation confirms that in a setting that models temporal data distributions, methods such as self-attention combined with non-stationary kernel approximation can boost the predictive power compared to the best base models (e.g. positional encoding and stationary kernels). However, the benefits from such methods are reduced under non-stationarity characteristics of EHRs. These findings emphasize the importance of accounting for the irregular data distributions when modeling patient representations in EHRs and the need for methods which can appropriately model the irregular time gaps between consecutive visits. Therefore, we show that using non-stationary kernels can be an appropriate alternative to positional encoding in self-attention models to model the complex non-stationary relationships in EHRs.

7.2. Future Directions

This section presents several potential future directions that can be followed based on the work presented in this thesis.

7.2.1. Reinforcement Learning

Reinforcement learning (RL) is a class of machine learning which executes a sequence of actions to increase the probability of achieving a predefined goal [211]. The considerable recent developments of RL motivates its utilization in the medical domain, where RL showed promising results for a variety of healthcare applications, especially those in which diagnosing decisions or treatment strategies are usually characterized by

a sequential decision-making procedure [212]–[214]. EHR includes two main features which motivate the use of reinforcement learning for EHR-based prediction tasks. First, in EHRs, decisions are made sequentially along a timeline, actions are based on the examined state, effects appear at later time points than the actions that generate them (time delay), and there is some notion of desired state(s). Therefore, combining EHRs data and reinforcement learning can provide an attractive framework for the continuous recommendation of treatments and intervention plans for each different medical state of the patient. Second, extracting labeled data from EHR is not always an easy task.

For example, if we would like to predict the onset of preterm delivery for an expecting mother as early as possible, standard supervised learning algorithms will require the training dataset to include labels for both the preterm outcome and the optimal time point to make the prediction. However, we are often given a training dataset where only the former label is present. In fact, it isn't a simple task even for medical specialists to identify the optimal time to make the decision. Reinforcement learning offers a potential solution for this issue, where labeled training data is not required as in supervised learning schemes, and we only need to define for the model the reward for making the decisions correctly. Therefore, a potential future contribution may address combining reinforcement learning with our code-level attention-based model (e.g. PredictPTB) for providing early predictions of outcomes complimented with interpretations explaining the decisions made by the reinforcement agent.

7.2.2. BERT-like Models

Given the success of deep sequence models and attention/self-attention mechanisms in EHR-based predictive modeling, a promising line of research may address adapting

transformer architectures [91] for EHR-based predictive modeling, while taking into account EHR-specific challenges and provide improvements on the prediction accuracy of these models.

Transformer models have been fundamentally studied for text processing, particularly the Bidirectional Encoder Representations from Transformers (BERT) model [215]. Adapting the transformer architecture to structured EHR data is a natural idea, since EHR and natural language text share a common analogy of being represented as sequential modalities for tokens from a large vocabulary, which was considered in BioBERT [216] for biomedical knowledge and clinicalBERT [217] for clinical text. However, to the best of our knowledge, there are only three relevant studies which adapted the transformer architecture into EHRs which are BEHRT [218], G-BERT [219], and Med-BERT [220]. These efforts developed interpretable risk prediction models, which can predict various types of diseases and integrate a wide range of EHR modalities/concepts in its modular architecture. Future work may consider evaluating alternative architectures for pre-training and fine-tuning on EHR, such as ELMo[221] and generative pre-training (GPT) [222].

7.2.3. Kernel Approximation Using Data-dependent Algorithms

Kernel approximation using random features-based algorithms proved effective in a broad range of machine learning tasks. Random features-based algorithms can be broadly grouped into two classes, data-independent algorithms and data-dependent algorithms. In this thesis, we focused on data-dependent approaches, which uses the training data to guide the selection of points and weights in the random features for better approximation quality and/or generalization performance. Given its considerable

success in modeling the non-stationarity of EHR data in the empirical evaluation conducted in Chapter 5 and the rapid growth of the related literature, we believe that an important direct extension of our work is to explore more data-dependent approaches for effective modeling of the time trajectories in EHRs. Specifically, we recommend future work to explore the following data-dependent classes (1) leverage score sampling, which replaces the original distribution $p(\omega)$ by a carefully chosen distribution $q(\omega)$ constructed using leverage scores (LS) [223]–[226], (2) re-weighted random feature selection, which is based on the idea of re-weighting the random features by solving a constrained optimization problem [227], [228], and (3) kernel learning by random features, which aims to learn the spectral distribution of kernel from the data to achieve better similarity representation and prediction [229], [230].

7.2.4. Integrating Heterogeneous Data Sources

Besides structured data (e.g. diagnosis codes, procedure codes, and medication codes), unstructured data (e.g. images, clinical notes, lab measures, and spectrograms) in EHRs can provide additional, precious information for more accurate predictions. However, the processes to analyze such unstructured data are often complex, time-consuming, and require excessive manual effort. In this dissertation, we focused on capturing the relationships among structured data. A rather straightforward, but nonetheless promising future work is to leverage more data modalities (structured and unstructured) from EHRs using deep learning techniques for an improved performance in various prediction tasks.

REFERENCES

[1] J. V. Selby and B. H. Fireman, "Building predictive models for clinical care—where to build and what to predict?" *JAMA Network Open*, vol. 4, no. 1, e2032539, Jan. 2021. DOI: 10.1001/jamanetworkopen.2020.32539. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2020.32539.

[2] F. Wang, L. P. Casalino, and D. Khullar, "Deep learning in medicine—promise, progress, and challenges," *JAMA Internal Medicine*, vol. 179, no. 3, p. 293, Mar. 2019. DOI: 10.1001/jamainternmed.2018.7117. [Online]. Available: https://doi.org/10.1001/jamainternmed.2018.7117.

[3] A. J. Steele, S. C. Denaxas, A. D. Shah, H. Hemingway, and N. M. Luscombe, "Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease," *PLOS ONE*, vol. 13, no. 8, T. R. Singh, Ed., e0202344, Aug. 2018. DOI: 10.1371/journal.pone.0202344. [Online]. Available: https://doi.org/10.1371/journal.pone.0202344.

[4] E. Mahmoudi, N. Kamdar, N. Kim, G. Gonzales, K. Singh, and A. K. Waljee, "Use of electronic medical records in development and validation of risk prediction models of hospital readmission: Systematic review," *BMJ*, p. m958, Apr. 2020. DOI: 10.1136/bmj.m958. [Online]. Available: https://doi.org/10.1136/bmj.m958.

[5] S. Kashyap, K. M. Corey, A. Kansal, and M. Sendak, "Machine learning for predictive analytics," in *Machine Learning in Cardiovascular Medicine*, Elsevier,

2021, pp. 45–69. DOI: 10.1016/b978-0-12-820273-9.00003-8. [Online]. Available: https://doi.org/10.1016/b978-0-12-820273-9.00003-8.

[6] B. Norgeot, B. S. Glicksberg, L. Trupin, *et al.*, "Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis," *JAMA Network Open*, vol. 2, no. 3, e190606, Mar. 2019. DOI: 10.1001/jamanetworkopen.2019.0606. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2019.0606.

[7] G. Shamai, Y. Binenbaum, R. Slossberg, I. Duek, Z. Gil, and R. Kimmel, "Artificial intelligence algorithms to assess hormonal status from tissue microarrays in patients with breast cancer," *JAMA Network Open*, vol. 2, no. 7, e197700, Jul. 2019. DOI: 10.1001/jamanetworkopen.2019.7700. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2019.7700.

[8] A. Rajkomar, E. Oren, K. Chen, *et al.*, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 1, May 2018. DOI: 10.1038/s41746-018-0029-1. [Online]. Available: https://doi.org/10.1038/s41746-018-0029-1.

[9] J. Wu, J. Roy, and W. F. Stewart, "Prediction modeling using EHR data," *Medical Care*, vol. 48, no. 6, S106–S113, Sep. 2010. DOI: 10.1097/mlr.0b013e3181de9e17. [Online]. Available: https://doi.org/10.1097/mlr.0b013e3181de9e17.

[10] "A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: A retrospective analysis of electronic medical records data,"

BMC Medical Informatics and Decision Making, vol. 18, no. 1, pp. 1–17, 2018, ISSN: 14726947. DOI: 10.1186/s12911-018-0620-z.

[11] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk Prediction with Electronic Health Records: A Deep Learning Approach," pp. 432–440, 2016. DOI: 10.1137/1.9781611974348.49.

[12] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1419–1428, Jun. 2018. DOI: 10.1093/jamia/ocy068. [Online]. Available: https://doi.org/10.1093/jamia/ocy068.

[13] Z. C. Lipton, D. C. Kale, C. Elkan, and R. C. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1511.03677.

[14] S.-J. Bang, Y. Wang, and Y. Yang, "Phased-LSTM Based Predictive Model for longitudinal EHR Data with Missing Values," 2016. [Online]. Available: https://www.cs.cmu.edu/%7B~%7Depxing/Class/10708-17/project-reports/project8.pdf.

[15] C. Esteban, O. Staeck, S. Baier, Y. Yang, and V. Tresp, "Predicting clinical events by combining static and dynamic information using recurrent neural networks," in *2016 IEEE International Conference on Healthcare Informatics*

(ICHI), IEEE, Oct. 2016. DOI: 10.1109/ichi.2016.16. [Online]. Available: https://doi.org/10.1109/ichi.2016.16.

[16] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "DeepCare: A deep dynamic memory model for predictive medicine," in *Advances in Knowledge Discovery and Data Mining*, Springer International Publishing, 2016, pp. 30–41. DOI: 10.1007/978-3-319-31750-2_3. [Online]. Available: https://doi.org/10.1007/978-3-319-31750-2_3.

[17] A. N. Jagannatha and H. Yu, "Bidirectional RNN for Medical Event Detection in Electronic Health Records," pp. 473–482, 2016.

[18] J. Liu, Z. Zhang, and N. Razavian, "Deep ehr: Chronic disease prediction using medical notes," in *Proceedings of the 3rd Machine Learning for Healthcare Conference*, F. Doshi-Velez, J. Fackler, K. Jung, *et al.*, Eds., ser. Proceedings of Machine Learning Research, vol. 85, PMLR, 2018, pp. 440–464. [Online]. Available: https://proceedings.mlr.press/v85/liu18b.html.

[19] S. Wunnava, X. Qin, T. Kakar, C. Sen, E. A. Rundensteiner, and X. Kong, "Adverse Drug Event Detection from Electronic Health Records Using Hierarchical Recurrent Neural Networks with Dual-Level Embedding," *Drug Safety*, vol. 42, no. 1, pp. 113–122, Jan. 2019, ISSN: 0114-5916. DOI: 10.1007/s40264-018-0765-9. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30649736%20http://link.springer.com/10.1007/s40264-018-0765-9.

[20] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–

15, Feb. 2018. DOI: 10.1016/j.dsp.2017.10.011. [Online]. Available: https://doi.org/10.1016/j.dsp.2017.10.011.

[21] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for HealthCare," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Aug. 2015. DOI: 10.1145/2783258.2788613. [Online]. Available: https://doi.org/10.1145/2783258.2788613.

[22] T. Holman, S. Wallask, and K. Lee, *What is icd-10 (international classification of diseases, tenth revision)?* Sep. 2018. [Online]. Available: https://searchhealthit.techtarget.com/definition/ICD-10#:~:text=Another%5C%20difference%5C%20is%5C%20the%5C%20number,10%5C%2DPCS%5C%20has%5C%2087%5C%2C000%5C%20codes..

[23] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," ser. NIPS'16, Barcelona, Spain: Curran Associates Inc., 2016, pp. 3512–3520, ISBN: 9781510838819.

[24] R. AlSaad, Q. Malluhi, and S. Boughorbel, "PredictPTB: An interpretable preterm birth prediction model using attention-based recurrent neural networks," *BioData Mining*, vol. 15, no. 1, Feb. 2022. DOI: 10.1186/s13040-022-00289-8. [Online]. Available: https://doi.org/10.1186/s13040-022-00289-8.

[25] R. AlSaad, Q. Malluhi, I. Janahi, and S. Boughorbel, "Interpreting patient-specific risk prediction using contextual decomposition of BiLSTMs: Application to children with asthma," *BMC Medical Informatics and Decision Making*,

vol. 19, no. 1, Nov. 2019. DOI: 10.1186/s12911-019-0951-4. [Online]. Available: https://doi.org/10.1186/s12911-019-0951-4.

[26] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *Journal of Clinical Epidemiology*, vol. 49, no. 11, pp. 1225–1231, Nov. 1996. DOI: 10.1016/s0895-4356(96)00002-9. [Online]. Available: https://doi.org/10.1016/s0895-4356(96)00002-9.

[27] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–14, 2017, ISSN: 21682194. DOI: 10.1109/JBHI.2017.2767063. arXiv: 1706.03446.

[28] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," *Journal of Biomedical Informatics*, vol. 35, no. 5-6, pp. 352–359, Oct. 2002. DOI: 10.1016/s1532-0464(03)00034-0. [Online]. Available: https://doi.org/10.1016/s1532-0464(03)00034-0.

[29] D. E. Adkins, "Machine Learning and Electronic Health Records: A Paradigm Shift," *American Journal of Psychiatry*, vol. 174, no. 2, pp. 93–94, 2017. DOI: 10.1176/appi.ajp.2016.16101169.

[30] X. Liu, K. Gao, B. Liu, *et al.*, "Advances in deep learning-based medical image analysis," *Health Data Science*, vol. 2021, pp. 1–14, Jun. 2021. DOI: 10.34133/

2021/8786793. [Online]. Available: https://doi.org/10.34133/2021/8786793.

[31] R. AlSaad, S. Al-maadeed, M. A. A. Mamun, and S. Boughorbel, "A deep learning based automatic severity detector for diabetic retinopathy," in *Machine Learning and Data Mining in Pattern Recognition*, Springer International Publishing, 2018, pp. 64–76. DOI: 10.1007/978-3-319-96136-1_6. [Online]. Available: https://doi.org/10.1007/978-3-319-96136-1_6.

[32] R. AlSaad, S. Boughorbel, and S. Al-Maadeed, "Automated classification of diabetic retinopathy severity: A deep learning approach," in *Qatar Foundation Annual Research Conference Proceedings Volume 2018 Issue 2*, Hamad bin Khalifa University Press (HBKU Press), 2018. DOI: 10.5339/qfarc.2018.hbpd1007. [Online]. Available: https://doi.org/10.5339/qfarc.2018.hbpd1007.

[33] L. Dai, L. Wu, H. Li, *et al.*, "A deep learning system for detecting diabetic retinopathy across the disease spectrum," *Nature Communications*, vol. 12, no. 1, May 2021. DOI: 10.1038/s41467-021-23458-5. [Online]. Available: https://doi.org/10.1038/s41467-021-23458-5.

[34] J. Vamathevan, D. Clark, P. Czodrowski, *et al.*, "Applications of machine learning in drug discovery and development," *Nature Reviews Drug Discovery*, vol. 18, no. 6, pp. 463–477, Apr. 2019. DOI: 10.1038/s41573-019-0024-5. [Online]. Available: https://doi.org/10.1038/s41573-019-0024-5.

[35] T. Lucas, M. Ferreira, R. Plachta, G. Ferreira, and K. Costa, "Non-fragmented network flow design analysis: Comparison IPv4 with IPv6 using path MTU dis-

covery," *Computers*, vol. 9, no. 2, p. 54, Jun. 2020. DOI: 10.3390/computers9020054. [Online]. Available: https://doi.org/10.3390/computers9020054.

[36] B. Rim, N.-J. Sung, S. Min, and M. Hong, "Deep learning in physiological signal data: A survey," *Sensors*, vol. 20, no. 4, p. 969, Feb. 2020. DOI: 10.3390/s20040969. [Online]. Available: https://doi.org/10.3390/s20040969.

[37] Y. Zhang, F. Tiryaki, M. Jiang, and H. Xu, "Parsing clinical text using the state-of-the-art deep learning based parsers: A systematic comparison," *BMC Medical Informatics and Decision Making*, vol. 19, no. S3, Apr. 2019. DOI: 10.1186/s12911-019-0783-2. [Online]. Available: https://doi.org/10.1186/s12911-019-0783-2.

[38] S. Wu, K. Roberts, S. Datta, *et al.*, "Deep learning in clinical natural language processing: A methodical review," *Journal of the American Medical Informatics Association*, vol. 27, no. 3, pp. 457–470, Dec. 2019. DOI: 10.1093/jamia/ocz200. [Online]. Available: https://doi.org/10.1093/jamia/ocz200.

[39] D. Plant and A. Barton, "Machine learning in precision medicine: Lessons to learn," *Nature Reviews Rheumatology*, vol. 17, no. 1, pp. 5–6, Nov. 2020. DOI: 10.1038/s41584-020-00538-2. [Online]. Available: https://doi.org/10.1038/s41584-020-00538-2.

[40] B. C. Kwon, M. Choi, J. T. Kim, *et al.*, "Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records," *CoRR*, vol. abs/1805.10724, 2018. arXiv: 1805.10724. [Online]. Available: http://arxiv.org/abs/1805.10724.

[41] C. Giordano, M. Brennan, B. Mohamed, P. Rashidi, F. Modave, and P. Tighe, "Accessing artificial intelligence for clinical decision-making," *Frontiers in Digital Health*, vol. 3, Jun. 2021. DOI: 10.3389/fdgth.2021.645232. [Online]. Available: https://doi.org/10.3389/fdgth.2021.645232.

[42] Z. Che, S. Purushotham, R. Khemani, and Y. Liu, "Interpretable Deep Models for ICU Outcome Prediction.," *AMIA Annual Symposium proceedings. AMIA Symposium*, vol. 2016, pp. 371–380, 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28269832%7B%5C%%7D0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5333206.

[43] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, 2018. DOI: 10.18653/v1/n18-1100. [Online]. Available: https://doi.org/10.18653/v1/n18-1100.

[44] C. Meng, L. Trinh, N. Xu, and Y. Liu, *Mimic-if: Interpretability and fairness evaluation of deep learning models on mimic-iv dataset*, 2021. arXiv: 2102.06761 [cs.LG].

[45] K. HAYRINEN, K. SARANTO, and P. NYKANEN, "Definition, structure, content, use and impacts of electronic health records: A review of the research literature," *International Journal of Medical Informatics*, vol. 77, no. 5, pp. 291–304, May 2008. DOI: 10.1016/j.ijmedinf.2007.09.001. [Online]. Available: https://doi.org/10.1016/j.ijmedinf.2007.09.001.

[46] M. C. Data, *Secondary Analysis of Electronic Health Records*. Springer International Publishing, 2016. DOI: 10.1007/978-3-319-43742-2. [Online]. Available: https://doi.org/10.1007/978-3-319-43742-2.

[47] S. M. Shah and R. A. Khan, "Secondary use of electronic health record: Opportunities and challenges," *IEEE Access*, vol. 8, pp. 136 947–136 965, 2020. DOI: 10.1109/access.2020.3011099. [Online]. Available: https://doi.org/10.1109/access.2020.3011099.

[48] R. Alsaad, R. Al-Ali, and R. Badji, "Towards a national electronic health record in qatar: Building on international experiences," in *Qatar Foundation Annual Research Conference Proceedings Volume 2016 Issue 1*, Hamad bin Khalifa University Press (HBKU Press), 2016. DOI: 10.5339/qfarc.2016.hbop3013. [Online]. Available: https://doi.org/10.5339/qfarc.2016.hbop3013.

[49] T. Bergquist, V. Pejaver, N. Hammarlund, S. D. Mooney, and S. J. Mooney, "Evaluation of the secondary use of electronic health records to detect seasonal, holiday-related, and rare events related to traumatic injury and poisoning," *BMC Public Health*, vol. 20, no. 1, Jan. 2020. DOI: 10.1186/s12889-020-8153-7. [Online]. Available: https://doi.org/10.1186/s12889-020-8153-7.

[50] I. Danciu, J. D. Cowan, M. Basford, *et al.*, "Secondary use of clinical data: The vanderbilt approach," *Journal of Biomedical Informatics*, vol. 52, pp. 28–35, Dec. 2014. DOI: 10.1016/j.jbi.2014.02.003. [Online]. Available: https://doi.org/10.1016/j.jbi.2014.02.003.

[51] N. G. Weiskopf, G. Hripcsak, S. Swaminathan, and C. Weng, "Defining and measuring completeness of electronic health records for secondary use," *Journal*

of Biomedical Informatics, vol. 46, no. 5, pp. 830–836, Oct. 2013. DOI: 10.1016/j.jbi.2013.06.010. [Online]. Available: https://doi.org/10.1016/j.jbi.2013.06.010.

[52] B. J. Wells, A. S. Nowacki, K. Chagin, and M. W. Kattan, "Strategies for handling missing data in electronic health record derived data," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, vol. 1, no. 3, p. 7, Dec. 2013. DOI: 10.13063/2327-9214.1035. [Online]. Available: https://doi.org/10.13063/2327-9214.1035.

[53] J. L. Fernández-Alemán, I. C. Señor, P. Á. O. Lozoya, and A. Toval, "Security and privacy in electronic health records: A systematic literature review," *Journal of Biomedical Informatics*, vol. 46, no. 3, pp. 541–562, Jun. 2013. DOI: 10.1016/j.jbi.2012.12.003. [Online]. Available: https://doi.org/10.1016/j.jbi.2012.12.003.

[54] C. Pagliari, D. Detmer, and P. Singleton, "Potential of electronic personal health records," *BMJ*, vol. 335, no. 7615, pp. 330–333, Aug. 2007. DOI: 10.1136/bmj.39279.482963.ad. [Online]. Available: https://doi.org/10.1136/bmj.39279.482963.ad.

[55] I. Keshta and A. Odeh, "Security and privacy of electronic health records: Concerns and challenges," vol. 22, no. 2, pp. 177–183, Jul. 2021. DOI: 10.1016/j.eij.2020.07.003. [Online]. Available: https://doi.org/10.1016/j.eij.2020.07.003.

[56] B. K. Atchinson and D. M. Fox, "From the field: The politics of the health insurance portability and accountability act," vol. 16, no. 3, pp. 146–150, May

1997. DOI: 10.1377/hlthaff.16.3.146. [Online]. Available: https://doi.org/10.1377/hlthaff.16.3.146.

[57] R. B. Ness, "Influence of the HIPAA privacy rule on health research," vol. 298, no. 18, p. 2164, Nov. 2007. DOI: 10.1001/jama.298.18.2164. [Online]. Available: https://doi.org/10.1001/jama.298.18.2164.

[58] Y. Lu, R. O. Sinnott, K. Verspoor, and U. Parampalli, "Privacy-preserving access control in electronic health record linkage," IEEE, Aug. 2018. DOI: 10.1109/trustcom/bigdatase.2018.00151. [Online]. Available: https://doi.org/10.1109/trustcom/bigdatase.2018.00151.

[59] V. K. Saxena and S. Pushkar, "Risk reduction privacy preserving approach for accessing electronic health records," vol. 16, no. 3, pp. 46–57, Jul. 2021. DOI: 10.4018/ijhisi.20210701.oa3. [Online]. Available: https://doi.org/10.4018/ijhisi.20210701.oa3.

[60] A. N. Kho, J. P. Cashy, K. L. Jackson, *et al.*, "Design and implementation of a privacy preserving electronic health record linkage tool in chicago," vol. 22, no. 5, pp. 1072–1080, Jun. 2015. DOI: 10.1093/jamia/ocv038. [Online]. Available: https://doi.org/10.1093/jamia/ocv038.

[61] F. N. Wirth, T. Meurers, M. Johns, and F. Prasser, "Privacy-preserving data sharing infrastructures for medical research: Systematization and comparison," vol. 21, no. 1, Aug. 2021. DOI: 10.1186/s12911-021-01602-x. [Online]. Available: https://doi.org/10.1186/s12911-021-01602-x.

[62] K. P. Andriole, "Security of electronic medical information and patient privacy: What you need to know," *Journal of the American College of Radiology*, vol. 11,

no. 12, pp. 1212–1216, Dec. 2014. DOI: 10.1016/j.jacr.2014.09.011. [Online]. Available: https://doi.org/10.1016/j.jacr.2014.09.011.

[63] I. Keshta and A. Odeh, "Security and privacy of electronic health records: Concerns and challenges," *Egyptian Informatics Journal*, vol. 22, no. 2, pp. 177–183, Jul. 2021. DOI: 10.1016/j.eij.2020.07.003. [Online]. Available: https://doi.org/10.1016/j.eij.2020.07.003.

[64] G. I. Hausvik, D. Thapa, and B. E. Munkvold, "Information quality life cycle in secondary use of EHR data," *International Journal of Information Management*, vol. 56, p. 102 227, Feb. 2021. DOI: 10.1016/j.ijinfomgt.2020.102227. [Online]. Available: https://doi.org/10.1016/j.ijinfomgt.2020.102227.

[65] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis, "Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review," *Journal of the American Medical Informatics Association*, vol. 24, no. 1, pp. 198–208, May 2016. DOI: 10.1093/jamia/ocw042. [Online]. Available: https://doi.org/10.1093/jamia/ocw042.

[66] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012, ISBN: 0262018020.

[67] J. R. A. Solares, F. E. D. Raimondi, Y. Zhu, *et al.*, "Deep learning for electronic health records: A comparative review of multiple deep neural architectures," *Journal of Biomedical Informatics*, vol. 101, p. 103 337, Jan. 2020. DOI: 10.1016/j.jbi.2019.103337. [Online]. Available: https://doi.org/10.1016/j.jbi.2019.103337.

[68] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific Reports*, vol. 6, no. 1, May 2016. DOI: 10.1038/srep26094. [Online]. Available: https://doi.org/10.1038/srep26094.

[69] A. Jagannatha and H. Yu, *Structured prediction models for rnn based sequence labeling in clinical text*, 2016. arXiv: 1608.00612 [cs.CL].

[70] D. Williams, H. Hornung, A. Nadimpalli, and A. Peery, "Deep learning and its application for healthcare delivery in low and middle income countries," *Frontiers in Artificial Intelligence*, vol. 4, Apr. 2021. DOI: 10.3389/frai.2021.553987. [Online]. Available: https://doi.org/10.3389/frai.2021.553987.

[71] C. Weng, N. H. Shah, and G. Hripcsak, "Deep phenotyping: Embracing complexity and temporality—towards scalability, portability, and interoperability," *Journal of Biomedical Informatics*, vol. 105, p. 103 433, May 2020. DOI: 10.1016/j.jbi.2020.103433. [Online]. Available: https://doi.org/10.1016/j.jbi.2020.103433.

[72] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent neural networks for time series forecasting: Current status and future directions," *International Journal of Forecasting*, vol. 37, no. 1, pp. 388–427, Jan. 2021. DOI: 10.1016/j.ijforecast.2020.06.008. [Online]. Available: https://doi.org/10.1016/j.ijforecast.2020.06.008.

[73] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific Reports*,

vol. 8, no. 1, Apr. 2018. DOI: 10.1038/s41598-018-24271-9. [Online]. Available: https://doi.org/10.1038/s41598-018-24271-9.

[74] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware LSTM networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Aug. 2017. DOI: 10.1145/3097983.3097997. [Online]. Available: https://doi.org/10.1145/3097983.3097997.

[75] W. Lu, L. Ma, H. Chen, X. Jiang, and M. Gong, "A clinical prediction model in health time series data based on long short-term memory network optimized by fruit fly optimization algorithm," *IEEE Access*, vol. 8, pp. 136 014–136 023, 2020. DOI: 10.1109/access.2020.3011721. [Online]. Available: https://doi.org/10.1109/access.2020.3011721.

[76] F. Khoshnevisan, J. Ivy, M. Capan, R. Arnold, J. Huddleston, and M. Chi, "Recent temporal pattern mining for septic shock early prediction," in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, Jun. 2018. DOI: 10.1109/ichi.2018.00033. [Online]. Available: https://doi.org/10.1109/ichi.2018.00033.

[77] D. N. A. Ningrum, W.-M. Kung, I.-S. Tzeng, *et al.*, "A deep learning model to predict knee osteoarthritis based on nonimage longitudinal medical record," *Journal of Multidisciplinary Healthcare*, vol. Volume 14, pp. 2477–2485, Sep. 2021. DOI: 10.2147/jmdh.s325179. [Online]. Available: https://doi.org/10.2147/jmdh.s325179.

[78] R. Y. Coley, J. M. Boggs, A. Beck, and G. E. Simon, "Predicting outcomes of psychotherapy for depression with electronic health record data," *Journal of Affective Disorders Reports*, vol. 6, p. 100 198, Dec. 2021. DOI: 10.1016/j.jadr.2021.100198. [Online]. Available: https://doi.org/10.1016/j.jadr.2021.100198.

[79] R. Vuokko, P. Mäkelä-Bengs, H. Hyppönen, M. Lindqvist, and P. Doupi, "Impacts of structuring the electronic health record: Results of a systematic literature review from the perspective of secondary use of patient data," *International Journal of Medical Informatics*, vol. 97, pp. 293–303, Jan. 2017. DOI: 10.1016/j.ijmedinf.2016.10.004. [Online]. Available: https://doi.org/10.1016/j.ijmedinf.2016.10.004.

[80] A. A. Ismail, M. K. Gunady, H. C. Bravo, and S. Feizi, "Benchmarking deep learning interpretability in time series predictions," in *NeurIPS*, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/47a3893cc405396a5c30d91320572d6d-Abstract.html.

[81] I. Gandin, A. Scagnetto, S. Romani, and G. Barbati, "Interpretability of time-series deep learning models: A study in cardiovascular patients admitted to intensive care unit," *Journal of Biomedical Informatics*, vol. 121, p. 103 876, Sep. 2021. DOI: 10.1016/j.jbi.2021.103876. [Online]. Available: https://doi.org/10.1016/j.jbi.2021.103876.

[82] D. Baric, P. Fumic, D. Horvatic, and T. Lipic, "Benchmarking attention-based interpretability of deep learning in multivariate time series predictions," *Entropy*,

vol. 23, no. 2, p. 143, Jan. 2021. DOI: 10.3390/e23020143. [Online]. Available: https://doi.org/10.3390/e23020143.

[83] C. H. Yoon, R. Torrance, and N. Scheinerman, "Machine learning in medicine: Should the pursuit of enhanced interpretability be abandoned?" *Journal of Medical Ethics*, medethics–2020–107 102, May 2021. DOI: 10.1136/medethics-2020-107102. [Online]. Available: https://doi.org/10.1136/medethics-2020-107102.

[84] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Muller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, O. D. Suarez, Ed., e0130140, Jul. 2015. DOI: 10.1371/journal.pone.0130140. [Online]. Available: https://doi.org/10.1371/journal.pone.0130140.

[85] K. Simonyan, A. Vedaldi, and A. Zisserman, *Deep inside convolutional networks: Visualising image classification models and saliency maps*, 2014. arXiv: 1312.6034 [cs.CV].

[86] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, *Smoothgrad: Removing noise by adding noise*, 2017. arXiv: 1706.03825 [cs.LG].

[87] P.-J. Kindermans, K. Schütt, K.-R. Müller, and S. Dähne, *Investigating the influence of noise and distractors on the interpretation of neural networks*, 2016. arXiv: 1611.07270 [stat.ML].

[88] M. Sundararajan, A. Taly, and Q. Yan, *Axiomatic attribution for deep networks*, 2017. arXiv: 1703.01365 [cs.LG].

[89] M. T. Ribeiro, S. Singh, and C. Guestrin, *"why should i trust you?": Explaining the predictions of any classifier*, 2016. arXiv: 1602.04938 [cs.LG].

[90] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," ser. NIPS17, Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4768–4777, ISBN: 9781510860964.

[91] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2017. arXiv: 1706.03762 [cs.CL].

[92] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020. DOI: 10.3390/e23010018. [Online]. Available: https://doi.org/10.3390/e23010018.

[93] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Muller, and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," in *ICANN*, 2016.

[94] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML17, Sydney, NSW, Australia: JMLR.org, 2017, pp. 3145–3153.

[95] Y. Sha and M. D. Wang, "Interpretable predictions of clinical outcomes with an attention-based recurrent neural network," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, ACM, Aug. 2017. DOI: 10.1145/3107411.3107445. [Online]. Available: https://doi.org/10.1145/3107411.3107445.

[96] J. Luo, M. Ye, C. Xiao, and F. Ma, "HiTANet: Hierarchical time-aware attention networks for risk prediction on electronic health records," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, Jul. 2020. DOI: 10.1145/3394486.3403107. [Online]. Available: https://doi.org/10.1145/3394486.3403107.

[97] Y. Xiang, H. Ji, Y. Zhou, *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*, vol. 22, no. 7, e16981, Jul. 2020. DOI: 10.2196/16981. [Online]. Available: https://doi.org/10.2196/16981.

[98] P. Chen, W. Dong, J. Wang, X. Lu, U. Kaymak, and Z. Huang, "Interpretable clinical prediction via attention-based neural network," *BMC Medical Informatics and Decision Making*, vol. 20, no. S3, Jul. 2020. DOI: 10.1186/s12911-020-1110-7. [Online]. Available: https://doi.org/10.1186/s12911-020-1110-7.

[99] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.0473.

[100] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proceedings of the 27th International Conference on Neural*

Information Processing Systems - Volume 2, ser. NIPS'14, Montreal, Canada: MIT Press, 2014, pp. 2204–2212.

[101] F. Wang, M. Jiang, C. Qian, *et al.*, "Residual attention network for image classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017. DOI: 10.1109/cvpr.2017.683. [Online]. Available: https://doi.org/10.1109/cvpr.2017.683.

[102] J. B. Lee, R. Rossi, and X. Kong, "Graph classification using structural attention," ACM, Jul. 2018. DOI: 10.1145/3219819.3219980. [Online]. Available: https://doi.org/10.1145/3219819.3219980.

[103] J. Cheng and M. Lapata, "Neural summarization by extracting sentences and words," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2016. DOI: 10.18653/v1/p16-1046. [Online]. Available: https://doi.org/10.18653/v1/p16-1046.

[104] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2015. DOI: 10.18653/v1/d15-1044. [Online]. Available: https://doi.org/10.18653/v1/d15-1044.

[105] H. Choi, K. Cho, and Y. Bengio, "Fine-grained attention mechanism for neural machine translation," *Neurocomputing*, vol. 284, pp. 171–176, Apr. 2018. DOI: 10.1016/j.neucom.2018.01.007. [Online]. Available: https://doi.org/10.1016/j.neucom.2018.01.007.

[106] A. F. T. Martins and R. F. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML16, New York, NY, USA: JMLR.org, 2016, pp. 1614–1623.

[107] H. Chen, M. Sun, C. Tu, Y. Lin, and Z. Liu, "Neural sentiment classification with user and product attention," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1650–1659. DOI: 10.18653/v1/D16-1171. [Online]. Available: https://aclanthology.org/D16-1171.

[108] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "GRAM: Graph-based Attention Model for Healthcare Representation Learning," pp. 1–15, 2016, ISSN: 03092402. DOI: 10.1145/3097983.3098126. arXiv: 1611.07012. [Online]. Available: http://arxiv.org/abs/1611.07012.

[109] W. Guo, W. Ge, L. Cui, H. Li, and L. Kong, "An interpretable disease onset predictive model using crossover attention mechanism from electronic health records," *IEEE Access*, vol. 7, pp. 134 236–134 244, 2019. DOI: 10.1109/access.2019.2928579. [Online]. Available: https://doi.org/10.1109/access.2019.2928579.

[110] Z. Lin, M. Feng, C. N. dos Santos, *et al.*, *A structured self-attentive sentence embedding*, 2017. arXiv: 1703.03130 [cs.CL].

[111] A. E. Johnson, T. J. Pollard, L. Shen, *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, May 2016. DOI: 10.1038/

sdata.2016.35. [Online]. Available: https://doi.org/10.1038/sdata.2016.35.

[112] N. Hou, M. Li, L. He, *et al.*, "Predicting 30-days mortality for MIMIC-III patients with sepsis-3: A machine learning approach using XGboost," *Journal of Translational Medicine*, vol. 18, no. 1, Dec. 2020. DOI: 10.1186/s12967-020-02620-5. [Online]. Available: https://doi.org/10.1186/s12967-020-02620-5.

[113] J. F. Rodrigues-Jr, G. Spadon, B. Brandoli, and S. Amer-Yahia, *Patient trajectory prediction in the mimic-iii dataset, challenges and pitfalls*, 2019. arXiv: 1909.04605 [cs.LG].

[114] J. Deasy, P. Liò, and A. Ercole, "Dynamic survival prediction in intensive care units from heterogeneous time series without the need for variable selection or curation," *Scientific Reports*, vol. 10, no. 1, Dec. 2020. DOI: 10.1038/s41598-020-79142-z. [Online]. Available: https://doi.org/10.1038/s41598-020-79142-z.

[115] N. Ding, C. Guo, C. Li, Y. Zhou, and X. Chai, "An artificial neural networks model for early predicting in-hospital mortality in acute pancreatitis in MIMIC-III," *BioMed Research International*, vol. 2021, D. Monlezun, Ed., pp. 1–8, Jan. 2021. DOI: 10.1155/2021/6638919. [Online]. Available: https://doi.org/10.1155/2021/6638919.

[116] V. L. Martucci, N. Liu, V. E. Kerchberger, *et al.*, "A clinical phenotyping algorithm to identify cases of chronic obstructive pulmonary disease in electronic

health records," Jul. 2019. DOI: 10.1101/716779. [Online]. Available: https://doi.org/10.1101/716779.

[117] X. Zhong, Z. Yin, G. Jia, *et al.*, "Electronic health record phenotypes associated with genetically regulated expression of CFTR and application to cystic fibrosis," *Genetics in Medicine*, vol. 22, no. 7, pp. 1191–1200, Apr. 2020. DOI: 10.1038/s41436-020-0786-5. [Online]. Available: https://doi.org/10.1038/s41436-020-0786-5.

[118] Cerner, *Real-world data solution*. [Online]. Available: https://www.cerner.com/solutions/real-world-data.

[119] A. Tatham, "The increasing importance of clinical coding," vol. 69, no. 7, pp. 372–373, Jul. 2008. DOI: 10.12968/hmed.2008.69.7.30409. [Online]. Available: https://doi.org/10.12968/hmed.2008.69.7.30409.

[120] N. Delvaux, B. Vaes, B. Aertgeerts, *et al.*, "Coding systems for clinical decision support: Theoretical and real-world comparative analysis," vol. 4, no. 10, e16094, Oct. 2020. DOI: 10.2196/16094. [Online]. Available: https://doi.org/10.2196/16094.

[121] J. C. Ferrao, M. D. Oliveira, F. Janela, H. M. G. Martins, and D. Gartner, "Can structured EHR data support clinical coding? a data mining approach," vol. 10, no. 2, pp. 138–161, Mar. 2020. DOI: 10.1080/20476965.2020.1729666. [Online]. Available: https://doi.org/10.1080/20476965.2020.1729666.

[122] V. Alonso, J. V. Santos, M. Pinto, *et al.*, "Health records as the basis of clinical coding: Is the quality adequate? a qualitative study of medical coders'

perceptions," *Health Information Management Journal*, vol. 49, no. 1, pp. 28–37, 2020, PMID: 30744403. DOI: 10.1177/1833358319826351. [Online]. Available: https://doi.org/10.1177/1833358319826351.

[123] S. Saigal and L. W. Doyle, "An overview of mortality and sequelae of preterm birth from infancy to adulthood," *The Lancet*, vol. 371, no. 9608, pp. 261–269, Jan. 2008. DOI: 10.1016/s0140-6736(08)60136-1. [Online]. Available: https://doi.org/10.1016/s0140-6736(08)60136-1.

[124] J. M. Tielsch, "Global incidence of preterm birth," in *Nestlé Nutrition Institute Workshop Series*, S. Karger AG, 2015, pp. 9–15. DOI: 10.1159/000365798. [Online]. Available: https://doi.org/10.1159/000365798.

[125] F. C. Barros, A. T. Papageorghiou, C. G. Victora, *et al.*, "The distribution of clinical phenotypes of preterm birth syndrome," *JAMA Pediatrics*, vol. 169, no. 3, p. 220, Mar. 2015. DOI: 10.1001/jamapediatrics.2014.3040. [Online]. Available: https://doi.org/10.1001/jamapediatrics.2014.3040.

[126] N. Marlow, D. Wolke, M. A. Bracewell, and M. Samara, "Neurologic and developmental disability at six years of age after extremely preterm birth," *New England Journal of Medicine*, vol. 352, no. 1, pp. 9–19, Jan. 2005. DOI: 10.1056/nejmoa041367. [Online]. Available: https://doi.org/10.1056/nejmoa041367.

[127] H. G. Taylor, N. Klein, N. M. Minich, and M. Hack, "Middle-school-age outcomes in children with very low birthweight," *Child Development*, vol. 71,

no. 6, pp. 1495–1511, Nov. 2000. DOI: 10.1111/1467-8624.00242. [Online]. Available: https://doi.org/10.1111/1467-8624.00242.

[128] R. W. I. Cooke, "Health, lifestyle, and quality of life for young adults born very preterm," *Archives of Disease in Childhood*, vol. 89, no. 3, pp. 201–206, Mar. 2004. DOI: 10.1136/adc.2003.030197. [Online]. Available: https://doi.org/10.1136/adc.2003.030197.

[129] J. Henderson, C. Carson, and M. Redshaw, "Impact of preterm birth on maternal well-being and women's perceptions of their baby: A population-based survey," *BMJ Open*, vol. 6, no. 10, e012676, Oct. 2016. DOI: 10.1136/bmjopen-2016-012676. [Online]. Available: https://doi.org/10.1136/bmjopen-2016-012676.

[130] V. Pierrat, L. Marchand-Martin, C. Arnaud, *et al.*, "Neurodevelopmental outcome at 2 years for preterm children born at 22 to 34 weeks' gestation in france in 2011: EPIPAGE-2 cohort study," *BMJ*, j3448, Aug. 2017. DOI: 10.1136/bmj.j3448. [Online]. Available: https://doi.org/10.1136/bmj.j3448.

[131] T. Cobo, M. Kacerovsky, and B. Jacobsson, "Risk factors for spontaneous preterm delivery," *International Journal of Gynecology & Obstetrics*, vol. 150, no. 1, pp. 17–23, Jun. 2020. DOI: 10.1002/ijgo.13184. [Online]. Available: https://doi.org/10.1002/ijgo.13184.

[132] H. Ren and M. Du, "Role of maternal periodontitis in preterm birth," *Frontiers in Immunology*, vol. 8, Feb. 2017. DOI: 10.3389/fimmu.2017.00139. [Online]. Available: https://doi.org/10.3389/fimmu.2017.00139.

[133] Z. A. O. Kaplan and A. S. Ozgu-Erdinc, "Prediction of preterm birth: Maternal characteristics, ultrasound markers, and biomarkers: An updated overview," *Journal of Pregnancy*, vol. 2018, pp. 1–8, Oct. 2018. DOI: 10.1155/2018/8367571. [Online]. Available: https://doi.org/10.1155/2018/8367571.

[134] H. Blencowe, S. Cousens, M. Z. Oestergaard, *et al.*, "National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: A systematic analysis and implications," *The Lancet*, vol. 379, no. 9832, pp. 2162–2172, Jun. 2012. DOI: 10.1016/s0140-6736(12)60820-4. [Online]. Available: https://doi.org/10.1016/s0140-6736(12)60820-4.

[135] R. J. Baer, M. R. McLemore, N. Adler, *et al.*, "Pre-pregnancy or first-trimester risk scoring to identify women at high risk of preterm birth," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 231, pp. 235–240, Dec. 2018. DOI: 10.1016/j.ejogrb.2018.11.004. [Online]. Available: https://doi.org/10.1016/j.ejogrb.2018.11.004.

[136] L. K. Goodwin, M. A. Iannacchione, W. E. Hammond, P. Crockett, S. Maher, and K. Schlitz, "Data mining methods find demographic predictors of preterm birth," *Nursing Research*, vol. 50, no. 6, pp. 340–345, Nov. 2001. DOI: 10.1097/00006199-200111000-00003. [Online]. Available: https://doi.org/10.1097/00006199-200111000-00003.

[137] D. E. Jesse, W. Seaver, and D. C. Wallace, "Maternal psychosocial risks predict preterm birth in a group of women from appalachia," *Midwifery*, vol. 19, no. 3,

pp. 191–202, Sep. 2003. DOI: 10.1016/s0266-6138(03)00031-7. [Online]. Available: https://doi.org/10.1016/s0266-6138(03)00031-7.

[138] L. K. Woolery and J. Grzymala-Busse, "Machine learning for an expert system to predict preterm birth risk," *Journal of the American Medical Informatics Association*, vol. 1, no. 6, pp. 439–446, Nov. 1994. DOI: 10.1136/jamia.1994.95153433. [Online]. Available: https://doi.org/10.1136/jamia.1994.95153433.

[139] A. Weber, G. L. Darmstadt, S. Gruber, *et al.*, "Application of machine-learning to predict early spontaneous preterm birth among nulliparous non-hispanic black and white women," *Annals of Epidemiology*, vol. 28, no. 11, 783–789.e1, Nov. 2018. DOI: 10.1016/j.annepidem.2018.08.008. [Online]. Available: https://doi.org/10.1016/j.annepidem.2018.08.008.

[140] H. Rawashdeh, S. Awawdeh, F. Shannag, *et al.*, "Intelligent system based on data mining techniques for prediction of preterm birth for women with cervical cerclage," *Computational Biology and Chemistry*, vol. 85, p. 107 233, Apr. 2020. DOI: 10.1016/j.compbiolchem.2020.107233. [Online]. Available: https://doi.org/10.1016/j.compbiolchem.2020.107233.

[141] E. Nodelman, J. Molitoris, and M. Holbert, "543: Using artificial intelligence to predict spontaneous preterm delivery," *American Journal of Obstetrics and Gynecology*, vol. 222, no. 1, S350, Jan. 2020. DOI: 10.1016/j.ajog.2019.11.559. [Online]. Available: https://doi.org/10.1016/j.ajog.2019.11.559.

[142] T. A. H. Rocha, E. B. A. F. de Thomaz, D. G. de Almeida, *et al.*, "Data-driven risk stratification for preterm birth in brazil: A population-based study to develop of a machine learning risk assessment approach," *The Lancet Regional Health - Americas*, p. 100 053, Aug. 2021. DOI: 10.1016/j.lana.2021.100053. [Online]. Available: https://doi.org/10.1016/j.lana.2021.100053.

[143] Z. Safi, N. Venugopal, H. Ali, M. Makhlouf, and S. Boughorbel, "Analysis of risk factors progression of preterm delivery using electronic health records," Sep. 2020. DOI: 10.21203/rs.3.rs-78033/v1. [Online]. Available: https://doi.org/10.21203/rs.3.rs-78033/v1.

[144] T. Wlodarczyk, S. Plotka, P. Rokita, *et al.*, "Spontaneous preterm birth prediction using convolutional neural networks," in *ASMUS/PIPPI@MICCAI*, 2020.

[145] T. Włodarczyk, S. Płotka, T. Trzciński, *et al.*, *Estimation of preterm birth markers with u-net segmentation network*, 2019. arXiv: 1908.09148 [eess.IV].

[146] C. Gao, S. Osmundson, D. R. V. Edwards, G. P. Jackson, B. A. Malin, and Y. Chen, "Deep learning predicts extreme preterm birth from electronic health records," *Journal of Biomedical Informatics*, vol. 100, p. 103 334, Dec. 2019. DOI: 10.1016/j.jbi.2019.103334. [Online]. Available: https://doi.org/10.1016/j.jbi.2019.103334.

[147] A. Abraham, B. Le, I. Kosti, *et al.*, "Dense phenotyping from electronic health records enables machine-learning-based prediction of preterm birth," Jul. 2020. DOI: 10.1101/2020.07.15.20154864. [Online]. Available: https://doi.org/10.1101/2020.07.15.20154864.

[148] H. He and Y. Ma, *Imbalanced learning : foundations, algorithms, and applications*. Hoboken, New Jersey: John Wiley & Sons, Inc, 2013, ISBN: 1118074629.

[149] A. Jayaram, C. H. Collier, and J. N. Martin, "Preterm parturition and pre-eclampsia: The confluence of two great gestational syndromes," *International Journal of Gynecology & Obstetrics*, vol. 150, no. 1, pp. 10–16, Jun. 2020. DOI: 10.1002/ijgo.13173. [Online]. Available: https://doi.org/10.1002/ijgo.13173.

[150] C. Carreno, B. Kase, L. Hart, S. Blackwell, B. Sibai, and B. Connealy, "A history of prior preeclampsia as a risk factor for preterm birth," *American Journal of Perinatology*, vol. 31, no. 06, pp. 483–488, Aug. 2013. DOI: 10.1055/s-0033-1353439. [Online]. Available: https://doi.org/10.1055/s-0033-1353439.

[151] L. Visser, C. Slaager, B. Kazemier, *et al.*, "Risk of preterm birth after prior term cesarean," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 127, no. 5, pp. 610–617, Feb. 2020. DOI: 10.1111/1471-0528.16083. [Online]. Available: https://doi.org/10.1111/1471-0528.16083.

[152] B. S. de Vries, J. P. Ludlow, and A. Cong, "Term cesarean delivery in the first pregnancy and increased risk for preterm delivery in the subsequent pregnancy," *American Journal of Obstetrics and Gynecology*, vol. 222, no. 6, pp. 635–636, Jun. 2020. DOI: 10.1016/j.ajog.2020.01.047. [Online]. Available: https://doi.org/10.1016/j.ajog.2020.01.047.

[153] L. Lecointre, N. Sananes, A. S. Weingertner, *et al.*, "Fetoscopic laser coagulation for twin-twin transfusion syndrome before 17 weeks' gestation: Laser data,

complications and neonatal outcome," *Ultrasound in Obstetrics & Gynecology*, vol. 44, no. 3, pp. 299–303, Aug. 2014. DOI: 10.1002/uog.13375. [Online]. Available: https://doi.org/10.1002/uog.13375.

[154] A. Malshe, S. Snowise, L. K. Mann, *et al.*, "Preterm delivery after fetoscopic laser surgery for twin-twin transfusion syndrome: Etiology and risk factors," *Ultrasound in Obstetrics & Gynecology*, vol. 49, no. 5, pp. 612–616, May 2017. DOI: 10.1002/uog.15972. [Online]. Available: https://doi.org/10.1002/uog.15972.

[155] J. Harris and E. Sheiner, "Does an upper respiratory tract infection during pregnancy affect perinatal outcomes? a literature review," *Current Infectious Disease Reports*, vol. 15, no. 2, pp. 143–147, Jan. 2013. DOI: 10.1007/s11908-013-0320-x. [Online]. Available: https://doi.org/10.1007/s11908-013-0320-x.

[156] M. Silasi, I. Cardenas, J.-Y. Kwon, K. Racicot, P. Aldo, and G. Mor, "Viral infections during pregnancy," *American Journal of Reproductive Immunology*, vol. 73, no. 3, pp. 199–213, Jan. 2015. DOI: 10.1111/aji.12355. [Online]. Available: https://doi.org/10.1111/aji.12355.

[157] M. S. Rose, G. Pana, and S. Premji, "Prenatal maternal anxiety as a risk factor for preterm birth and the effects of heterogeneity on this relationship: A systematic review and meta-analysis," *BioMed Research International*, vol. 2016, pp. 1–18, 2016. DOI: 10.1155/2016/8312158. [Online]. Available: https://doi.org/10.1155/2016/8312158.

[158] C. Lilliecreutz, J. Larén, G. Sydsjö, and A. Josefsson, "Effect of maternal stress during pregnancy on the risk for preterm birth," *BMC Pregnancy and Childbirth*, vol. 16, no. 1, Jan. 2016. DOI: 10.1186/s12884-015-0775-x. [Online]. Available: https://doi.org/10.1186/s12884-015-0775-x.

[159] M. Shimaoka, Y. Yo, K. Doh, *et al.*, "Association between preterm delivery and bacterial vaginosis with or without treatment," *Scientific Reports*, vol. 9, no. 1, Jan. 2019. DOI: 10.1038/s41598-018-36964-2. [Online]. Available: https://doi.org/10.1038/s41598-018-36964-2.

[160] K. J. Lee, J. Yoo, Y.-H. Kim, *et al.*, "The clinical usefulness of predictive models for preterm birth with potential benefits: A KOrean preterm collaboratE network (KOPEN) registry-linked data-based cohort study," *International Journal of Medical Sciences*, vol. 17, no. 1, pp. 1–12, 2020. DOI: 10.7150/ijms.37626. [Online]. Available: https://doi.org/10.7150/ijms.37626.

[161] C. Espinosa, M. Becker, I. Marić, *et al.*, "Data-driven modeling of pregnancy-related complications," *Trends in Molecular Medicine*, vol. 27, no. 8, pp. 762–776, Aug. 2021. DOI: 10.1016/j.molmed.2021.01.007. [Online]. Available: https://doi.org/10.1016/j.molmed.2021.01.007.

[162] R. A. Belaghi, J. Beyene, and S. D. McDonald, "Clinical risk models for preterm birth less than 28 weeks and less than 32 weeks of gestation using a large retrospective cohort," *Journal of Perinatology*, Jun. 2021. DOI: 10.1038/s41372-021-01109-3. [Online]. Available: https://doi.org/10.1038/s41372-021-01109-3.

[163] J. M. Schaaf, A. C. Ravelli, B. W. J. Mol, and A. Abu-Hanna, "Development of a prognostic model for predicting spontaneous singleton preterm birth," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 164, no. 2, pp. 150–155, Oct. 2012. DOI: 10.1016/j.ejogrb.2012.07.007. [Online]. Available: https://doi.org/10.1016/j.ejogrb.2012.07.007.

[164] N.-H. Morken, K. Källen, and B. Jacobsson, "Predicting risk of spontaneous preterm delivery in women with a singleton pregnancy," *Paediatric and Perinatal Epidemiology*, vol. 28, no. 1, pp. 11–22, Oct. 2013. DOI: 10.1111/ppe.12087. [Online]. Available: https://doi.org/10.1111/ppe.12087.

[165] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, ACM, Aug. 2018. DOI: 10.1145/3233547.3233667. [Online]. Available: https://doi.org/10.1145/3233547.3233667.

[166] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: definitions, methods, and applications," pp. 1–11, 2019. arXiv: 1901.04592. [Online]. Available: http://arxiv.org/abs/1901.04592.

[167] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, and N. Elhadad, "Multi-Label Classification of Patient Notes a Case Study on ICD Code Assignment," 2017. arXiv: 1709.09587. [Online]. Available: http://arxiv.org/abs/1709.09587.

[168] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks," 2017, ISSN: 0029-7828. DOI: 10.1145/3097983.3098088. arXiv: 1706.05764. [Online]. Available: http://arxiv.org/abs/1706.05764%7B%5C%%7D0Ahttp://dx.doi.org/10.1145/3097983.3098088.

[169] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, and L. E. Barnes, "Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record," *IEEE Access*, vol. 6, pp. 65 333–65 346, 2018, ISSN: 21693536. DOI: 10.1109/ACCESS.2018.2875677. arXiv: 1810.04793.

[170] Y. Xu, S. Biswal, S. R. Deshpande, K. O. Maher, and J. Sun, "RAIM: Recurrent Attentive and Intensive Model of Multimodal Patient Monitoring Data," 2018. DOI: 10.1145/3219819.3220051. arXiv: 1807.08820. [Online]. Available: http://arxiv.org/abs/1807.08820.

[171] Z. Che, S. Purushotham, R. Khemani, and Y. Liu, "Distilling Knowledge from Deep Networks with Applications to Healthcare Domain," Dec. 2015. arXiv: 1512.03542. [Online]. Available: http://arxiv.org/abs/1512.03542.

[172] W. J. Murdoch, P. J. Liu, and B. Yu, "Beyond word importance: Contextual decomposition to extract interactions from lstms," *arXiv preprint arXiv:1801.05453*, 2018.

[173] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, Dec. 1997. DOI: 10.1162/neco.1997.9.8.1735.

[174] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997. DOI:

10.1109/78.650093. [Online]. Available: https://doi.org/10.1109/78.650093.

[175] J. P. DeShazo and M. A. Hoffman, "A comparison of a multistate inpatient EHR database to the HCUP nationwide inpatient sample," *BMC Health Services Research*, vol. 15, no. 1, Jun. 2015. DOI: 10.1186/s12913-015-1025-7. [Online]. Available: https://doi.org/10.1186/s12913-015-1025-7.

[176] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. arXiv: 1412.6980 [cs.LG].

[177] M. Szumilas, "Explaining odds ratios.," *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, vol. 19, no. 3, pp. 227–9, Aug. 2010, ISSN: 1719-8429. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20842279%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2938757.

[178] M. Morais-Almeida, A. Gaspar, G. Pires, S. Prates, and J. Rosado-Pinto, "Risk factors for asthma symptoms at school age: an 8-year prospective study.," *Allergy and asthma proceedings*, vol. 28, no. 2, pp. 183–9, 2007, ISSN: 1088-5412. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17479602.

[179] A. Bjerg and E. Rönmark, "Asthma in school age: prevalence and risk factors by time and by age," *The Clinical Respiratory Journal*, vol. 2, pp. 123–126, Oct. 2008, ISSN: 17526981. DOI: 10.1111/j.1752-699X.2008.00095.x. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20298361%20http://doi.wiley.com/10.1111/j.1752-699X.2008.00095.x.

[180] S. S. Szentpetery, O. Gruzieva, E. Forno, *et al.*, "Combined effects of multiple risk factors on asthma in school-aged children.," *Respiratory medicine*, vol. 133, pp. 16–21, Dec. 2017, ISSN: 1532-3064. DOI: 10.1016/j.rmed.2017.11.002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29173444%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5728683.

[181] C. Hur, J. Wi, and Y. Kim, "Facilitating the development of deep learning models with visual analytics for electronic health records," *International Journal of Environmental Research and Public Health*, vol. 17, no. 22, p. 8303, Nov. 2020. DOI: 10.3390/ijerph17228303. [Online]. Available: https://doi.org/10.3390/ijerph17228303.

[182] S. S. Abdullah, N. Rostamzadeh, K. Sedig, A. X. Garg, and E. McArthur, "Visual analytics for dimension reduction and cluster analysis of high dimensional electronic health records," *Informatics*, vol. 7, no. 2, p. 17, May 2020. DOI: 10.3390/informatics7020017. [Online]. Available: https://doi.org/10.3390/informatics7020017.

[183] J. Carriere, H. Shafi, K. Brehon, *et al.*, "Case report: Utilizing AI and NLP to assist with healthcare and rehabilitation during the COVID-19 pandemic," *Frontiers in Artificial Intelligence*, vol. 4, Feb. 2021. DOI: 10.3389/frai.2021.613637. [Online]. Available: https://doi.org/10.3389/frai.2021.613637.

[184] M. Nauta, M. van Putten, M. C. Tjepkema-Cloostermans, J. Bos, M. van Keulen, and C. Seifert, "Interactive explanations of internal representations of neural net-

work layers: An exploratory study on outcome prediction of comatose patients," in *Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence, KDH@ECAI 2020, Santiago de Compostela, Spain & Virtually, August 29-30, 2020*, K. Bach, R. C. Bunescu, C. Marling, and N. Wiratunga, Eds., ser. CEUR Workshop Proceedings, vol. 2675, CEUR-WS.org, 2020, pp. 5–11. [Online]. Available: http://ceur-ws.org/Vol-2675/paper1.pdf.

[185] Y. Juhn, A. Kung, R. Voigt, and S. Johnson, "Characterisation of children's asthma status by ICD-9 code and criteria-based medical record review," *Primary Care Respiratory Journal*, vol. 20, no. 1, pp. 79–83, Nov. 2010. DOI: 10.4104/pcrj.2010.00076. [Online]. Available: https://doi.org/10.4104/pcrj.2010.00076.

[186] J. Gao, C. Xiao, Y. Wang, W. Tang, L. M. Glass, and J. Sun, "StageNet: Stage-aware neural networks for health risk prediction," in *Proceedings of The Web Conference 2020*, ACM, Apr. 2020. DOI: 10.1145/3366423.3380136. [Online]. Available: https://doi.org/10.1145/3366423.3380136.

[187] X. Peng, G. Long, T. Shen, S. Wang, J. Jiang, and M. Blumenstein, "Temporal self-attention network for medical concept embedding," *2019 IEEE International Conference on Data Mining (ICDM)*, Nov. 2019. DOI: 10.1109/icdm.2019.00060. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2019.00060.

[188] S. A. Kamal, C. Yin, B. Qian, and P. Zhang, "An interpretable risk prediction model for healthcare with pattern attention," *BMC Medical Informatics and*

Decision Making, vol. 20, no. S11, Dec. 2020. DOI: 10.1186/s12911-020-01331-7. [Online]. Available: https://doi.org/10.1186/s12911-020-01331-7.

[189] X. Peng, G. Long, T. Shen, S. Wang, and J. Jiang, *Self-attention enhanced patient journey understanding in healthcare system*, 2020. arXiv: 2006.10516 [cs.LG].

[190] P. Kar and H. Karnick, "Random feature maps for dot product kernels," in *AISTATS*, 2012.

[191] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *NIPS*, 2007.

[192] J. Chen, D. Cheng, and Y. Liu, *On bochner's and polya's characterizations of positive-definite kernels and the respective random feature maps*, 2016. arXiv: 1610.08861 [stat.ML].

[193] D. Xu, C. Ruan, S. Kumar, E. Korpeoglu, and K. Achan, *Self-attention with functional time representation learning*, 2019. arXiv: 1911.12864 [cs.LG].

[194] Z. Song, D. Woodruff, Z. Yu, and L. Zhang, "Fast sketching of polynomial kernels of polynomial degree," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, Jul. 2021, pp. 9812–9823. [Online]. Available: https://proceedings.mlr.press/v139/song21c.html.

[195] J. Pennington, F. X. Yu, and S. Kumar, "Spherical random features for polynomial kernels," in *Proceedings of the 28th International Conference on Neural*

Information Processing Systems - Volume 2, ser. NIPS'15, Montreal, Canada: MIT Press, 2015, pp. 1846–1854.

[196] H. Avron, H. L. Nguyen, and D. P. Woodruff, "Subspace embeddings for the polynomial kernel," in *NIPS*, 2014.

[197] D. Patel, W. Hsu, and M. L. Lee, "Mining relationships among interval-based events for classification," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data - SIGMOD '08*, ACM Press, 2008. DOI: 10.1145/1376616.1376658. [Online]. Available: https://doi.org/10.1145/1376616.1376658.

[198] R. Moskovitch and Y. Shahar, "Classification of multivariate time series via temporal abstraction and time intervals mining," *Knowledge and Information Systems*, vol. 45, no. 1, pp. 35–74, Oct. 2014. DOI: 10.1007/s10115-014-0784-5. [Online]. Available: https://doi.org/10.1007/s10115-014-0784-5.

[199] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal phenotyping from longitudinal electronic health records: A graph based framework," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Aug. 2015. DOI: 10.1145/2783258.2783352. [Online]. Available: https://doi.org/10.1145/2783258.2783352.

[200] J. Zhao and A. Henriksson, "Learning temporal weights of clinical events using variable importance," *BMC Medical Informatics and Decision Making*, vol. 16, no. S2, Jul. 2016. DOI: 10.1186/s12911-016-0311-6. [Online]. Available: https://doi.org/10.1186/s12911-016-0311-6.

[201] Y. Shahar, "A framework for knowledge-based temporal abstraction," *Artificial Intelligence*, vol. 90, no. 1-2, pp. 79–133, Feb. 1997. DOI: 10.1016/s0004-3702(96)00025-2. [Online]. Available: https://doi.org/10.1016/s0004-3702(96)00025-2.

[202] R. Moskovitch, C. Walsh, F. Wang, G. Hripcsak, and N. Tatonetti, "Outcomes prediction via time intervals related patterns," in *2015 IEEE International Conference on Data Mining*, IEEE, Nov. 2015. DOI: 10.1109/icdm.2015.143. [Online]. Available: https://doi.org/10.1109/icdm.2015.143.

[203] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," vol. 116, no. 44, pp. 22 071–22 080, Oct. 2019. DOI: 10.1073/pnas.1900654116. [Online]. Available: https://doi.org/10.1073/pnas.1900654116.

[204] F. Fan, J. Xiong, M. Li, and G. Wang, *On interpretability of artificial neural networks: A survey*, 2021. arXiv: 2001.02522 [cs.LG].

[205] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, *Explaining explanations: An overview of interpretability of machine learning*, 2019. arXiv: 1806.00069 [cs.AI].

[206] S. Serrano and N. A. Smith, "Is attention interpretable?" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2931–2951. DOI: 10.18653/v1/P19-1282. [Online]. Available: https://aclanthology.org/P19-1282.

[207] D. G. Altman and P. Royston, "What do we mean by validating a prognostic model?" *Statistics in Medicine*, vol. 19, no. 4, pp. 453–473, Feb. 2000. DOI: 10.1002/(sici)1097-0258(20000229)19:4<453::aid-sim350>3.0.co;2-5. [Online]. Available: https://doi.org/10.1002/(sici)1097-0258(20000229)19:4%3C453::aid-sim350%3E3.0.co;2-5.

[208] A. C. Justice, "Assessing the generalizability of prognostic information," *Annals of Internal Medicine*, vol. 130, no. 6, p. 515, Mar. 1999. DOI: 10.7326/0003-4819-130-6-199903160-00016. [Online]. Available: https://doi.org/10.7326/0003-4819-130-6-199903160-00016.

[209] J. ZHONG, C. BLAUM, J. YU, *et al.*, "1470-p: EHR-based vs. population-based CVD risk predictions for diabetes patients," *Diabetes*, vol. 68, no. Supplement 1, 1470–P, Jun. 2019. DOI: 10.2337/db19-1470-p. [Online]. Available: https://doi.org/10.2337/db19-1470-p.

[210] B. A. Goldstein, A. M. Navar, and M. J. Pencina, "Risk prediction with electronic health records," *JAMA Cardiology*, vol. 1, no. 9, p. 976, Dec. 2016. DOI: 10.1001/jamacardio.2016.3826. [Online]. Available: https://doi.org/10.1001/jamacardio.2016.3826.

[211] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, "Human-level control through deep reinforcement learning," vol. 518, no. 7540, pp. 529–533, Feb. 2015. DOI: 10.1038/nature14236. [Online]. Available: https://doi.org/10.1038/nature14236.

[212] M. Tejedor, A. Z. Woldaregay, and F. Godtliebsen, "Reinforcement learning application in diabetes blood glucose control: A systematic review," *Artifi-*

cial Intelligence in Medicine, vol. 104, p. 101 836, 2020, ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2020.101836. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365718304548.

[213] A. Coronato, M. Naeem, G. D. Pietro, and G. Paragliola, "Reinforcement learning for intelligent healthcare applications: A survey," vol. 109, p. 101 964, Sep. 2020. DOI: 10.1016/j.artmed.2020.101964. [Online]. Available: https://doi.org/10.1016/j.artmed.2020.101964.

[214] S. He, L. G. Leanse, and Y. Feng, "Artificial intelligence and machine learning assisted drug delivery for effective treatment of infectious diseases," vol. 178, p. 113 922, Nov. 2021. DOI: 10.1016/j.addr.2021.113922. [Online]. Available: https://doi.org/10.1016/j.addr.2021.113922.

[215] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423. [Online]. Available: https://doi.org/10.18653/v1/n19-1423.

[216] J. Lee, W. Yoon, S. Kim, *et al.*, "Biobert: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, J. Wren, Ed.,

Sep. 2019, ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btz682. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btz682.

[217] E. Alsentzer, J. Murphy, W. Boag, *et al.*, "Publicly available clinical," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Association for Computational Linguistics, 2019. DOI: 10.18653/v1/w19-1909. [Online]. Available: https://doi.org/10.18653/v1/w19-1909.

[218] Y. Li, S. Rao, J. R. A. Solares, *et al.*, "BEHRT: Transformer for electronic health records," vol. 10, no. 1, Apr. 2020. DOI: 10.1038/s41598-020-62922-y. [Online]. Available: https://doi.org/10.1038/s41598-020-62922-y.

[219] J. Shang, T. Ma, C. Xiao, and J. Sun, "Pre-training of graph augmented transformers for medication recommendation," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, Aug. 2019. DOI: 10.24963/ijcai.2019/825. [Online]. Available: https://doi.org/10.24963/ijcai.2019/825.

[220] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," vol. 4, no. 1, May 2021. DOI: 10.1038/s41746-021-00455-y. [Online]. Available: https://doi.org/10.1038/s41746-021-00455-y.

[221] M. Peters, M. Neumann, M. Iyyer, *et al.*, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, 2018. DOI: 10.18653/v1/n18-1202. [Online]. Available: https://doi.org/10.18653/v1/n18-1202.

[222] T. B. Brown, B. Mann, N. Ryder, *et al.*, *Language models are few-shot learners*, 2020. arXiv: 2005.14165 [cs.CL].

[223] H. Avron, M. Kapralov, C. Musco, C. Musco, A. Velingker, and A. Zandieh, *Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees*, 2018. arXiv: 1804.09893 [cs.LG].

[224] F. Bach, "On the equivalence between kernel quadrature rules and random feature expansions," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 714–751, 2017, ISSN: 1532-4435.

[225] Z. Li, J.-F. Ton, D. Oglic, and D. Sejdinovic, "Towards a unified analysis of random fourier features," *J. Mach. Learn. Res.*, vol. 22, 108:1–108:51, 2019.

[226] F. Liu, X. Huang, Y. Chen, J. Yang, and J. Suykens, "Random fourier features via fast surrogate leverage weighted sampling," vol. 34, no. 04, pp. 4844–4851, Apr. 2020. DOI: 10.1609/aaai.v34i04.5920. [Online]. Available: https://doi.org/10.1609/aaai.v34i04.5920.

[227] H. Avron, V. Sindhwani, J. Yang, and M. W. Mahoney, "Quasi-monte carlo feature maps for shift-invariant kernels," *Journal of Machine Learning Research*, vol. 17, no. 120, pp. 1–38, 2016. [Online]. Available: http://jmlr.org/papers/v17/14-538.html.

[228] T. Dao, C. D. Sa, and C. Ré, "Gaussian quadrature for kernel features," in *Proceedings of the 31st International Conference on Neural Information Processing*

Systems, ser. NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6109–6119, ISBN: 9781510860964.

[229] C.-L. Li, W.-C. Chang, Y. Mroueh, Y. Yang, and B. Póczos, "Implicit kernel learning," *ArXiv*, vol. abs/1902.10214, 2019.

[230] B. Bullins, C. Zhang, and Y. Zhang, *Not-so-random features*, 2018. arXiv: 1710.10230 [cs.LG].

APPENDIX A: ETHICAL APPROVAL

Data was obtained under a data use agreement between Sidra Medicine and Cerner Corporation. Data access was approved through the institutional review board at Sidra Medicine (protocol number 1804023464). Cerner Health Facts data is de-identified and is HIPAA-compliant to protect the identity/privacy of patients as well as organizations. Informed consent was exempted because of the retrospective nature of this research. Rawan AlSaad was affiliated with Sidra Medicine during the time of this work. A copy of the IRB protocol letter is included in Figure A.1.

سدرة للطب
Sidra Medicine

SIDRA MEDICINE

Tel: +974-4003-7747
Email: irb@sidra.org

Sidra IRB MOPH Registration: MOPH -Sidra-IRB-099
Sidra IRB DHHS Registration: IRB00009930
Sidra IRB MOPH Assurance: MOPH-A-Sidra-00100
Sidra IRB DHHS Assurance: FWA00022378

To: Ibrahim Janahi
Date: 9th May 2018
Committee Action: Exemption Granted
IRB Protocol #: 1804023464
Study Title: DATA-DRIVEN PREDICTIVE MODELS FOR ASTHMA DEVELOPMENT IN CHILDREN

The above-referenced protocol is considered exempt after review by the Institutional Review Board Office pursuant to Policies, Regulations and Guidelines for Research Involving Human Subjects as amended, for the exemption from IRB review. The protocol is granted exemption under the below category:

Category 2: Research involving the use of education test (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, observation of public behavior, unless the information is obtained and recorded in such a manner that human subjects can be identified, directly or through identifiers linked to subjects; and any disclosure of the human subjects' responses outside the research could reasonable place the subject at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

If you wish to make any changes to the study that might affect whether it qualifies for this exemption, you must notify the IRB Office and obtain approval before proceeding. You should retain a copy of this letter for your records.

Sincerely yours,

Chiara Cugno

Chiara Cugno
Acting Chair, Sidra IRB

Sidra Medical & Research Center
IRB
Exemption Granted
Date: 10/5/18

PO Box 26999
Doha - Qatar
Sidra.org

Figure A.1. Institutional Review Board (IRB) Letter

APPENDIX B: DETAILED RESULTS

Table B.1. Detailed Results for the Evaluation of PredictPTB Model and Baselines

	Experiment details					Mean					Standard deviation				
	Model	**Modality**	**Experiment**	**Window**	**Prediction Point (days)**	**ACC**	**AUC**	**PRUC**	**Sensitivity**	**Specificity**	**ACC**	**AUC**	**PRUC**	**Sensitivity**	**Specificity**
1	MLP	all	retro	short-term	30	92.42	79.26	34.27	70.77	70.59	0.1	0.61	1.55	0.5	1.36
2	MLP	all	retro	long-term	30	92.03	74.35	27.46	66.91	67.02	0.11	0.75	1.01	0.36	1.16
3	MLP	all	full	long-term	30	92	73.91	26.63	66.82	66.58	0.12	0.52	0.8	0.24	1.02
4	MLP	diagnosis	full	long-term	30	91.88	72.95	24.84	65.88	65.89	0.07	0.37	0.67	0.67	0.98
5	MLP	diagnosis	retro	long-term	30	91.96	72.78	24.85	65.87	65.27	0.11	0.47	0.61	0.38	1.23
6	MLP	all	retro	long-term	90	91.92	70.42	21.05	64.72	63.77	0.17	0.61	0.99	0.57	0.79
7	MLP	all	full	long-term	90	91.92	70.29	20.66	64.32	63.96	0.09	0.62	0.89	0.38	1.1
8	MLP	all	retro	long-term	180	92.04	69.42	18.77	63.9	63.72	0.15	0.75	0.62	0.88	1.48
9	MLP	all	full	long-term	180	91.99	69.34	18.44	63.68	64.21	0.09	0.34	0.38	0.79	1
10	MLP	diagnosis	full	long-term	90	91.93	69.34	19.78	63.51	63.24	0.06	0.75	0.41	1.06	1.06
11	MLP	diagnosis	retro	long-term	90	91.95	69.47	20.35	63.41	63.4	0.16	0.51	0.8	0.52	0.76
12	MLP	diagnosis	retro	long-term	180	92.1	67.88	16.96	62.23	62.26	0.09	0.76	0.54	0.45	1.08
13	MLP	diagnosis	full	long-term	180	91.95	67.79	16.84	62.18	62.26	0.09	0.63	0.61	0.44	1.16
14	MLP	lab orders	retro	long-term	180	92.13	64.97	14.87	60.92	60.4	0.23	0.45	0.68	0.86	0.85
15	MLP	lab orders	retro	long-term	30	91.54	64.28	16.24	60.63	59.78	0.19	0.52	0.62	0.64	1.22
16	MLP	lab orders	full	long-term	180	92.08	64.84	14.66	60.44	61.09	0.12	0.52	0.44	0.77	1.44
17	MLP	lab orders	full	long-term	30	91.7	64.19	16.16	60.37	59.77	0.18	0.55	0.71	0.5	1.12
18	MLP	lab orders	retro	long-term	90	91.89	64.11	14.9	60.2	60.19	0.17	0.52	0.41	0.38	1.07
19	MLP	lab orders	full	long-term	90	91.86	63.88	14.87	59.86	59.42	0.13	0.44	0.43	0.95	0.91
20	MLP	all	full	long-term	270	91.49	64.36	16.63	59.8	59.45	0.11	0.66	0.4	0.59	0.9
21	MLP	all	retro	long-term	270	91.49	64.73	17.15	59.76	59.91	0.16	0.82	0.62	1.02	1.62
22	MLP	diagnosis	retro	long-term	270	91.41	62.87	15.42	59.01	58.1	0.18	0.94	0.63	0.87	1.44
23	MLP	diagnosis	full	long-term	270	91.41	62.95	15.27	58.83	58.18	0.08	1.03	0.6	0.81	1.68
24	MLP	medications	full	long-term	30	91.12	63.57	18.11	58.76	58.91	0.12	0.47	0.1	0.42	0.92
25	MLP	medications	retro	long-term	30	91.11	63.86	18.16	58.74	59.37	0.14	0.43	0.6	0.84	1.05
26	MLP	medications	retro	long-term	270	91.43	62.27	14.95	58.59	57.71	0.15	0.57	0.8	0.3	1.22
27	MLP	procedures	retro	long-term	30	91.08	62.29	16.22	58.5	58.04	0.14	0.38	0.42	0.69	0.58
28	MLP	procedures	full	long-term	30	90.95	61.94	16.17	58.2	57.71	0.13	0.66	0.52	0.92	1.42
29	MLP	medications	retro	long-term	180	91.51	62.19	14.88	58.16	58.34	0.16	0.92	0.6	0.76	1.24
30	MLP	medications	full	long-term	270	91.34	62.5	15.21	58.11	58.39	0.26	0.74	0.68	0.83	1.55
31	MLP	lab orders	full	long-term	270	91.41	61.15	14.79	58.03	58.08	0.17	0.68	0.53	0.83	1
32	MLP	lab orders	retro	long-term	270	91.43	61.93	14.63	58.02	59.05	0.24	0.66	0.52	0.8	1.01
33	MLP	medications	full	long-term	180	91.51	62.06	15.26	57.74	58.58	0.17	0.57	0.61	1.12	1.74
34	MLP	medications	retro	long-term	90	91.38	61.88	15.18	57.56	57.98	0.09	0.74	0.53	0.47	1.07
35	MLP	procedures	retro	long-term	90	91.28	61.07	14.16	57.44	57.43	0.14	0.58	0.48	0.59	1.12
36	MLP	procedures	full	long-term	180	91.6	60.2	12.64	57.32	56.64	0.13	0.75	0.39	0.75	1.48
37	MLP	procedures	full	long-term	90	91.18	60.16	13.77	57.28	56.62	0.22	0.74	0.59	0.64	1.03
38	MLP	procedures	retro	long-term	180	91.61	60.77	12.79	57.25	57.12	0.08	0.67	0.6	0.95	2.06
39	MLP	medications	full	long-term	90	91.35	61.87	15.32	57.24	58.17	0.12	0.43	0.34	0.67	1.14
40	MLP	procedures	full	long-term	270	91.43	59.26	12.24	56.65	56.11	0.11	0.65	0.42	0.76	1.3
41	MLP	procedures	retro	long-term	270	91.43	59.72	12.54	56.61	57.24	0.2	0.84	0.31	0.53	1.2
42	MLP	surgeries	full	long-term	30	87.93	52.25	15.78	39.18	65.91	2.4	5.92	3.62	13.95	16.58
43	MLP	surgeries	retro	long-term	270	87.88	51.14	17.4	38.13	62.12	2.9	6.35	4.85	20.58	21.87
44	MLP	surgeries	full	long-term	180	89.6	49.47	13.24	34.54	60.23	1.64	9.04	4.13	21.13	29.3
45	MLP	surgeries	full	long-term	90	89.13	46.29	13.67	31.06	61.86	2.85	6.13	6.03	20.48	25.53
46	MLP	surgeries	retro	long-term	90	89.64	50.19	13.66	25.34	74.08	3.18	10.26	6.62	18.15	23.73
47	MLP	surgeries	retro	long-term	180	89.43	51.62	14.87	24.83	72.23	2.36	6.82	3.4	20.23	18.31
48	MLP	surgeries	retro	long-term	30	87.86	48.39	15.03	23.74	72.41	2.02	5.32	3.32	15.83	14.62
49	MLP	surgeries	full	long-term	270	85.34	51.82	17.77	19.55	80.35	1.84	4.7	3.56	12.85	15.34
50	MLP	all	full	short-term	30	92.41	79.02	33.47	70.52	70.52	0.18	0.37	1.01	0.63	0.94

	Experiment details					Mean					Standard deviation				
	Model	**Modality**	**Experiment**	**Window**	**Prediction Point (days)**	**ACC**	**AUC**	**PRUC**	**Sensitivity**	**Specificity**	**ACC**	**AUC**	**PRUC**	**Sensitivity**	**Specificity**
51	MLP	all	full	short-term	180	94.3	76.9	21.83	69.81	68.37	0.16	0.8	1.07	1.21	1.35
52	MLP	all	retro	short-term	90	92.64	76.22	26.78	69.08	68.37	0.1	0.46	1.11	0.77	1.1
53	MLP	all	retro	short-term	180	94.26	76.72	23.24	68.76	69.61	0.06	0.85	0.98	1.14	1.87
54	MLP	diagnosis	retro	short-term	30	92.15	76.75	30.44	68.7	68.5	0.16	0.42	0.9	0.46	0.93
55	MLP	all	full	short-term	90	92.63	75.65	25.83	68.66	68.09	0.13	0.25	0.65	0.22	0.72
56	MLP	diagnosis	retro	short-term	180	94.17	76.63	22.06	68.62	69.89	0.16	0.62	1.27	1.08	1.68
57	MLP	diagnosis	full	short-term	180	94.37	75.49	20.64	68.62	68.04	0.13	0.55	0.58	0.41	0.85
58	MLP	diagnosis	full	short-term	30	92.22	76.64	29.93	68.52	68.48	0.13	0.42	0.62	0.49	0.96
59	MLP	diagnosis	full	short-term	90	92.5	74.23	24.6	67.23	67.04	0.13	0.3	0.98	0.45	1.24
60	MLP	diagnosis	retro	short-term	90	92.52	74.37	24.96	66.9	67.61	0.15	0.69	0.55	0.55	1.37
61	MLP	lab orders	retro	short-term	180	94.55	68.63	14.54	64.61	62.57	0.16	1	1.24	0.95	1.38
62	MLP	procedures	retro	short-term	180	92.76	69.97	22.07	64.37	62.4	0.43	2.41	1.41	3.61	5.68
63	MLP	lab orders	full	short-term	180	94.67	69.48	13.97	63.7	64.71	0.13	1.5	0.7	1.4	2.5
64	MLP	procedures	full	short-term	180	92.51	70.29	20.29	63.51	63.37	0.4	1.97	2.39	2.38	4.18
65	MLP	procedures	full	short-term	30	88.32	67.96	28.49	63.48	61.27	0.45	0.88	0.81	0.66	2.25
66	MLP	procedures	retro	short-term	90	89.22	67.83	27.04	62.26	61.73	0.21	1.15	1.34	1.88	2.47
67	MLP	procedures	retro	short-term	30	88.41	69.12	29.3	62.25	63.8	0.27	0.97	1.46	1.13	1.6
68	MLP	lab orders	retro	short-term	90	92.59	67.41	15.76	62.13	61.46	0.16	0.55	0.74	0.99	1.74
69	MLP	lab orders	full	short-term	90	92.6	67.36	15.41	61.96	61.79	0.13	0.56	0.61	0.74	0.87
70	MLP	procedures	full	short-term	90	89	67.01	24.61	61.62	61.42	0.32	1.36	2.38	1.43	2.84
71	MLP	lab orders	full	short-term	30	91.81	66.24	17.96	61.19	61.68	0.18	0.69	0.5	0.54	1.22
72	MLP	lab orders	retro	short-term	30	91.74	66.33	18.05	60.92	62	0.12	0.4	0.63	0.72	0.68
73	MLP	medications	retro	short-term	30	89.59	64.62	24.9	59.11	60.11	0.2	0.94	1.07	1.13	1.21
74	MLP	medications	full	short-term	30	89.48	64.76	24.73	58.83	60.55	0.25	0.53	0.88	1.36	1.39
75	MLP	medications	full	short-term	180	91.55	64.2	20.52	57.8	61.46	0.55	1.08	2.26	2.87	3.09
76	MLP	medications	retro	short-term	180	91.38	63.45	22.13	56.68	60.52	0.44	1.74	2.91	2.77	2.25
77	MLP	medications	retro	short-term	90	90.1	60.99	17.12	55.78	58.57	0.28	1.54	0.95	1.82	3.39
78	MLP	medications	full	short-term	90	89.91	60.59	17.73	55.34	57.74	0.3	1.29	0.8	1.9	1.49
79	PredictPTB	all	retro	long-term	30	91.71	78.91	34.85	70.44	70.22	0.14	0.49	0.98	0.58	0.77
80	PredictPTB	all	full	long-term	30	91.74	78.77	34.74	70.2	70.03	0.15	0.49	1.05	0.68	0.76
81	PredictPTB	diagnosis	retro	long-term	30	91.81	76.76	31.84	68.92	68.64	0.24	0.51	0.61	0.39	0.96
82	PredictPTB	diagnosis	full	long-term	30	91.7	76.47	31.59	68.16	68.35	0.24	0.41	0.73	0.52	0.44
83	PredictPTB	all	full	long-term	90	91.54	74.43	26.11	67.32	66.83	0.15	0.59	0.56	0.29	0.78
84	PredictPTB	all	retro	long-term	90	91.54	75	27.13	66.96	67.89	0.12	0.39	0.73	0.53	1.15
85	PredictPTB	all	retro	long-term	180	91.79	73	21.55	66.48	66.01	0.16	0.7	0.46	0.63	1.04
86	PredictPTB	diagnosis	full	long-term	90	91.51	72.52	24.83	66	65.67	0.12	0.56	0.52	0.42	1.24
87	PredictPTB	diagnosis	retro	long-term	90	91.49	72.2	24.76	65.85	66.16	0.25	0.32	0.22	0.37	1.16
88	PredictPTB	all	full	long-term	180	91.68	72.61	20.99	65.78	66.38	0.19	0.48	0.5	0.57	1.61
89	PredictPTB	diagnosis	retro	long-term	180	91.89	70.85	19.6	65.37	64.45	0.15	0.38	0.43	0.49	1.19
90	PredictPTB	diagnosis	full	long-term	180	91.8	70.73	19.28	64.85	64.79	0.14	0.49	0.44	0.52	0.81
91	PredictPTB	lab orders	retro	long-term	180	92.08	67.77	16.78	62.79	62.48	0.11	0.72	0.86	0.66	0.72
92	PredictPTB	lab orders	full	long-term	180	92.11	67.76	16.1	62.53	63.05	0.16	0.51	0.45	0.69	1.2
93	PredictPTB	lab orders	full	long-term	30	91.44	67.19	18.36	62.46	61.27	0.17	0.41	0.49	0.45	0.58
94	PredictPTB	lab orders	retro	long-term	30	91.49	67.14	18.39	62.15	62.05	0.18	0.27	0.49	0.79	0.77
95	PredictPTB	lab orders	full	long-term	90	91.72	67.17	16.53	61.86	62.22	0.03	0.37	0.52	0.8	0.91
96	PredictPTB	lab orders	retro	long-term	90	91.84	67.05	16.38	61.73	61.83	0.15	0.5	0.5	0.42	1.33
97	PredictPTB	all	full	long-term	270	91.3	65.23	17.28	60.47	60.06	0.15	0.66	0.72	0.23	1
98	PredictPTB	all	retro	long-term	270	91.32	65.11	17.31	60.2	60.11	0.16	0.4	0.5	0.62	0.71
99	PredictPTB	medications	retro	long-term	30	90.78	64.94	19.96	59.62	59.2	0.14	0.56	0.85	0.44	1.14
100	PredictPTB	medications	full	long-term	30	90.91	64.97	20.08	59.27	59.66	0.17	0.58	0.69	0.47	0.65

	Experiment details					Mean					Standard deviation				
	Model	**Modality**	**Experiment**	**Window**	**Prediction Point (days)**	**ACC**	**AUC**	**PRUC**	**Sensitivity**	**Specificity**	**ACC**	**AUC**	**PRUC**	**Sensitivity**	**Specificity**
101	PredictPTB	lab orders	retro	long-term	270	91.33	63.24	15.45	59.23	59.34	0.24	0.48	0.54	0.55	0.61
102	PredictPTB	medications	full	long-term	180	91.4	63.4	16.1	59.08	58.78	0.16	0.52	0.75	0.71	1.12
103	PredictPTB	diagnosis	full	long-term	270	91.51	62.5	15.58	59.05	59.15	0.2	0.42	0.42	0.4	0.71
104	PredictPTB	diagnosis	retro	long-term	270	91.4	63.14	16.06	59	59.12	0.11	0.42	0.36	0.47	1.04
105	PredictPTB	medications	retro	long-term	180	91.55	63.75	16.08	58.98	59.13	0.13	0.59	0.58	0.56	1.3
106	PredictPTB	medications	retro	long-term	270	91.29	63.75	16.4	58.86	59.18	0.12	0.19	0.4	0.94	0.81
107	PredictPTB	lab orders	full	long-term	270	91.36	62.88	14.85	58.84	58.97	0.13	0.27	0.52	0.53	0.87
108	PredictPTB	medications	full	long-term	270	91.33	63.41	15.92	58.74	58.37	0.16	0.38	0.67	0.8	0.95
109	PredictPTB	medications	retro	long-term	90	91.4	63.14	15.83	58.66	58.72	0.18	0.41	0.54	0.4	0.77
110	PredictPTB	procedures	retro	long-term	30	90.73	62.82	18.11	58.43	58.12	0.25	0.67	0.63	1.02	0.67
111	PredictPTB	medications	full	long-term	90	91.32	62.76	15.75	58.38	57.67	0.09	0.64	0.49	0.8	1.09
112	PredictPTB	procedures	full	long-term	90	91.19	61.25	15.05	58.08	57.2	0.11	0.96	0.45	0.47	1.16
113	PredictPTB	procedures	full	long-term	30	91	62.76	17.2	57.88	58.61	0.14	0.66	0.3	0.56	0.96
114	PredictPTB	procedures	retro	long-term	90	91.21	61.1	15.42	57.87	56.79	0.17	0.68	0.8	1.02	1.13
115	PredictPTB	procedures	retro	long-term	180	91.57	60.56	13	57.11	57.03	0.17	0.73	0.47	0.99	1.56
116	PredictPTB	procedures	full	long-term	270	91.55	59.71	12.44	56.92	56.1	0.24	0.49	0.42	0.69	1.26
117	PredictPTB	procedures	full	long-term	180	91.68	60.56	12.85	56.65	57.23	0.11	0.94	0.49	0.42	1.58
118	PredictPTB	procedures	retro	long-term	270	91.52	59.49	12.32	56.53	56.83	0.14	0.67	0.43	0.67	0.79
119	PredictPTB	surgeries	retro	long-term	270	88.73	51.03	14.29	52.18	49.2	3.26	5.21	5.34	14.56	25.73
120	PredictPTB	surgeries	retro	long-term	180	89.76	49	11.73	41.4	66.76	3.18	8.07	3.97	18.33	14.83
121	PredictPTB	surgeries	full	long-term	30	87.93	49.81	15.76	41.32	61.71	1.76	6.42	5.24	16.27	18.14
122	PredictPTB	surgeries	retro	long-term	30	88.94	50.8	13.56	38.43	55.9	2.58	8.64	3.22	17.48	18.65
123	PredictPTB	surgeries	full	long-term	180	87.5	52.9	16.36	37.26	55.66	3.32	5.11	7.58	14.9	17.94
124	PredictPTB	surgeries	full	long-term	90	86.85	50.83	17.04	36.16	58.11	2.66	9.38	4.06	14.18	24.1
125	PredictPTB	surgeries	full	long-term	270	88.64	56.6	15.89	32.82	64.74	1.57	8.16	3.72	12.57	20.12
126	PredictPTB	surgeries	retro	long-term	90	85.9	53.73	19.09	31.6	68.86	2.26	9.27	5.91	18.25	11.07
127	PredictPTB	all	retro	short-term	30	92.08	82.49	41.09	73.28	73.32	0.13	0.39	0.84	0.27	1.24
128	PredictPTB	all	full	short-term	30	91.98	82.16	40.41	73.26	73.15	0.12	0.32	1.25	0.38	0.66
129	PredictPTB	diagnosis	retro	short-term	30	91.88	79.99	37.14	71.72	71.48	0.05	0.24	1.01	0.72	0.7
130	PredictPTB	diagnosis	full	short-term	30	91.91	80.09	36.69	71.48	70.89	0.14	0.12	0.76	0.24	0.58
131	PredictPTB	all	retro	short-term	180	93.86	78.54	24.57	71.45	71.12	0.26	0.57	1.13	1.14	2.33
132	PredictPTB	all	retro	short-term	90	92.34	78.93	32.27	71.06	70.54	0.15	0.47	0.83	0.44	0.82
133	PredictPTB	all	full	short-term	180	93.8	77.88	23.81	70.99	69.87	0.37	0.86	1.84	0.7	1.3
134	PredictPTB	all	full	short-term	90	92.22	78.81	31.14	70.05	71.08	0.26	0.34	0.7	0.87	1
135	PredictPTB	diagnosis	full	short-term	180	94.07	76.14	21.03	69.52	69.37	0.25	1.09	2.24	1.01	1.98
136	PredictPTB	diagnosis	retro	short-term	180	93.94	76.09	22.67	69.25	70.1	0.14	1.04	1.72	0.78	2.04
137	PredictPTB	diagnosis	retro	short-term	90	92.09	76.59	30.04	69.04	69.11	0.21	0.68	1.24	0.86	1.16
138	PredictPTB	diagnosis	full	short-term	90	92.26	76.14	29.26	69	68.36	0.14	0.44	0.76	0.74	1.26
139	PredictPTB	procedures	retro	short-term	180	92.46	69.54	22.08	65.88	65.81	0.27	4.51	5.09	1.84	3.63
140	PredictPTB	procedures	full	short-term	180	92.26	68.02	17.99	65.66	64.09	0.44	2.39	2.69	3.36	4.7
141	PredictPTB	lab orders	retro	short-term	180	94.49	70.36	15.99	64.67	66.34	0.2	0.93	0.58	0.93	1.41
142	PredictPTB	lab orders	full	short-term	180	94.26	67.98	14.86	64.1	63.41	0.04	0.73	2.07	0.61	0.67
143	PredictPTB	lab orders	full	short-term	90	92.35	68.98	16.23	63.79	62.79	0.18	0.4	0.57	0.51	1.27
144	PredictPTB	lab orders	retro	short-term	90	92.42	70.03	16.72	63.67	64.51	0.19	0.54	0.91	0.79	0.96
145	PredictPTB	lab orders	full	short-term	30	91.54	68.42	19.42	63.29	62.72	0.15	0.72	0.79	0.23	0.96
146	PredictPTB	procedures	full	short-term	30	86.9	69.7	31.5	63.22	63.78	0.56	1.35	1.74	1.21	2.91
147	PredictPTB	lab orders	retro	short-term	30	91.5	68.41	19.78	63.08	62.72	0.15	0.6	0.62	0.62	1.28
148	PredictPTB	procedures	retro	short-term	30	87.23	68.84	31.49	62.95	63.93	0.48	1.5	1.48	0.68	2.75
149	PredictPTB	procedures	retro	short-term	90	88.4	68.5	27.69	62.88	63.39	0.57	1.2	2.77	1.12	2.23
150	PredictPTB	procedures	full	short-term	90	88.32	67.29	26.04	62.31	62.87	0.32	1.86	3.14	1.14	3.64

	Experiment details					Mean					Standard deviation				
	Model	**Modality**	**Experiment**	**Window**	**Prediction Point (days)**	**ACC**	**AUC**	**PRUC**	**Sensitivity**	**Specificity**	**ACC**	**AUC**	**PRUC**	**Sensitivity**	**Specificity**
151	PredictPTB	medications	retro	short-term	180	90.8	66.5	23.86	60.45	59.4	0.24	2.12	2.71	1.82	2.92
152	PredictPTB	medications	full	short-term	180	91.08	66.28	21.23	59.59	62.32	0.36	1.76	2.57	2.35	3.09
153	PredictPTB	medications	retro	short-term	30	88.88	65.33	26.05	59.52	60.63	0.36	1.08	1.19	0.85	1.41
154	PredictPTB	medications	full	short-term	30	88.83	65.09	25.81	59.43	60.06	0.28	1.17	1.3	0.78	2.29
155	PredictPTB	medications	full	short-term	90	89.92	61.69	17.63	57.87	58.46	0.14	1.36	1.03	1.41	2.73
156	PredictPTB	medications	retro	short-term	90	89.87	61.75	17.85	57.54	57.83	0.17	0.86	1	1.63	2.19
157	RETAIN	all	retro	long-term	30	92.27	76.94	31.53	68.31	68.98	0.12	0.44	0.88	0.4	0.59
158	RETAIN	all	full	long-term	30	92.14	76.44	31.11	68.1	67.95	0.07	0.47	0.7	0.38	0.67
159	RETAIN	diagnosis	retro	long-term	30	92.08	75.17	28.42	67.36	67.04	0.11	0.47	0.39	0.5	1.26
160	RETAIN	diagnosis	full	long-term	30	92.12	75.09	28.7	67.35	66.88	0.09	0.49	0.73	0.41	0.88
161	RETAIN	all	full	long-term	90	92	71.86	22.56	65.21	64.79	0.12	0.37	0.56	0.45	0.57
162	RETAIN	all	retro	long-term	90	92.02	72.24	22.67	65.07	65.27	0.1	0.7	1.35	0.48	1.06
163	RETAIN	all	retro	long-term	180	92.09	71.22	19.24	64.94	64.79	0.13	0.42	0.58	0.6	0.89
164	RETAIN	all	full	long-term	180	92.17	71.1	18.84	64.93	64.69	0.09	0.56	0.84	0.35	0.94
165	RETAIN	diagnosis	retro	long-term	90	92.12	71.08	22.05	64.68	64.11	0.12	0.51	0.62	0.55	0.67
166	RETAIN	diagnosis	full	long-term	90	92.07	70.97	21.71	64.36	64.31	0.15	0.35	0.67	0.43	1.07
167	RETAIN	diagnosis	retro	long-term	180	92.1	69.9	17.68	64.05	64.13	0.15	0.63	0.98	0.65	1.45
168	RETAIN	diagnosis	full	long-term	180	92.17	69.67	17	63.9	64.13	0.08	0.46	0.35	0.36	1.12
169	RETAIN	lab orders	full	long-term	180	92.16	67.36	16.08	62.37	62.44	0.09	0.37	0.64	0.51	1.02
170	RETAIN	lab orders	retro	long-term	180	92.08	67.22	16.27	62.12	62.82	0.12	0.5	0.73	0.47	1.03
171	RETAIN	lab orders	retro	long-term	90	91.85	65.87	15.72	61.73	60.22	0.18	0.46	0.65	0.81	1.19
172	RETAIN	lab orders	full	long-term	30	91.67	66.05	17.63	61.28	61.4	0.09	0.41	0.37	1.09	1.09
173	RETAIN	lab orders	full	long-term	90	91.92	65.92	15.5	61.26	60.66	0.07	0.39	0.52	0.47	0.77
174	RETAIN	lab orders	retro	long-term	30	91.68	66	17.72	60.86	61.74	0.08	0.57	0.76	0.72	0.76
175	RETAIN	all	full	long-term	270	91.52	63.43	15.18	59.47	58.57	0.19	0.73	0.63	0.65	1.1
176	RETAIN	medications	retro	long-term	30	91.22	64.18	19.31	59.34	59.06	0.09	0.28	0.88	0.81	1.06
177	RETAIN	all	retro	long-term	270	91.5	63.42	15.38	59.34	58.54	0.19	0.86	0.82	0.66	1.75
178	RETAIN	diagnosis	full	long-term	270	91.53	62.8	14.85	58.99	57.86	0.14	0.84	0.5	0.65	1.36
179	RETAIN	diagnosis	retro	long-term	270	91.59	62.94	14.73	58.81	58.12	0.1	0.28	0.39	0.76	0.91
180	RETAIN	medications	retro	long-term	180	91.56	61.97	14.9	58.79	57.04	0.18	0.79	0.65	0.59	1.66
181	RETAIN	medications	full	long-term	30	91.32	64.26	19.2	58.75	59.59	0.14	0.4	0.48	0.31	0.99
182	RETAIN	medications	retro	long-term	270	91.42	62.18	14.47	58.32	57.61	0.14	0.8	0.74	0.46	1.57
183	RETAIN	lab orders	full	long-term	270	91.47	61.46	14.12	58.26	57.84	0.1	0.75	0.57	1.03	0.95
184	RETAIN	medications	full	long-term	270	91.36	62.56	14.8	58.14	58.09	0.17	0.9	0.68	0.73	1.36
185	RETAIN	procedures	retro	long-term	30	91.18	61.32	15.63	58.07	57.45	0.22	0.7	0.54	0.78	1.3
186	RETAIN	medications	full	long-term	180	91.52	62.06	14.46	58.02	57.95	0.21	0.63	0.8	0.78	0.83
187	RETAIN	procedures	full	long-term	30	91.03	61.1	15.84	57.81	57.1	0.15	0.45	0.58	0.63	1.43
188	RETAIN	lab orders	retro	long-term	270	91.46	61.58	14.24	57.78	58.46	0.13	0.71	0.98	0.87	1.56
189	RETAIN	surgeries	retro	long-term	30	86.43	47.84	15.85	57.46	39.1	1.67	8.64	5.25	9.71	19.06
190	RETAIN	medications	full	long-term	90	91.43	61.74	14.36	57.42	57.98	0.12	0.5	0.68	0.43	1.34
191	RETAIN	medications	retro	long-term	90	91.4	61.85	14.55	57.22	58.06	0.09	0.46	0.38	0.6	1.04
192	RETAIN	procedures	retro	long-term	90	91.35	60.12	13.71	56.91	56.87	0.19	0.57	0.7	1.04	1
193	RETAIN	procedures	full	long-term	90	91.25	59.61	13.78	56.86	56.35	0.15	0.67	0.94	0.63	0.99
194	RETAIN	procedures	full	long-term	180	91.57	57.93	10.81	56.7	54.9	0.21	0.71	0.54	0.65	1.05
195	RETAIN	procedures	retro	long-term	180	91.61	58.95	11.23	56.11	56.86	0.19	0.6	0.48	0.91	1.47
196	RETAIN	procedures	retro	long-term	270	91.67	58.36	10.84	55.83	56.2	0.1	0.71	0.49	0.88	1.41
197	RETAIN	procedures	full	long-term	270	91.54	57.07	10.61	55.66	54.57	0.13	1.15	0.48	0.73	1.87
198	RETAIN	surgeries	full	long-term	180	86.62	48.62	16.49	48.1	50.7	3.44	10.66	5.21	18.23	24.69
199	RETAIN	surgeries	full	long-term	90	89.24	47.87	11.79	44.82	53.94	2.17	9.32	3.2	10.69	9.66
200	RETAIN	surgeries	retro	long-term	180	88.87	51.25	13.01	44.08	57.02	2.25	9.19	2.93	18.72	27.8

	Experiment details					Mean					Standard deviation				
	Model	**Modality**	**Experiment**	**Window**	**Prediction Point (days)**	**ACC**	**AUC**	**PRUC**	**Sensitivity**	**Specificity**	**ACC**	**AUC**	**PRUC**	**Sensitivity**	**Specificity**
200	RETAIN	surgeries	retro	long-term	180	88.87	51.25	13.01	44.08	57.02	2.25	9.19	2.93	18.72	27.8
201	RETAIN	surgeries	full	long-term	30	89.35	55.34	14.21	42.97	61.83	1.73	5.14	3.16	17.29	21.66
202	RETAIN	surgeries	retro	long-term	270	88.89	48.98	14.88	37.34	56.53	4.22	7.99	6.59	17.89	22.8
203	RETAIN	surgeries	full	long-term	270	86.82	48.19	15.05	37.1	61.06	2.03	9.19	3.7	17.84	23.64
204	RETAIN	surgeries	retro	long-term	90	87.27	47.29	15.22	29.2	70.41	3.56	3.35	3.77	18.52	16.23
205	RETAIN	all	retro	short-term	30	92.47	80.01	36.55	71.18	71.07	0.16	0.61	0.78	0.57	0.94
206	RETAIN	all	full	short-term	30	92.46	79.52	35.52	71.07	70.14	0.13	0.51	0.85	0.72	0.53
207	RETAIN	diagnosis	retro	short-term	30	92.3	78.59	33.2	69.87	70.12	95.99	0.57	1.34	0.69	1.06
208	RETAIN	diagnosis	full	short-term	30	92.28	78.43	33.68	69.8	69.89	0.11	0.42	0.75	0.6	0.98
209	RETAIN	all	retro	short-term	180	94.39	76.37	20.86	69.31	68.22	0.12	0.59	0.68	0.62	1.84
210	RETAIN	diagnosis	retro	short-term	180	94.39	76.06	20.89	69.01	68.57	0.17	0.6	0.93	0.73	1.68
211	RETAIN	all	full	short-term	90	92.78	76.48	26.13	68.87	68.22	0.24	0.49	0.82	0.45	0.7
212	RETAIN	diagnosis	full	short-term	180	94.34	76.36	20.26	68.74	69.87	0.11	1.13	1.58	1.07	2.29
213	RETAIN	all	retro	short-term	90	92.67	76.17	26.67	68.39	68.37	0.2	0.76	1.4	0.35	1.18
214	RETAIN	all	full	short-term	180	94.35	76	19.56	68.29	69.01	0.23	0.61	0.78	1.19	0.86
215	RETAIN	diagnosis	full	short-term	90	92.66	74.52	25.84	67.51	66.25	0.09	0.55	1.37	0.54	1.27
216	RETAIN	diagnosis	retro	short-term	90	92.62	74.84	26.53	67.11	67.12	0.12	0.58	0.94	0.54	1.26
217	RETAIN	procedures	retro	short-term	180	92.39	68.44	19.13	64.35	60.72	0.52	2.06	4.15	3.67	3.9
218	RETAIN	lab orders	retro	short-term	180	94.86	69.13	13.64	64.17	63.93	0.1	0.94	0.73	1.19	2.38
219	RETAIN	procedures	full	short-term	180	92.59	69.5	17.74	63.02	64.03	0.25	2.77	2.88	2.02	4.57
220	RETAIN	lab orders	full	short-term	90	92.52	68.42	15.19	62.91	62.68	0.1	0.98	0.53	0.79	1.15
221	RETAIN	lab orders	retro	short-term	90	92.45	68.56	15.39	62.87	62.94	0.2	0.52	0.29	0.94	1.68
222	RETAIN	lab orders	full	short-term	180	94.72	67.96	12.85	62.76	63.38	0.23	0.75	0.88	1.06	2.01
223	RETAIN	procedures	full	short-term	30	88.62	66.58	26.93	62.42	60.64	0.33	0.62	1.07	1.3	1.4
224	RETAIN	procedures	retro	short-term	30	88.53	66.83	27.71	62.15	61.45	0.22	1	1.71	1.3	2.38
225	RETAIN	lab orders	full	short-term	30	91.85	67.28	17.86	62.14	62.1	0.13	0.58	0.67	0.73	0.81
226	RETAIN	lab orders	retro	short-term	30	91.82	67.28	18.2	62.1	62.01	0.11	0.71	0.61	0.69	1.27
227	RETAIN	procedures	retro	short-term	90	89.26	67.09	25.47	61.96	61.98	0.34	1.82	1.73	1.02	2.98
228	RETAIN	procedures	full	short-term	90	89.34	65.23	23.96	61.18	60.17	0.26	1.13	1.67	1.78	1.5
229	RETAIN	medications	full	short-term	180	91.76	61.65	18.49	60.25	56.83	0.39	2.01	1.53	3.37	5.01
230	RETAIN	medications	retro	short-term	180	91.83	62.95	20.41	59.88	57.25	0.23	0.87	1.43	2.44	3.88
231	RETAIN	medications	retro	short-term	30	89.71	64.31	25.63	59.47	59.65	0.21	0.63	0.54	1.24	1.27
232	RETAIN	medications	full	short-term	30	89.77	64.02	24.59	59.21	59.92	0.26	0.94	1.29	1.28	1.64
233	RETAIN	medications	full	short-term	90	89.94	58.86	16.19	56.95	54.14	0.21	1.17	0.89	2.43	3.09
234	RETAIN	medications	retro	short-term	90	90.13	59.47	16.13	56.37	56.14	0.24	0.9	0.94	1.46	2.03