MODELING ACUTE CARE UTILIZATION FOR INSOMNIA PATIENTS

by

Zitong Zhu

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Computer and Information Science at IUPUI

Indianapolis, Indiana

August 2023

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Shiaofen Fang, Chair

Department of Computer and Information Science

Dr. Zina Ben Miled

Department of Electrical and Computer Engineering

Dr. Yuni Xia

Department of Computer and Information Science

Dr. Jiangyu Zheng

Department of Computer and Information Science

Approved by:

Dr. Shiaofen Fang

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Fang for his guidance and support throughout my MS degree. I am grateful to have had the opportunity to work and learn from Dr. Ben Miled. Her commitment to excellence and passion for teaching have been an inspiration to me. This work would not have been possible without the valuable weekly discussions with Drs. Malaz Boustani, Paul Dexter, Rezaul Karim Khandker, and Farid Chekani. My sincere gratitude goes to the members of my thesis committee for their review and feedback.

This work was supported in part by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., 90 E Scott Ave, Rahway, NJ 07065, USA. I would like to thank Jarod Baker and Ravan Carter of the Regenstrief Institute for their support. The high-performance computing services used in this study are supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Machine learning (ML) models can help improve health care services. However, they need to be practical to gain wide adoption. A methodology is proposed in this study to evaluate the utility of different data modalities and cohort segmentation strategies when designing these models. The methodology is used to compare models that predict emergency department (ED) and inpatient hospital (IH) visits. The data modalities include socio-demographics, diagnosis and medications and cohort segmentation is based on age group and disease severity. The proposed methodology is applied to models developed using a cohort of insomnia patients and a cohort of general non-insomnia patients under different data modalities and segmentation strategies. All models are evaluated using the traditional intra-cohort testing. In addition, to establish the need for disease-specific segmentation, transfer testing is recommended where the same insomnia test patients used for intra-cohort testing are submitted to the general-patient model. The results indicate that using both diagnosis and medications as a source of data does not generally improve model performance and may increase its overhead. For insomnia patients, the best ED and IH models using both data modalities or either one of the modalities achieved an area under the receiver operating curve (AUC) of 0.71 and 78, respectively. Our results also show that an insomnia-specific model is not necessary when predicting future ED visits but may have merit when predicting IH visits. As such, we recommend the evaluation of disease-specific models using transfer testing. Based on these initial findings, a language model was pretrained using diagnosis codes. This model can be used for the prediction of future ED and IH visits for insomnia and non-insomnia patients.

CHAPTER 1. INTRODUCTION

Interest in machine learning (ML) has significantly increased in the health sector over the past decade. The fields of neurology, radiology, ophthalmology, anesthesia, intensive care medicine and oncology have all explored ML[1]. While these ML models show potential in experimental settings, their use in real-world practice is limited. To successfully translate ML models from research to practice, the constraints, and limitations of the models in a production environment must be fully understood.

Most ML models in healthcare focus on two main outcomes: chronic disease risk prediction[2] and acute care utilization[3]. The sources of the data for these models vary considerably as does the focus on either diseases or procedures. Data modalities ranging from gene expression, signal data from mobile devices, to laboratory and electronic health record (EHR) data have been used in previous studies[4][5]. Disease-specific random forest (RF) and convolutional neural network (CNN) models have successfully been developed to detect sleep disorder in asthma patients and achieved accuracies of 0.81 and 0.95, respectively[6]. Procedure-specific models tend to focus on post-operative risks. For example, a logistic regression (LR) model was used to predict 90 days readmission after total joint arthroplasty with an area under the receiver operating curve (AUC) of 0.65[7].

Disease-specific models are common, but often not practical. Using training and validation data from a highly specific cohort of patients may improve model accuracy but it may also reduce generalization to different populations[5]. Because disease-specific data is limited, especially if it is sourced from a single health care institution, models developed using these small datasets are prone to overfitting, making them unsuitable for unobserved patient populations[4]. Moreover, when considering the deployment of the models in practice, use of disease-specific models over general ones may require health care institutions to deploy multiple models, one for each disease of interest. On the one hand, the large number of chronic diseases and maintenance of associated models is a high burden for the health care institutions. On the other hand, selectively choosing among the available ML models goes against the intended ubiquity promise of digital health.

Developing general patient models to predict the risk for multiple disease conditions is difficult. However, a general patient model that can predict acute care utilization (ACU) may be possible if a pragmatic design methodology is followed. ACU is defined as the use of hospital

services in the form of emergency department (ED) or inpatient hospital (IH) visits. The first part of the present study introduces insomnia-specific and general patient models to predict ED and IH visits using data from multiple healthcare institutions. These models use data from several modalities including patient's socio-demographics, history of medications and disease conditions. The models are then compared using a structured methodology.

Insomnia was selected as the focus of the present study because it is one of the most prevalent sleep disorders, affecting approximately 20–30% of the adult population in the United States[16][17]. It was also shown to increase health care utilization among Medicare beneficiaries[18]. The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) defines insomnia as dissatisfaction with sleep quantity or quality that is associated with at least one of the following symptoms: difficulty initiating sleep, difficulty maintaining sleep characterized by frequent awakenings or problems returning to sleep after awakenings, or early-morning awakening with the inability to return to sleep[19]. Insomnia often has a significant negative impact on patient quality of life and has been associated with impaired daytime functioning, increased risk of workplace injury and absenteeism, and fatal injuries[20][21]. In addition to its direct negative impact on health-related quality of life and economic productivity, the selection of insomnia is motivated by the fact that this disease can manifest as a primary medical condition or a common secondary comorbid condition to a wide range of chronic conditions such as depressive or anxiety disorders, Alzheimer Disease and other related dementia, and alcohol and substance use disorders[21][22][23]. Moreover, pharmacologic therapy for insomnia should be limited in duration and cognitive behavioral therapy is preferred[23]. The complexity of the disease condition and the restricted treatment options are added incentives for the identification of insomnia patients that may become high acute care users.

The second part of the present thesis investigates transfer learning for ACU prediction using diagnosis data. Transfer learning is a powerful technique that can be employed to address the issues of generalizability and overfitting resulting from limited amount of training data. In comparison with conventional machine learning models, transfer learning frequently achieves better performance and results in practical applications[52]. More specifically, a self-supervised learning approach can be used over large, general datasets to derive a general-purpose pre-trained model that captures the intrinsic structure of the data. This pre-trained model can then be fine-tuned for specific tasks and datasets. The pre-training fine-tuning paradigm has been proven to be effective

for natural language processing (NLP) applications. For instance, the approach was used to create language models (LM) such as the Bidirectional encoder representations from transformers (BERT) [28]. BERT is pre-trained with general English text from BookCorpus and Wikipedia. Subsequently, other LMs were developed using the same methodology and model architecture but pre-training was performed over a medical corpus. Examples of these domain-specific LMs include BEHRT[41] and Med-BERT[40].

In this study, an LM is developed to enable ACU prediction. This LM uses only the patient's history of diagnosis. Diagnosis codes are represented using one-hot encoding and used as input for the language model. Two tasks are then applied during the pre-training phase to capture the semantics of the data. The downstream task focuses on ACU prediction. The proposed LM is tested on both insomnia and non-insomnia patients' data.

CHAPTER 2. BACKGROUND

Machine learning is being increasingly used in healthcare to predict disease risk and improve ACU models. Specifically, traditional models like support vector machine (SVM), RF, deep learning models like recurrent neural network (RNN), long short-term memory networks (LSTM), have shown promise in accurately predicting disease risk based on various patient data, such as genetic information, medical history, and lifestyle factors. Machine learning algorithms have also been utilized in support of ACU prediction, primarily for the identification of patients at high risk of hospital readmission or future emergency department visits. By identifying high-risk patients, healthcare providers can take proactive measures to reduce ACU. Choi et al. (2016) [42] found that a machine learning model outperformed traditional risk prediction models in identifying patients at high risk of hospital readmission.

The present thesis focuses on ACU prediction using machine learning. The goal is to develop models that accurately predict the risk of hospital admission or emergency department visits using the patients' medical history and socio-demographic factors. By improving ACU prediction, healthcare providers can attend to the patients with high risk level.

The remainder of this chapter offers a review of previous applications in modeling ACU as well as the relevant machine learning techniques that can support these applications.

2.1 Modeling ACU

Managing ACU is important for both health care institutions and patients. For health care institutions, the concern is operational efficiency, and the aim is to manage surges in demand by increasing resources, reducing the length of stay, or redirecting patients. Few operational support models have been proposed to predict hospitalizations, and most are specific to individual healthcare institutions. For example, a neural network model was developed to predict surges in demand for ED services by patients with chronic pulmonary diseases[8]. This short-term forecast model relies on weather and environmental conditions specific to the geo-location of the health care institution. A long-term forecast model for ED visits by patients with non-communicable diseases was also proposed[9]. This model is based on two assumptions: (a) non-communicable diseases are the most important reason for ED visits and (b) the incidence of this type of disease

increases with age. Other models focus on predicting the transition from ED to IH visits[10][11]. Although these transition models include patients' demographics and medical history as exposure variables, they rely heavily on non-generalizable social determinants (e.g., zip code, mode of arrival) thereby making the models less transferable to other communities[10].

Identifying the determinants of high ACU for each individual patient was the focus of numerous previous studies. A systematic review of statistical models of ED visits by older adults found that measures of need such as prior ED or IH visits are statistically significant determinants of future ED visits[12]. Other LR models for ACU in the general patient population used exposure variables defined in the adjusted clinical group (AGC) scoring instrument[13]. These variables include demographics, utilization, diagnosis and medication with values derived from claims data.

Institution- and patient-centric models for ACU are complementary and serve different purposes. The institution-centric model is unlikely to transfer to other health care institutions. The patient-centric model, if designed properly, may transfer to patients in other health care institutions. Both institution- and patient-centric models can be designed for the general patient population or a subset of this population (e.g., specific disease condition or age group). However, the more specific the target population, the higher is the deployment burden on the health care institution.

The ability to understand, estimate or potentially reduce ED and IH visits in any patient population can have a significant impact on the health care system. Differences in cost between inpatient and outpatient health services are easily measurable and provide a strong supporting argument for adopting ACU models into clinical practice. Moreover, early identification of the small percentage of patients that are high acute care users is critical because they account for a large proportion of the health care costs[14]. However, the proposed models must be scalable and preferably applicable to the wider patient population. If unnecessary, targeting a specific disease condition or age group may perpetuate some of the practices specific to survey-based instruments, introduce unneeded operational burden and infringe on the added benefit afforded by new modeling techniques.

One of the few earlier studies that compares general patient ML models to disease-specific models aims to predict hospital readmission within 30 days[3]. This outcome is different from the two outcomes considered in the present study. However, this earlier study shares some of the practical deployment concerns with the present study; concerns which are often ignored by other ML studies.

The above-mentioned readmission study collected 3.3 million hospital admissions from the New Zealand hospital system over a 6-year period starting from 2006. These data were used to develop models from a multi-modal feature space that included:

- Socio-demographics: sex, age, and race of the patient; number of hospital visits within the past 365 days; and type of health care institution (i.e., public versus private).
- Diagnosis: disease codes are collected for each patient following the International Classification of Diseases-10-AM (ICD-10-AM)[15], after removal of infrequent codes.
- Procedure: list of procedures associated with each hospital admission are also derived from the ICD-10-AM classification.

A subset of 280 disease groups of interest were identified from a total of 815. The selection criteria were based on high cost/penalty for readmission within the New Zealand health care system. Several ML techniques including LR, RF and SVM were considered. When all 280 disease groups are modeled collectively, the best AUC (0.82) was achieved by the LR model. When each disease group was considered independently, the AUCs of the best disease-specific models range from 0.57 to 0.95. As a result, the authors concluded that while some disease-specific models may show better performance, they are more likely to suffer from limited sample sizes making the general patient cohort model more practical. In the specific case of five chronic diseases (Chronic obstructive pulmonary disorder, Heart failure, Pneumonia, Acute myocardial infarction, Total hip arthroplasty/total knee arthroplasty), the study was also extended to include a deep neural network (NN) model with three hidden layers. The results indicate that the AUC of these deep NN models was similar to that of their LR counterparts.

2.2 ML Techniques

As discussed above, machine learning models can help improve health services and in particular they can help identify high acute care users. These models can be created using various techniques, each with its own strengths and limitations.

For example, SVM have been utilized for classification and prediction tasks in various medical applications, including cancer diagnosis, gene expression analysis, and medical image processing [44][45]. Random Forest (RF) has shown promise in predicting heart failure and identifying factors associated with medication non-adherence [46][47]. Similarly, Neural

Networks (NN) have been used to predict the risk for developing Alzheimer's disease and for readmission in patients with heart failure [48][49].

In the case of future ED and IH visits, machine learning models have been developed to predict these outcomes using various techniques, including SVM, RF and NN [8]. These models were trained using various types of data, including demographics, medical history, laboratory test results, and medication data [18].

In the first part of the present thesis, SVM, RF, NN, LR, and Extreme Gradient Boosting (XGB) are used to develop prediction models for ED and IH visits. The performance of the resulting models is compared in order to identify the most effective ML technique for ACU prediction. Different data modalities are also evaluated to assess the contribution of each modality to the predictive performance of the models. Moreover, transfer testing is used to compare the performance of the models for insomnia and non-insomnia patients.

In the second part of the present thesis, a transformer model is pre-trained using the disease conditions modality. The pre-trained language model (LM) is then fine-tuned to predict the risk for future ACU. By comparing the performance of this pre-trained model to the traditional machine learning techniques developed in the first part of the thesis, we aim to evaluate the potential of the proposed LM for predicting ACU as it offers the ability for adaptation through fine-tuning to other patient cohorts and health care institutions.

Previous LM derived for medical data include BEHRT[41] and Med-BERT[40]. BEHRT[41] was developed in 2019 and was an initial attempt at creating EHR domain-specific token embeddings. The model leveraged the wealth of information available in EHR data, including patient encounter sequence, diagnostic codes, and age at each encounter, to create token embeddings. The pre-training stage of BEHRT utilized the Masked Language Model (MLM) task. This task involves masking a proportion of the input tokens and predicting the masked tokens based on their surrounding tokens. The goal of this task is to a) capture the contextual dependencies between tokens and b) the development of semantic representations of the tokens. Three downstream tasks related to disease prediction were then used to evaluate the performance of BEHRT:

1. prediction of concepts in the next visit,
2. prediction of the occurrence of diseases in the next six months, and
3. prediction of the occurrence of diseases in the next 12 months.

The downstream tasks were defined as input-output pair for each patient. Specifically, a random length of continuous visit records (greater than 3) is selected from the patient's diagnosis history as the model input. The model output is a one-hot encoded multi-segment vector where each bit position represents diseases that exist in patient future's record over different prediction horizons (i.e., the next visit, within 6 months, with 12 months).

BEHRT was trained and evaluated on data from nearly 1.6 million individuals. It demonstrated an improvement of 8 to10% in terms of Average Precision Score when predicting the onset of more than 300 disease conditions compared previous models including DeepR[54] and RETAIN[51].

Subsequently in 2020, a Med-BERT[40] was introduced. This LM is an adaptation of the BERT architecture to structured EHR. Med-BERT uses disease codes and visit sequences. It was pre-trained on a dataset containing 28,490,650 patient records. Moreover, the pre-training procedure employed a novel task focused on predicting prolonged hospital stays, which encouraged the model to capture contextual information among visit sequences. The performance of Med-BERT was evaluated for three specific disease prediction tasks after fine-tuning. The results show an improvement of 1 to 6% in Area Under the ROC curve (AUC) compared to baseline models such as GRU [50], RETAIN[51], and LR. This modest improvement may be attributed to the complexity of the multi-label disease prediction task. That said, this earlier study was able to show that the BERT architecture is able to learn the meaning of disease codes and their respective taxonomies under different contexts and with different vocabularies[40]. Therefore, this architecture may prove valuable in facilitating downstream tasks such as ACU prediction. This aspect is explored in the second part of the present thesis. Following the Med-BERT architecture, a transformer model is pre-trained using diagnosis data. The model is then fine-tuned for the specific tasks of predicting future IH visits.

CHAPTER 3. METHODS

The patients in the present study are served by any healthcare system providing data to the Indiana Network for Patient Care (INPC), an operational community-wide electronic medical record. The National Library of Medicine and the Agency for Healthcare Research and Quality have supported the initial development of the INPC. The system currently includes data from 19 hospitals in seven health systems, the Marion County Health Department, regional laboratories, radiology centers, and various physician practices. These hospitals account for over 95% of all beds and ED visits in Indianapolis. The data collected include demographics, laboratory results, emergency department, inpatient, and outpatient encounter data, free-text chief complaint, coded diagnoses and procedures, vital signs, and other data, but not all these data elements are available for every participant. These data are accessible to care providers and researchers through the standardized and centralized Medical Record System of the Regenstrief institute. One of the advantages of the Regenstrief Medical Record System is that patients with multiple medical records from different health care institutions are linked. For instance, when a patient is seen in any of the 19 emergency departments operated by the consortium of hospitals, the information from all these health care institutions about one patient can be presented as one virtual medical record.

All procedures performed in the study were in accordance with the ethical standards of the institutional review board of Indiana University and with the 1,964 Helsinki declaration and its later amendments or comparable ethical standards. The experimental protocol was approved by the Institutional Review Board of Indiana University (IRB Number: 11,732). The study was exempt from patient informed consent by the Institutional Review Board of Indiana University due to the non-interventional and retrospective nature of the study.

3.1 Evaluating disease specific ACU models

The models developed in the present study are derived from EHR data and use different ML techniques to predict two outcomes: IH and ED visits. We selected an insomnia patient cohort for the purpose of disease-specific modeling because of the prevalence of this disease condition and routine care EHR as a source of data because it is widely available for all patients. Relying on

exposure variables from signal or imaging data (e.g., polysomnography) will limit the number of patients available for the study[1]. More importantly, it will restrict the application of the proposed model to a limited cohort of patients for whom the results of these special diagnostic instruments are available since insomnia diagnosis and follow up do not often require complementary methods, and polysomnography is rarely performed[24].

3.1.1 Study population

Two different cohorts are considered: Insomnia patients and general patients. All patients in the two cohorts must be at least 18 years of age and have at least one encounter on record in the Medical Record System every year from 2010 to 2019. This ensures that selected patients are adults, enrolled in the network and are using the health services over the study period. The insomnia patients are identified according to two criteria:

- A new insomnia diagnosis, or
- A new insomnia medication order.

Insomnia diagnosis is established using a list of ICD[25] codes in the patient records. Similarly, a list of FDA and non-FDA approved insomnia medications was collected from the literature and used to identify patients that have medication orders consistent with the target list. This list includes zolpidem, suvorexant, butabarbital, quazepam, estazolam, flurazepam, triazolam, tasimelteon, eszopiclone, temazepam, ramelteon, secobarbital, zaleplon, chloral hydrate, and melatonin. Extending the medication list to both FDA and non-FDA approved medications allows the inclusion of undetected insomnia patients in the study cohort. The index date for each insomnia patient is defined as the date of the earliest one of the above two selection criteria. Only patients with index date between January 1st, 2011, and December 31st, 2019 (the study period) were considered in order to ensure that each patient had at least one year of history data prior to the index date.

The proposed models focus on estimating ACU for incident cases of insomnia. As such, only new cases of insomnia are considered. Moreover, antidepressants (e.g., doxepin, trazodone, and mirtazapine) and low-dose antipsychotics (e.g., quetiapine and olanzapine) with hypnotic properties were excluded from the list of medications used for insomnia diagnosis. These off-label treatments may not be specific to insomnia as they are also prescribed for depression, pain,

psychosis and other medical conditions. Verification of off-label medication use for insomnia treatment requires a review of the medical notes.

The second cohort consists of general patients that do not satisfy any of the above two criteria over the study period. For each insomnia patient, a matching general patient was selected from the Medical Record System. The match was performed based on age at index date (within one year), sex and race. When multiple matches are available, one patient is selected at random.

For notation simplicity, the insomnia patients are labeled "cases" and the general patients are labeled "controls" in the remainder of the present thesis. Moreover, the entire cohort of either cases or controls is labeled "all", the subset of cases with an insomnia diagnosis (i.e., excluding cases that have been prescribed insomnia medications with no documented insomnia diagnosis in their EHR) is labeled "ICD" and the subset of patients 65 and older is labeled (65+).

3.1.2 Exposure variables

Exposure variables were extracted from routine EHR data and grouped according to the following three modalities:

- Socio-demographic: Variables in this modality include age, sex, race, ethnicity, insurance type (i.e., government, commercial, self-pay, and other/unknown) and the neighborhood deprivation index[26] which is derived from the last patient's address on file at the end of the study period (2019). Three variables are added to this list: the number of ED and IH visits during the year prior to the index date and the Elixhauser's comorbidity score[7]. This score corresponds to the weighted sum of the 31 Elixhauser's comorbidity categories[27].
- Diagnosis: The disease conditions in the patient's medical record are collected over a period of one year prior to the index date (non-inclusive). The disease ICD codes are then aggregated and mapped to the 31 Elixhauser's comorbidity categories which range from congestive heart failure to depression. This specific comorbidity index was selected because it is based on a widely used disease code taxonomy and includes disease categories related to mental health. Alternative comorbidity indices were considered. Both the Charlson Comorbidity Index[31] and the NCI Comorbidity Index[32] have limited emphasis on mental health diseases and the latter was customized for cancer patients.

- Medications: A similar data aggregation approach to that used for the diagnosis modality is applied to the medication orders. For each patient, the medications in the patient's record during the year prior to the index date are mapped to the appropriate Anatomical Therapeutic Chemical (ATC)[33] first level subgroup. There are fourteen main groups under the ATC taxonomy. Each group corresponds to a biological system (e.g., nervous system or respiratory system) and may have from 3 to 16 subgroups. For example, the nervous system group has seven subgroups including anesthetics and analgesics. A total of 86 medication exposure variables are developed where each variable corresponds to an ATC medication subgroup. The mapping was performed by using the RxNav API[34]. Some of the ATC subgroups were omitted because they are not relevant to this study (e.g., diagnostic agents, contrast media, diagnostic radiopharmaceuticals, etc.). Although other medication taxonomies are available (e.g., the Generic Product Identifier [58], the USP Drug Classification and the First Databank's Drug Classification [59]), the choice of the ATC taxonomy was motivated by its wide-use and hierarchical structure. The first characteristic was necessary to support the design of a generalizable model for patients from multiple health care institutions and the second was needed to facilitate the aggregation of different types of medications.

3.1.3 Predictive models

The exposure variables described in the previous section are used to predict the ACU for cases and controls during the year following the index date. Since the control patients are selected randomly, this cohort is representative of the general population of patients affiliated with all the health care institutions contributing data to INPC. Moreover, matching between cases and controls was based on age, sex and race. Therefore, potential biases inherent to the insomnia population due to these variables are reduced. Two outcomes are analyzed: ED and IH visits. Each outcome is stratified into two classes: (1) zero visit and (2) one or more visits. This stratification corresponds to a binary classification. A regression model where the outcome is the number of visits was also considered. However, given the low number of patients with one or more visits compared to the number of patients with zero visits in the study cohorts, binary classification was best suited for identifying at risk patients of ACU. Using this binary stratification, different models were developed by varying the cohort, the exposure variables, and the outcome variables.

In a first step the socio-demographics variables are used to train models using cases and controls independently. These models are compared to models that are developed using the combinations of (a) the socio-demographic and the diagnosis variables and (b) the socio-demographic and the medications variables. The purpose of this comparison is to identify differences in entropy among the socio-demographics, diagnosis, and medication modalities. A global model that combines all the exposure variables is also developed. Again, the purpose of this step is to identify the modality and the variables with the highest predictive power for ACU in insomnia and control patients.

In previous studies, the prevalence of insomnia was found to vary depending on age[23]. Moreover, previous research suggest that more disease-specific cohorts tend to be more homogeneous[3]. In order to validate these findings through sensitivity analysis, additional models are developed for two sub-cohorts: 65+ years patients and ICD cases.

For all scenarios, the models induced by the insomnia cases are compared to the models induced by the control patients. The goal of this comparison is to establish whether an insomnia-specific model for ACU is necessary. This assessment is validated using a technique inspired by transfer learning[35]. As mentioned in Chapter 2, transfer learning is a process that aims at transferring knowledge from one domain to a related domain. For instance, transfer learning was used to fine tune an image classification model for dementia[36]. Since a limited number of MRI image samples are available for dementia patients, the authors started with a pre-trained model for classification of general-purpose images. They subsequently fine-tuned this model to the specific task of identifying MRI images for dementia patients using a limited number of samples. In the present study, the idea of transfer from one domain to another is used for the purposes of validation and testing. Transfer testing is used to evaluate the performance of models trained using a general patient cohort when tested on a cohort of insomnia patients. If the observed performance is equivalent to that of a model natively trained on insomnia patients, then one can stipulate that an insomnia specific ACU model is not needed.

For all the above scenarios five different ML techniques are investigated: SVM, RF, Neural Networks (NN), LR and Extreme Gradient Boost (XGB) in order to assess the findings with respect to a wide range of ML techniques. Moreover, all models are evaluated using a single metric: AUC. This metric was selected because it is a measure of the ability of the models to discriminate between the two outcome classes independent of any threshold value. The architecture and hyper-

parameters of the models are included in Table 1. For each model, the hyper-parameters were fined-tuned using a cross-validated grid search over the training dataset with AUC as the optimization metric.

Table 3.1 Hyper-parameters and architecture of the models. Optimal hyper-parameters values were obtained using a grid search over cross-validated models for each architecture.

Model	Hyper-parameters
SVM	Kernel = RBF, C = [0.1–10], Gamma = [0.001–5]
RF	Number of estimators = [50–300], Maximum depth = [5–30], Minimum samples split = [2–15], Minimum samples leaf = [1–15]
NN	Number of layers = [1–5], Number of nodes per layer = [10–100], learning rate = [0.01–0.05], Optimizer = Adam
LR	L2 penalty with C=1.0
XGB	Number of estimators = [30–200], Maximum depth = [3–5], Minimum child weight = [1–10], Gamma = [0.5–5], Learning rate = [0.01–0.05]

3.2 Pretrained language model for ACU

In the second phase of this research, a more complex model architecture is explored. Specifically, the Med-BERT architecture is pre-trained and used to develop classifiers for ACU. As discussed in Chapter 2 the original Med-BERT architecture was used to develop an LM based on ICD9 and ICD10 disease codes. This LM was then fine-tuned and evaluated on risk prediction for a given disease condition from the list of current comorbid condition of the patient. In the present study, the Med-BERT architecture is pre-trained using ICD 10 codes towards the objective of predicting ACU. The study is limited to ICD 10 codes because ICD 9 codes are no longer used by health care systems. Several pre-training tasks are investigated, and a down-stream classifier is generated for IH visits for insomnia patients.

3.2.1 Dataset

The same study population and disease cohort introduced in Section 3.1.1. is also used to derive the data needed to pre-train and evaluate the proposed LM. Additionally, a general patient cohort was introduced in support of pre-training. Namely, the three cohorts of patients include:

- Insomnia patients (cases), identified through ICD diagnostic codes and insomnia medication orders.
- Matching controls by age, sex, and race.
- General patients which are selected randomly from INPC without any disease's specific exclusion or inclusion criteria.

The data related to the control and general patients cohorts are used to pre-train the LM and the data from the cases cohort are used to fine-tune and validate a IH prediction model specific to insomnia patients. The result of the first phase of the research, as described in Section 3.1., was used to guide the second phase. These results indicated that using both Dx and Rx data do not improve model performance and that Dx data tend to be more complete within the EHR and more generalizable than Rx data across health care institutions. Different health care institutions may use different medications taxonomies (e.g., ATC, GPI, et…) making the translation of medications from one health care institution to another difficult. Based on these findings, Dx was selected as the data modality of choice for the development of the proposed LM.

For the diagnoses, we utilized the general equivalence mappings [57] provided by the National Bureau of Economic Research to map all ICD-9 codes to ICD-10 codes. We focused on primary diagnoses only. We also limited the diagnosis codes to three characters, following the methodology introduced in [55] and [56]. To avoid bia due to recurring codes during long-term IH visits, we de-duplicated per-visit diagnoses. Lastly, we denoised patient Dx code records by dropping any code with a frequency lower than 100 among the dataset.

Additionally, exposure variables that have been verified as predictive in the preliminary evaluation were also considered when pre-training the proposed LM. These include age, sex, race, history of IH visit, and history of ED visit. The impact of each of these variables towards ACU risk prediction is investigated using the ablation method.

The results of phase 1 also indicated that a disease-specific IH model is more beneficial than a disease-specific ED model, since the general patient population ED model can be efficiently applied to insomnia cases. Therefore, the second phase of this study focuses on evaluating the efficiency of the pre-trained LM model on facilitating a fine-tuned IH risk prediction model.

To be consistent with the Med-BERT architecture, only diagnosis codes are used to pre-train the proposed LM. These include ICD10 codes, as well as the priority of these disease codes in a given visit, the admission time, the discharge time, the encounter type (i.e., emergency, inpatient, outpatient), insurance information. Each patient record is treated as a sequence of events that represent the patient's diagnosis history. Within this sequence, the ICD code for each diagnosis is one-hot encoded and sorted according to the encounter time. The order of the codes in each visit will is determined by the Dx priority. The patient's medical history over 10 years is used to train the model.

During pre-training, we attempted to improve the performance of the LM by using a data augmentation method. Specifically, we segmented the patient's medical history using each encounter as a longitudinal event. Each segment was then treated as an independent data sample with an observation period of one year. This approach increased the size of the dataset by nearly 50 times. Moreover, limiting the observation period to one year helped meet the maximum length limit of the input sequence imposed by the BERT architecture.

3.2.2 Model architecture

The architecture of the proposed LM is based on the Med-BERT architecture, which incorporates bidirectional transformers and multi-level embeddings. Four types of embeddings are used: the diagnosis codes, the priority of the codes within each visit, and the order of the visits. This is illustrated in Figure 1, where:

- The first segment represents a unique patient identifier,
- The second segment is a binary label that represents different meanings depending on the pre-training task being implemented. For instance, when the pre-training task is length of stay (LOS), a label of 1/0 is used to indicate whether or not the patient had an LOS of more than 7 days/3 days during the observation period.

- The third segment is a sequence of token IDs where each code represents an ICD10 code of a specific disease condition. For example, token IDs 263 and 23 may correspond to diabetes and hypertension, respectively. Moreover, the order of the codes follows a decreasing order of priority. Therefore, diabetes has higher priority than hypertension if they occur within the same visit as described next, and
- The fourth segment of the vector indicates the order of the visits and the diagnosis codes that are assigned to the same visit. For instance, diabetes and hypertension with token IDS 263 and 23, respectively occurred in the same visit (i.e., visit 1).

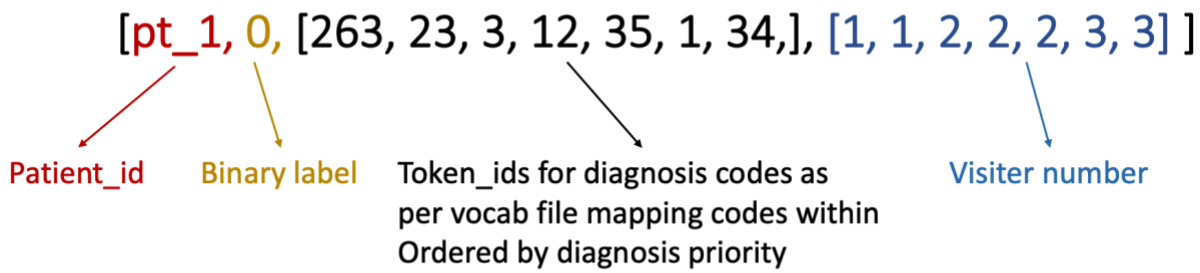

Figure 3.1 Four basic segments of embeddings used to encode the patient's medical history.

Additional embeddings are considered following a similar encoding methodology. The embedding encode age, sex, race, prior IH visits, prior ED visits; and Dx code frequency as follows:

- Socio-Demographic embedding: age information is treated as an independent segment, which has the same length as the Token_ids sequence where for each corresponding ICD code the patient's age at the time of encounter is recorded. The same encoding method was used for sex and race.
- Previous IH/ED visits embedding: This embedding represents the number of IH/ED visits that occurred during the one year observation period. This embedding is only used to supplement the history of encounters when the pre-training task is the number of encounters in the next year. For the LOS pre-training task, an observation period of ten years is used and the previous IH/ED visits embedding is not used.

- Dx frequency embedding: This embedding is an encoding of the 31 Elixhauser's comorbidity categories and the count of the conditions under each category over the observation period. Using this embedding instead of the third segment of the original MedBERT embedding discussed above can greatly reduce the length of the input sequence, thereby reducing the resources and time required for pre-training.

3.2.3 Pre-training

Three different pre-training tasks are considered during the pre-training of the proposed diagnosis LM.

- **Masked language model (MLM)**: This task is at the core of the contextualized embedding model. The MLM task was first introduced in the BERT model. Since then, several other models, including Med-BERT and BEHRT have also used it as a pre-training task in the medical domain. For each input sample, up to two codes in each input sequence are randomly substituted. The selected code for substitution has an 80% probability of being replaced by [MASK], a 10% probability of being replaced by a random code, and a 10% probability of remaining unchanged. By masking out arbitrary codes in each input sequence and then tasking the model with the prediction of the missing codes, the model must rely on the left and right contexts to make the necessary inference.
- **Prediction of prolonged length of stay (Prolonged LOS):** As opposed to BERT, which uses question-answer or classification tasks for pre-training, Med-BERT introduced the prolonged LOS task. In this case, the LM is trained to predict whether a patient will have a prolonged stay in the hospital which is defined as a stay longer than 7 days. This task takes in a sequence of 10 years of history diagnosis codes for a given patient and generates a binary classification indicating whether the patient is likely to have a prolonged stay or not.
- **Number of Encounters in the next year (>10):** This pre-training task was introduced in this study to improve the ability of the proposed LM to predict ACU. It takes the patient's history of diagnosis codes in the past year as input and predicts whether the patient will have more than n encounters in the upcoming year. In the present study,

the value of n was selected to create a balanced dataset between positive (i.e., patients with more than n encounters) and negative (i.e., patients with less than n encounters) samples. A value of n=10 was selected. Future work will explore methods for optimizing this hyperparameter.

Compared to predicting length of stay (LOS) which can be applied to an observation period of any length, it was necessary when predicting the number of encounter to limit the observation period to one year in order to allow for data augmentation. Two pre-training approaches are evaluated, both use the MLM task and in addition either the prolonged LOS or the encounter frequency. The latter two pre-training tasks embody the trade-off between data quality and data quantity when training LMs. In fact, previous studies have shown that pre-trained models are largely impacted by the volume of data used during pre-training [53].

3.2.4 Fine-tuning

First, a Recurrent Neural Network (RNN) prediction head is added to the Med-BERT architecture as recommended in [40]. Second, the contextualized embedding generated from the pre-training model described above is loaded. Third, the parameters of both the pre-training model and the RNN prediction head are updated using gradient descent. The LM model is pre-trained on non-insomnia patients (i.e., the general population). It is then fine-tuned on the downstream task of predicting IH visits for both insomnia and non-insomnia patients. This allows for a better understanding of the benefits of the BERT architecture over traditional machine learning models when predicting patients at risk of high ACU.

CHAPTER 4. RESULT

4.1 Evaluating disease specific ACU models

The selection criteria described in Section 3.1.1 identified 22,120 insomnia cases (Table 4.1). These were matched with 22,120 controls. Of the total number of cases, 65% had no ED visits (Class 0) during the year after the index date and the remaining 35% had one or more ED visits (Class 1). Similarly, of the total number of controls, 78% had no ED visits and the remaining 22% had one or more ED visits. This class imbalance was also observed for the IH visits as shown in Table 4.1. Adequate training of ML models is best achieved with balanced datasets. Moreover, to support a rigorous comparative analysis between disease-specific and general patient models, the number of patient records used to develop and test the models had to be consistent. Therefore, and as dictated by the minority class, the development and testing of the ED and IH models included 4,782 and 3,080 patients from each class for a total of 9,564 and 6,160, respectively.

Table 4.1 Number of patients in Class 0 (0 visits) and Class 1 (one or more visit) for the insomnia patient cohort (cases) and the general population cohort (controls).

	Cases	Controls
ED visits		
Class 0	13,659 (65%)	17,338 (78%)
Class 1	7,541 (35%)	**4,782** (22%)
IH visits		
Class 0	15,299 (73%)	19,040 (87%)
Class 1	5,901 (27%)	**3,080** (13%)

Typically for ML models, 80% of the data is used for training while 20% is used for testing. In this study, because of the attrition due to the sensitivity analysis for patients 65+ years old and the insomnia cases with ICD diagnosis code, a 70/30 split was adopted to allow for sufficient test samples. The split in number of patients for the age-specific models (65+) and the (ICD) cases models are obtained by applying the appropriate selection criteria to the (all) cases or controls datasets. The number of development and test patients for all models are shown in Table 4.2.

Table 4.2 Number of patients in each class for each cohort and outcome.

	Cases (all)	Controls (all)	Cases (65+)	Controls (65+)	ICD (cases)
ED visits					
Class 0	6,694	6,694	2,403	2,403	3,431
Class 1	2,870	2,870	1,031	1,031	1,471
IH visits					
Class 0	4,312	4,312	1,789	1,789	2,654
Class 1	1,848	1,848	767	767	1,138

Table 4.3 includes the AUC of the models that use only the sociodemographic (So) variables. The AUCs and the 95% confidence interval (CI) using bootstrapping are obtained over the test datasets. The top section of the table is dedicated to ED visits while the bottom section reports the classification results for IH visits. The AUC for the ED cases models varies between 0.57 and 0.62 compared to a range between 0.62 and 0.66 for the ED control models. A drop in performance is observed when the patient population is limited to the age group 65+ for both cases and controls. This drop may be due to the difference in number of samples used in the (all) and the (65+) models. Table 4.3 also shows that the AUC of the models induced by the (all) cases and (ICD) cases cohorts are comparable despite the lower number of samples used to develop the latter models. Moreover, the results in Table 4.3 indicate that the (all) cases and controls models have a CI with a range less than 5% whereas for the 65+ models, the CI range is greater than 5% indicating that the (all) models have a higher level of certainty. Similar trends are observed for the IH visits.

Table 4.3 Intra-cohort evaluation: AUC and 95% confidence interval of test samples for models trained on sociodemographic variables (So).

	Cases (all)	Controls (all)	Cases (65+)	Controls (65+)	Cases (ICD)
ED visits					
SVM	0.61[0.59,0.63]	0.64[0.62,0.66]	0.58[0.55,0.62]	0.58[0.54,0.61]	0.61[0.58,0.64]
RF	0.61[0.59,0.63]	0.66[0.64,0.68]	0.57[0.54,0.61]	0.54[0.50,0.57]	0.61[0.58,0.63]
NN	0.57[0.54,0.59]	0.62[0.60,0.64]	0.54[0.50,0.57]	0.55[0.51,0.58]	0.59[0.57,0.62]
LR	0.57[0.54,0.58]	0.60[0.58,0.62]	0.57[0.54,0.61]	0.58[0.54,0.61]	0.56[0.54,0.59]
XGB	0.62[0.60,0.64]	0.66[0.64,0.68]	0.56[0.53,0.60]	0.58[0.55,0.61]	0.61[0.58,0.64]
IH visits					
SVM	0.58[0.56,0.61]	0.57[0.55,0.60]	0.55[0.52,0.59]	0.59[0.55,0.62]	0.58[0.55,0.61]
RF	0.59[0.56,0.61]	059.[0.56,0.61]	0.53[0.49,0.58]	0.57[0.53,0.60]	0.61[0.58,0.64]
NN	0.58[0.55,0.61]	0.57[0.54,0.60]	0.56[0.52,0.61]	0.59[0.56,0.63]	0.60[0.56,0.63]
LR	0.57[0.54,0.59]	0.51[0.49,0.54]	0.56[0.52,0.61]	0.60[0.56,0.64]	0.59[0.55,0.61]
XGB	0.59[0.56,0.61]	0.59[0.56,0.61]	0.54[0.50,0.58]	0.57[0.53,0.60]	0.62[0.59,0.65]

The AUC of the models that use the disease conditions variables (Dx) in addition to the So variables are reported in Table 4.4. When compared to the So-only counter-part models (Table 4.3), the results show an improvement in AUC for all models. The increase in AUC is 0.10 or higher for most models. The cases (ICD) models benefited from the highest improvement with an increase of more than 0.20 in AUC for IH visits. This increase in performance is attributed to the added entropy from the disease conditions modality. The CI range of the models after the addition of the Dx modality is consistent with the baseline So model. Moreover, the ED cases models (columns 1 and 3) still have a lower performance than the ED controls models (columns 2 and 4). For the IH visits, with the addition of the Dx variables, the cases models show higher AUCs than the corresponding controls models. This was not applicable for the models that only used the So variables. The best IH models are the (ICD) cases models with AUCs greater than 0.80 for all ML techniques.

Table 4.4 Intra-cohort evaluation: AUC and 95% confidence interval of test samples for models trained on sociodemographic variables (So) and Elixhauser comorbidity conditions (Dx).

	Cases (all)	Controls (all)	Cases (65+)	Controls (65+)	Cases (ICD)
ED visits					
SVM	0.70[0.68,0.72]	0.74[0.72,0.76]	0.67[0.63,0.70]	0.70[0.66,0.73]	0.72[0.69,0.74]
RF	0.70[0.68,0.72]	0.76[0.75,0.78]	0.68[0.65,0.71]	0.68[0.64,0.71]	0.72[0.70,0.74]
NN	0.70[0.68,0.72]	0.76[0.74,0.77]	0.65[0.62,0.68]	0.76[0.63,0.70]	0.72[0.70,0.74]
LR	0.69[0.68,0.71]	0.74[0.72,0.75]	0.67[0.63,0.70]	0.69[0.67,0.72]	0.71[0.70,0.74]
XGB	0.70[0.68,0.72]	0.76[0.74,0.77]	0.67[0.63,0.70]	0.69[0.66,0.73]	0.72[0.70,0.75]
IH visits					
SVM	0.76[0.74,0.78]	0.72[0.70,0.74]	0.73[0.70,0.76]	0.71[0.67,0.75]	0.83[0.81,0.85]
RF	0.76[0.74,0.78]	0.75[0.72,0.77]	0.73[0.70,0.76]	0.71[0.68,0.75]	0.84[0.82,0.86]
NN	0.73[0.73,0.77]	0.71[0.68,0.73]	0.73[0.70,0.76]	0.68[0.64,0.72]	0.83[0.81,0.85]
LR	0.74[0.71,0.76]	0.71[0.68,0.73]	0.74[0.71,0.77]	0.70[0.67,0.74]	0.83[0.81,0.85]
XGB	0.77[0.75,0.79]	0.74[0.72,0.76]	0.72[0.68,0.75]	0.72[0.68,0.76]	0.84[0.82,0.86]

The performance of the models that combine the So variables with the ATC medication groups variables (Rx) is shown in Table 4.5. Compared to the baseline So models (Table 4.3), the addition of the Rx variables leads to an increase in performance similar to that observed after the addition of the Dx variables. Moreover, most of the trends observed with the addition of the Dx variables are maintained when the So variables are combined with the Rx variables. The predictive performance of the models in Table 4.4 is also equivalent to that of the corresponding models in Table 4.5. This finding is interesting but not surprising since medications and disease conditions are clinically related. However, the practical implication of this result is significant because it indicates that one of these two modalities is sufficient. To confirm this finding, models that use all three modalities are developed and the corresponding AUC are shown in Table 4.6.

Table 4.5 Intra-cohort evaluation: AUC and 95% confidence interval of test samples for models trained on sociodemographic variables (So) and ATC medication groups (Rx).

	Cases (all)	Controls (all)	Cases (65+)	Controls (65+)	Cases (ICD)
ED visits					
SVM	0.69[0.70,0.71]	0.75[0.73,0.77]	0.66[0.63,0.69]	0.68[0.63,0.70]	0.72[0.69,0.75]
RF	0.70[0.68,0.72]	0.77[0.75,0.78]	0.68[0.65,0.70]	0.71[0.67,0.74]	0.72[0.70,0.76]
NN	0.66[0.65,0.68]	0.75[0.73,0.77]	0.66[0.63,0.69]	0.67[0.62,0.70]	0.71[0.69,0.74]
LR	0.68[0.67,0.70]	0.74[0.72,0.76]	0.66[0.63,0.70]	0.68[0.64,0.71]	0.72[0.69,0.74]
XGB	0.71[0.69,0.73]	0.77[0.75,0.78]	0.68[0.66,0.71]	0.72[0.67,0.74]	0.73[0.71,0.76]
IH visits					
SVM	0.75[0.73,0.77]	0.68[0.65,0.70]	0.72[0.69,0.76]	0.65[0.61,0.68]	0.80[0.78,0.83]
RF	0.78[0.76,0.80]	0.74[0.73,0.76]	0.76[0.72,0.79]	0.70[0.67,0.74]	0.83[0.81,0.85]
NN	0.75[0.72,0.77]	0.70[0.67,0.72]	0.72[0.69,0.77]	0.65[0.61,0.70]	0.81[0.79,0.84]
LR	0.75[0.73,0.77]	0.67[0.65,0.70]	0.71[0.67,0.75]	0.66[0.62,0.70]	0.80[0.77,0.83]
XGB	0.78[0.76,0.80]	0.74[0.72,0.77]	0.74[0.70,0.78]	0.72[0.68,0.75]	0.83[0.81,0.86]

The first three columns of Table 4.6 represent the AUC of the (all) cases and control models and the (ICD) cases models that combine the So, Dx, and Rx variables. The age-specific models are omitted from this experiment because they were found in Tables 4.4 and 4.5 to not improve prediction for ED or IH visits. The results in the first three columns of Table 4.6 confirm that there is no clear benefit from including both the Dx and Rx variables when predicting ED or IH visits. In order to gain further insight into which exposure variables are most predictive when all three modalities are combined, the top 5 features of the (all) cases, (all) controls and (ICD) cases RF models are examined (Table 4.7). Most of the top variables are So variables. This suggests that while So variables alone have a low entropy (Table 4.3), they are important when combined with other clinical variables. Moreover, the history of previous ED and IH visits are important in all models. This observation aligns with previously reported findings[3][12]. That said, variables from the medication and disease groups still participate in the classification and some of the medication variables are among the top five (e.g., Analgesics and Psychoanaleptics).

Table 4.6 Intra-cohort and transfer testing evaluation: AUC and 95% confidence interval of test samples for models using sociodemographic (So), Elixhauser comorbidity conditions (Dx) and ATC medication groups (Rx) variables.

	Intra-Cohort			Transfer-testing	
	Cases (all)	Controls (all)	Cases (ICD)	Case (all)	Cases (ICD)
ED visits					
SVM	0.69[0.67,0.71]	0.75[0.73,0.77]	0.72[0.69,0.75]	0.68[0.66,0.70]	0.71[0.69,0.74]
RF	0.70[0.69,0.72]	0.77[0.75,0.78]	0.73[0.71,0.76]	0.69[0.67,0.71]	0.74[0.72,0.77]
NN	0.69[0.67,0.71]	0.71[0.69,0.73]	0.72[0.69,0.74]	0.68[0.66,0.70]	0.70[0.67,0.72]
LR	0.69[0.67,0.70]	0.74[0.72,0.76]	0.71[0.68,0.74]	0.68[0.66,0.70]	0.71[0.69,0.74]
XGB	0.71[0.69,0.73]	0.77[0.75,0.79]	0.73[0.71,0.76]	0.70[0.68,0.71]	0.74[0.71,0.77]
IH visits					
SVM	0.75[0.73,0.77]	0.68[0.65,0.70]	0.79[0.76,0.81]	0.71[0.69,0.73]	0.75[0.73,0.78]
RF	0.78[0.76,0.80]	0.74[0.72,0.77]	0.83[0.80,0.85]	0.77[0.75,0.79]	0.82[0.80,0.84]
NN	0.75[0.72,0.77]	0.71[0.69,0.73]	0.79[0.77,0.82]	0.76[0.74,0.78]	0.71[0.68,0.74]
LR	0.75[0.73,0.77]	0.68[0.65,0.70]	0.80[0.78,0.82]	0.74[0.71,0.76]	0.79[0.76,0.81]
XGB	0.78[0.76,0.80]	0.74[0.72,0.76]	0.83[0.81,0.85]	0.77[0.75,0.79]	0.81[0.78,0.84]

Two important observations related to the structure of the feature space were derived from tables 5 and 6. First, control models for ED visits have higher predictive power than cases models. Second, cases models for IH visits have better or similar performance than equivalent controls models. These observations can be confirmed using the transfer testing technique discussed in Section 2.1 . Basically, the models trained using controls are tested on cases. The last two columns of Table 4.6 show the result of these tests. They include the AUC for the ED visits prediction when the 2,870 (all) and the 1,471 (ICD) test cases are submitted to the ED controls models. Similarly, the lower half of the last two columns includes the AUC of the IH visits prediction for the 1,848 (all) and the 1,138 (ICD) test cases when processed by the IH controls models. To ensure that results are in fact comparable, the test cases used in the transfer testing are the same test patients used in the evaluation of the (all) and (ICD) cases models in Table 4.6 (intra-cohort testing).

Table 4.7 Top 5 variables identified in RF for the Intra-cohort models that use the sociodemographic (So), Elixhauser comorbidity conditions (Dx) and ATC medication groups (Rx) exposure variables.

ED visits	
Cases (all)	History of ED visits, Age, Deprivation index, Analgesics, History of IH visits
Controls (all)	History of ED visits, Age, Deprivation index, Race, Antibactibacterials for systemic use
Cases (ICD)	History of ED visits, Age, Deprivation index, Psychoanaleptics, History of IH visits
IH visits	
Cases (all)	History of IH visits, Age, Deprivation index, History of ED visits, Analgesics
Controls (all)	Age, History of IH visits, Deprivation index, History of ED visits, Insurance type History of IH visits, Age, History of ED visits, Deprivation index, Fluid and electrolyte disorders
Cases (ICD)	History of IH visits, Age, Deprivation index, History of ED visits, Analgesics

For all ML techniques under consideration, the AUCs of the (all) and (ICD) cases intra-cohort testing are higher or equal to the AUCs of the corresponding transfer testing (i.e., column 1 versus column 4 and column 3 versus column 5). However, the gain in AUC varies depending on the technique and the outcome. For the highest performing technique, XGB, only a difference of 1% in AUC is observed for the (all) and (ICD) cases ED visits between intra-cohort and transfer testing (column 1 versus column 4 and column 3 versus column 5, respectively). For the IH visits, the drop in performance starts to increase reaching more than 5% for some of the ML techniques. This may indicate that the IH outcome may benefit from a disease-specific model especially for (ICD) cases where a high AUC was obtained with the So and Dx data modalities (Table 4.4). In general, for each technique and outcome, it is recommended to implement transfer-testing in order to decide if the gain in performance justifies the deployment of a disease specific model.

4.2 Pretrained language model for ACU

As described in Section 3.2.2, the Med-BERT architecture is pre-trained with ICD diagnosis codes and other related variables. The model architecture follows the transformer structure with

six hidden layers and six attention heads. Moreover, as in Med-BERT, the embedding dimension was set to 192 and the feed-forward/filter size was set to 64. The maximum length allowed for the input sequence was set to 128 tokens where each token represents a disease code. The default BERT optimizer, Adam Weight decay optimizer, the recommended learning rate of 5e-5 and 100,000 epochs are used to pre-train the model.

In order to adapt the model to ACU risk prediction, different combinations of pre-training tasks were attempted. The two pre-training tasks already proposed by Med-BERT were considered. In addition, a new pre-training task which aligns better with ACU risk prediction is introduced. The results established using different combinations of pre-training tasks are presented next.

4.2.1 Masked LM + Prolonged LOS

First, the two pre-training tasks used in Med-BERT were considered. The threshold for LOS was however lowered from 7 to 3 in order to increase the size of the positive samples. Table 4.8 shows that the number of patients nearly doubled with a lower threshold. Moreover, with a larger dataset the performance of the two pre-training tasks improved. It should be noted that for this LM, the combination of insomnia cases and controls is used for pre-training.

Table 4.8 Number of patients and vocabulary size with different thresholds for prolonged LOS.

	LOS > 7	LOS > 3
Number of patients	13,000	25,000
Vocabulary size	1,072	1,072

Table 4.9 Evaluation of the LM models after pre-training with Masked LM and Prolonged LOS over a hold-out test dataset

		Threshold LOS > 7	Threshold LOS > 3
Task1 Masked LM	Accuracy	0.13	0.17
	AUC	0.59	0.78
Task 2 Prolonged LOS	Accuracy	0.69	0.72
	AUC	0.60	0.68

Table 4.9 shows the accuracy of the LM model which is pretrained with Masked LM and Prolonged LOS with different thresholds over a hold-out dataset for the two tasks. The results indicate that as mentioned above, a lower threshold increased the size of the training dataset and

improved both accuracy and AUC for the two tasks. Moreover, Table 4.9 shows that higher accuracy is obtained for the prolonged LOS compared to the masked LM.

4.2.2 Masked LM + Number of Encounters

To augment the data, we divided the disease code history for each patient into independent yearly segments. For each segment, the patient's number of encounters in the following year is predicted based on the disease code sequence during the observation period of the previous year. After data augmentation, 99,000,000 data samples were available (Table 4.10).

Table 4.10 Number of patients and vocabulary size for the two pre-training tasks Masked LM and Number of Encounters.

Number of patients	99,000,000
Vocabulary size	1,013

Table 4.11 Evaluation of the LM models after pre-training with Masked LM and Number of Encounters over a hold-out test dataset.

	Accuracy	AUC
Task1 Masked LM	0.29	0.90
Task 2 number of encounters (>10)	0.66	0.66

Table 4.11 shows the accuracy of the LM model which is pretrained with Masked LM and Number of Encounters over a hold-out dataset for the two tasks. As in the case of the previous LM, a higher accuracy is obtained for the Number of Encounters task compared to the masked LM. Moreover, compared to Table 4.9, Table 4.11 indicates that a larger training dataset is needed to increase accuracy for the Masked LM task.

4.2.3 IH risk prediction

Tables 4.9 and 4.11 show a significant improvement in accuracy (i.e., from 0.17 to 0.29) for the Masked LM task with data segmentation. Therefore, the LM model pretrained with

Masked LM and number of encounters was selected and fine-tuned for the downstream task of IH risk prediction. The LM of choice was evaluated with different embeddings as shown below:

- LM1: Dx code, and visit order sequence.
- LM2: Dx code, visit order sequence, age, sex, and race.
- LM3: Dx code, visit order sequence, count of prior IH visits, and count of prior ED visits.
- LM4: Dx code, Dx frequency, count of prior IH visits, and count of prior ED visits.

The ML models discussed in Section 4.1 are used as a baseline for comparison. All models are compared with respect to their performances when predicting next one year IH visits among insomnia patients (Table 4.12, Table 4.13). Table 4.12 compares the classifier fine-tuned over LM1 to the corresponding baseline ML models that only used the 31 Elixhauser's comorbidity categories as input variables. Table 4.13 compares the classifiers developed using LM2 through LM4 to the corresponding ML models that use the exposure variables derived from the Dx, socio-demographics and encounters history modalities. This provides for a better comparison since LM2 through LM3 also include embeddings from different data modalities.

Table 4.12 AUC over a hold-out test dataset for IH risk prediction models derived from only using the Dx modality.

	Cases (all)	Cases (ICD)
Med-BERT_LM1	**0.66**	**0.70**
SVM	0.59	0.65
RF	0.62	0.68
NN	0.55	0.61
LR	0.54	0.61
XGB	0.56	0.60

Table 4.12 shows that the classifier fine-tuned from LM1 has higher AUC compared to the baseline ML models for all insomnia cases as well as insomnia patients with documented diagnosis.

Table 4.13 AUC over a hold-out test dataset for IH risk prediction models derived from multiple modalities.

	Cases (all)	Cases (ICD)
Med-BERT_LM2	**0.65**	**0.69**
Med-BERT_LM3	**0.74**	**0.77**
Med-BERT_LM4	**0.75**	**0.77**
SVM	0.71	0.75
RF	0.77	0.82
NN	0.76	0.71
LR	0.74	0.79
XGB	0.77	0.81

Compared to Table 4.12, the results of Table 4.13 show that higher performance is obtained with additional modalities. The results for the classifier finetuned over LM2 compared to the one finetuned over LM1 indicate that adding the socio-demographic embedding has no added benefits. However, adding the embedding that encodes the encounter history (LM3) can lead to a 12% improvement in AUC. A similar increase in performance is observed with LM4 which uses Dx frequency embedding instead of the visit order embedding. This aspect of the result is important since LM4 uses a reduced-size embedding and limited computational resources compared to LM3.

CHAPTER 5. DISCUSSION

Access to ML models that are developed from EHR routine care data can help improve health services. However, efficient modeling using EHR data, even for structured data modalities, is difficult because of missing data and inadequate representation of the general patient population in the training and validation datasets[4]. This complexity is compounded by health institution-specific data distributions and patient cohort segmentation along disease conditions or age groups. For ML models to be adopted by health care systems, they must be widely applicable and carry minimum burden. These design considerations are often overlooked. In this study, modeling ACU is used to demonstrate the importance of feature selection, cohort segmentation, and model architecture design decisions when developing ML models for practical use.

First, it is critical to select a limited number of widely available exposure variables. Standard demographic variables include age, sex, race and ethnicity. These variables have limited inconsistencies and are widely available. However, augmenting demographic variables with other social determinants must be done with care. In this study, previous ED and IH visits were included because they were shown to be strong predictors of future need for health services[3]. Social determinants such as insurance or a deprivation index must be appropriately stratified for ease of transfer to other health care institutions. For instance, the stratification of the insurance type in the present study followed government, commercial, self-pay and other/unknown. An alternative insurance stratification is public versus private[3]. The Elixhauser's commorbidity score is also generalizable since it is calculated using ICD diagnosis codes. Using variables such as zip codes or mode of transportation as social determinants[10] may limit the generalizability of the model.

In contradiction with previous findings[13], the present study suggests that the combination of Socio-demographic variables and either Rx or Dx variables is best. Using all three modalities is not necessary. Choosing between Rx and Dx variables can have an impact on the generalizability of the model to other health care institutions. Diagnosis conditions (Dx) are often documented using ICD codes. Exposure variables derived from ICD codes are likely to be interoperable across health care systems. Moreover, diagnosis conditions are fewer, likely to be persistent and

consistent from one health care provider to another. Finally, the diagnosis codes are aggregated by disease groups which further reduces potential for variability.

In contrast, Rx variables may not seamlessly generalize for several reasons. Treatment plans may differ from one health care provider to another even for the same disease condition. Documentation of patients' medications in the EHR of health care institutions may rely on different classification taxonomies (e.g., GPI, ATC and UPS). Translating from one classification to another is also not trivial and requires substantial manual efforts and guidance from subject matter experts because one-to-one mappings for medication groups in one taxonomy to another taxonomy is not possible. Moreover, the Rx modality evolves as new drugs are introduced and old drugs are discontinued or modified. Therefore, even if a common medication taxonomy is possible, models derived from Rx exposure variables have to be updated on a regular basis. Finally, due to the movement of patients from one health care provider to another, the history of medications compared to the history of diagnosis is more likely to be incomplete in the EHR thereby leading to classification errors. That said, if Rx variables are omitted, some of the acute care utilization directly related to specific medications may be missed.

There are other modalities that have not been considered in the present study. These include medical notes, procedures and laboratory results. Medical notes have been used to predict patients at high risk for ACU[14][37]. However, the transfer of this modality to other patient populations remains an open research question. Laboratory results may not be widely available as in the case of polysomnography data for insomnia patients[1].

Second, the present study indicates that:

- An insomnia-specific model for predicting ED visits is not recommended. In fact, the general patient model is sufficiently accurate for the insomnia test patients.
- Predicting IH visits can benefit from a disease-specific model especially for severe cases of insomnia (i.e., patients that will eventually progress to a documented diagnosis).
- An age-specific model for older adults does not have better performance than the general patient model for predicting either ED or IH visits.

These findings can guide the development of ACU models, general patient models are more practical, cohort segmentation can limit the utility of the models and reduce the data available for their development. Previously, disease-specific prediction models for hospital re-admission were found to outperform general patient models[3]. This finding is consistent with the insomnia-specific model for IH visits presented in this study. However, for ED visits this may not be applicable.

The need for cohort segmentation should be evaluated for every model. As demonstrated in the present study, this can be achieved by comparing the cohort-specific model to an equivalent general patient model. In addition, transfer testing between the two models should be performed to confirm that the predictive performance is improved with segmentation. In the current study, transfer testing was able to confirm that disease-specific segmentation was only beneficial for IH models.

This study shows that it is possible to leverage the generalizability of general patient population models and the specificity of disease-specific models by using pre-trained language models. Moreover, the ability of these language models was demonstrated for the task of predicting IH visits for insomnia patients. The idea is to develop a general language model that can then be fine-tuned for specific cohorts according to the needs of different healthcare institutions. In this study pre-training a language model for the general population and fine-tuning it for a disease-specific cohort was shown to have similar performance to traditional disease-specific machine learning models. Better performance can be achieved with sufficient data for pre-training.

Finally, this study investigated the inclusion of socio-demographic variables (e.g., age), which have been shown to be important in traditional ML models, as an embedding in the language model. However, the results indicate that there was little improvement because of this inclusion. Future research is needed to understand if this lack of performance is due to the embedding representation of the socio-demographics variables or to the way data from this modality is fused with other modalities.

CHAPTER 6. CONCLUSION

Understanding the deployment constraints of ML models is necessary for their adoption by health care providers. These models can enhance disease diagnosis and enable the systematic classification of disease risks and health care services utilization. However, the design and engineering of these models must take into consideration their added burdens on the already overburdened health care system. This burden can increase with the inclusion of unnecessary and hard to obtain exposure variables and with cohort segmentation. In the present study, these design considerations are illustrated by comparing an insomnia patient cohort to a general patient cohort. The focus of the study is predicting future ED and IH visits for these patients.

Our results indicate that it is not necessary to include both diagnosis and medication history variables when predicting future ED and IH visits. Disease- and age-specific models for predicting future ED visits are also not needed. In contrast, the IH disease-specific models outperform the general patient models. The broader question of whether disease-specific models are necessary remains open. However, this study identified at least a scenario where it is not the case. The utility of any ML model developed for a segment of the population should not be proposed in isolation. We recommend that its predictive performance be compared to the general patient equivalent model. Transfer testing must also be performed to confirm that the cohort-specific model outperforms the general patient model.

It may be possible to avoid the high maintenance of cohort-specific models by developing a language model for the general population that can be fine-tuned for specific outcomes and patient cohorts. A preliminary investigation of this approach was performed in this study. An LM based on the Med-BERT architecture and pre-trained on the Masked LM and next year encounters tasks was developed. This LM was then fine-tuned to predict the risk for future IH visits and was shown to have comparable performance to baseline models developed using traditional ML techniques. Potential for improvement of the LM is also possible with additional training data. Future work will extend the present study to other patient populations, disease conditions and health care services in order to further ascertain the utility of the pre-trained LM approach.

REFERENCES

[1] Kim, S. S. (2021). Recent Trends of Artificial Intelligence and Machine Learning for Insomnia Research. Chronobiology in Medicine, 3(1), 16-19.

[2] Ahuja, R., Vivek, V., Chandna, M., Virmani, S., & Banga, A. (2022). Comparative study of various machine learning algorithms for prediction of Insomnia. In Research Anthology on Machine Learning Techniques, Methods, and Applications (pp. 776-799). IGI Global.

[3] Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. Journal of biomedical informatics, 56, 229-238.

[4] Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2016). Deep learning for health informatics. IEEE journal of biomedical and health informatics, 21(1), 4-21.

[5] Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., ... & Pattichis, C. S. (2020). AI in medical imaging informatics: current challenges and future directions. IEEE journal of biomedical and health informatics, 24(7), 1837-1857.

[6] Phan, D. V., Yang, N. P., Kuo, C. Y., & Chan, C. L. (2021). Deep learning approaches for sleep disorder prediction in an asthma cohort. Journal of Asthma, 58(7), 903-911.

[7] Goltz, D. E., Ryan, S. P., Howell, C. B., Attarian, D., Bolognesi, M. P., & Seyler, T. M. (2019). A weighted index of Elixhauser comorbidities for predicting 90-day readmission after total joint arthroplasty. The Journal of arthroplasty, 34(5), 857-864.

[8] Khatri, K. L., & Tamil, L. S. (2017). Early detection of peak demand days of chronic respiratory diseases emergency department visits using artificial neural networks. IEEE journal of biomedical and health informatics, 22(1), 285-290.

[9] Shojaei, E., Wong, A., Rexachs, D., Epelde, F., & Luque, E. (2020). A method for projections of the emergency department behaviour by non-communicable diseases from 2019 to 2039. IEEE journal of biomedical and health informatics, 24(9), 2490-2498.

[10] Parker, C. A., Liu, N., Wu, S. X., Shen, Y., Lam, S. S. W., & Ong, M. E. H. (2019). Predicting hospital admission at the emergency department triage: A novel prediction model. The American journal of emergency medicine, 37(8), 1498-1504.

[11] Luo, G., Stone, B. L., Nkoy, F. L., He, S., & Johnson, M. D. (2019). Predicting appropriate hospital admission of emergency department patients with bronchiolitis: secondary analysis. JMIR medical informatics, 7(1), e12591.

[12] McCusker, J., Karp, I., Cardin, S., Durand, P., & Morin, J. (2003). Determinants of emergency department visits by older adults: a systematic review. Academic Emergency Medicine, 10(12), 1362-1370.

[13] Lemke, K. W., Weiner, J. P., & Clark, J. M. (2012). Development and validation of a model for predicting inpatient hospitalization. Medical care, 131-139.

[14] Frost, D. W., Vembu, S., Wang, J., Tu, K., Morris, Q., & Abrams, H. B. (2017). Using the electronic medical record to identify patients at high risk for frequent emergency department visits and high system costs. The American journal of medicine, 130(5), 601-e17.

[15] Frost, D. M., Vembu, S., Wang, J., Tu, K., Morris, Q., & Abrams, H. (2017a). Using the Electronic Medical Record to Identify Patients at High Risk for Frequent Emergency Department Visits and High System Costs. The American Journal of Medicine, 130(5), 601.e17-601.e22. https://doi.org/10.1016/j.amjmed.2016.12.008

[16] Ford, E. S., Cunningham, T. J., Giles, W. H., & Croft, J. B. (2015). Trends in insomnia and excessive daytime sleepiness among US adults from 2002 to 2012. Sleep medicine, 16(3), 372-378.

[17] LeBlanc, M., Mérette, C., Savard, J., Ivers, H., Baillargeon, L., & Morin, C. M. (2009). Incidence and risk factors of insomnia in a population-based sample. Sleep, 32(8), 1027-1037.

[18] Wickwire, E. M., Tom, S. E., Scharf, S. M., Vadlamani, A., Bulatao, I. G., & Albrecht, J. S. (2019). Untreated insomnia increases all-cause health care utilization and costs among Medicare beneficiaries. Sleep, 42(4), zsz007.

[19] Edition, F. (2013). Diagnostic and statistical manual of mental disorders. Am Psychiatric Assoc, 21(21), 591-643.

[20] Laugsand, L. E., Strand, L. B., Vatten, L. J., Janszky, I., & Bjørngaard, J. H. (2014). Insomnia symptoms and risk for unintentional fatal injuries—the HUNT study. Sleep, 37(11), 1777-1786.

[21] Winkelman, J. W. (2015). Insomnia disorder. New England Journal of Medicine, 373(15), 1437-1444.

[22] Chakravorty, S., Chaudhary, N. S., & Brower, K. J. (2016). Alcohol dependence and its relationship with insomnia and other sleep disorders. Alcoholism: Clinical and Experimental Research, 40(11), 2271-2282.

[23] Khachatryan, S. G. (2021). Insomnia burden and future perspectives. Sleep Medicine Clinics, 16(3), 513-521.

[24] Schutte-Rodin, S., Broch, L., Buysse, D., Dorsey, C., & Sateia, M. (2008). Clinical guideline for the evaluation and management of chronic insomnia in adults. Journal of clinical sleep medicine, 4(5), 487-504.

[25] World Health Organization. (1978). International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index. World Health Organization.

[26] Singh, G. K. (2003). Area deprivation and widening inequalities in US mortality, 1969–1998. American journal of public health, 93(7), 1137-1143.

[27] Elixhauser, A., Steiner, C., Harris, D. R., & Coffey, R. M. (1998). Comorbidity measures for use with administrative data. Medical care, 8-27.

[28] Menendez, M. E., Neuhaus, V., van Dijk, C. N., & Ring, D. (2014). The Elixhauser comorbidity method outperforms the Charlson index in predicting inpatient death after orthopaedic surgery. Clinical Orthopaedics and Related Research®, 472, 2878-2886.

[29] Tang, P. L., Lin, H. S., & Hsu, C. J. (2021). Predicting in-hospital mortality for dementia patients after hip fracture surgery–A comparison between the Charlson Comorbidity Index (CCI) and the Elixhauser Comorbidity Index. Journal of Orthopaedic Science, 26(3), 396-402.

[30] Li, P., Kim, M. M., & Doshi, J. A. (2010). Comparison of the performance of the CMS Hierarchical Condition Category (CMS-HCC) risk adjuster with the Charlson and Elixhauser comorbidity measures in predicting mortality. BMC health services research, 10(1), 1-10.

[31] Charlson, M. E., Pompei, P., Ales, K. L., & MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of chronic diseases, 40(5), 373-383.

[32] Klabunde, C. N., Potosky, A. L., Legler, J. M., & Warren, J. L. (2000). Development of a comorbidity index using physician claims data. Journal of clinical epidemiology, 53(12), 1258-1267.

[33] World Health Organization. (1996). Guidelines for ATC classification and DDD assignment.

[34] Figueredo, A. J., & Wolf, P. S. A. (2009). Assortative Pairing and Life History Strategy. Human Nature, 20(3), 317–330. https://doi.org/10.1007/s12110-009-9068-2

[35] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ... & He, Q. (2020). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1), 43-76.

[36] Khan, N. M., Abraham, N., & Hon, M. (2019). Transfer learning with intelligent training data selection for prediction of Alzheimer's disease. IEEE Access, 7, 72726-72735.

[37] Topaz, M., Woo, K., Ryvicker, M., Zolnoori, M., & Cato, K. (2020). Home Health Care Clinical Notes Predict Patient Hospitalization and Emergency Department Visits. Nursing research, 69(6), 448.

[38] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports, 6(1), 1-10.

[39] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[40] Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine, 4(1), 86.

[41] Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., ... & Salimi-Khorshidi, G. (2020). BEHRT: transformer for electronic health records. Scientific reports, 10(1), 1-12.

[42] Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association, 24(2), 361-370.

[43] Choi, E., Bahadori, M. T., Sun, J., Kulas, J. A., & Schuetz, A. (2018). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In Advances in neural information processing systems (pp. 3504-3513).

[44] Liu, J., Huang, J., Zhang, T., & Guo, Y. (2019). Support vector machines for cancer classification using multiple gene expression profiles. Journal of biomedical informatics, 96, 103244.

[45] Moradi, P., Bayat, M., GholamHosseini, H., & Mohagheghian, E. (2021). A novel framework for multi-class classification of gene expression data based on support vector machines. Journal of biomedical informatics, 119, 103812.

[46] Zolfaghar, K., Mansournia, M. A., Ashrafi-Asgarabad, A., Bagherzadeh-Khiabani, F., & Sadeghi, R. (2018). Predicting outcomes of patients with heart failure using random forest and cox regression models. Iranian Journal of Public Health, 47(2), 231-239.

[47] Sankaranarayanan, J., Lulu, V. S., & Sathyabama, V. (2020). Exploring factors associated with medication non-adherence among patients with heart failure: a random forest analysis. Journal of Medical Systems, 44(7), 135.

[48] Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Esteva, A. (2018). Scalable and accurate deep learning with electronic health records. NPJ digital medicine, 1(1), 1-10.

[49] Chen, L., Wang, Y., He, T., & Liu, G. (2020). A two-stage convolutional neural network for Alzheimer's disease prediction using hippocampus segmentations. Frontiers in neuroscience, 14, 28.

[50] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling.

[51] Choi, E., Bahadori, M., Sun, J., Kulas, J., Schuetz, A., & Stewart, WalterF. (2015). RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism.

[52] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

[53] Jurafsky, D. (2000). Speech & Language Processing. Pearson Education India.

[54] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific Reports, 6(1). https://doi.org/10.1038/srep26094.

[55] Meng, Y., Speier, W., Ong, M. K., & Arnold, C. W. (2021). Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. IEEE journal of biomedical and health informatics, 25(8), 3121-3129.

[56] Choi, E., Bahadori, M., Schuetz, A., Stewart, WalterF., & Sun, J. (2015). Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. arXiv: Learning.

[57] Rupp, M., Peter, O., & Pattipaka, T. (2023). ExBEHRT: Extended Transformer for Electronic Health Records to Predict Disease Subtypes & Progressions. arXiv preprint arXiv:2303.12364.

[58] Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016, December). Doctor ai: Predicting clinical events via recurrent neural networks. In Machine learning for healthcare conference (pp. 301-318). PMLR.

[59] Saitwal, H., Qing, D., Jones, S., Bernstam, E. V., Chute, C. G., & Johnson, T. R. (2012). Cross-terminology mapping challenges: a demonstration using medication terminological systems. Journal of biomedical informatics, 45(4), 613-625.

PUBLICATION

Chekani, F., Zhu, Z., Khandker, R. K., Ai, J., Meng, W., Holler, E., Dexter, P., Boustani, M. and Ben Miled, Z. (2023). Modeling acute care utilization: practical implications for insomnia patients. Scientific Reports, 13(1), 2185.