HUMAN-CENTERED AI FOR PRECISION MEDICINE: METHODS AND APPLICATIONS

A Dissertation

by

SIRUI DING

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Xia Hu
Committee Members,	Xiaoning Qian
	Theodora Chaspari
	Ruihong Huang
Head of Department,	Scott Schaefer

May 2024

Major Subject: Computer Science

ABSTRACT

AI has emerged as a powerful tool in the healthcare and biomedical domains. In the field of medicine, AI must demonstrate strong performance while adhering to human ethics. Throughout my Ph.D., I focused on designing and developing human-centered AI tools for precision medicine, with a special emphasis on addressing ethical concerns in medical AI. To bridge the gap between AI and medicine, I delved into cutting-edge AI methods, including knowledge distillation, reinforcement learning, multi-task learning, multi-modality learning and contrastive learning. To make our contribution focused, I specialized in the phenotyping, disease diagnosis, organ transplant, and health event prediction scenario which are essential medical tasks. Four-folds challenges of designing human-centered AI framework towards precision medicine are summarized as (1).Trade-off between performance and fairness, (2).AI integration in clinical workflow, (3).Multi-task prediction on related medical indicators. (4).Multi-modality EHR data. To comprehensively investigate the fairness issue in the clinical prediction algorithm, I conduct extensive experiments on the disease diagnosis to benchmark the performance and bias in the electronic phenotyping. I design a two-step debiasing strategy with unbiased knowledge distillation to predict the graft failure after liver transplant fairly and precisely. In order to support the doctor's clinical decision, FairAlloc framework is proposed to directly generate accurate and unbiased patient prioritization decisions with reinforcement learning. To simultaneously predict highly related medical indicators, CoD-MTL is designed to take advantage of multiple highly related tasks to predict multiple cause-of-death after the liver transplant. When it comes to multimodal EHR data, I design a cross-modality knowledge distillation framework to distill the knowledge from LLM into the predictive model on structured EHR. My research efforts paved a way to design powerful and trustworthy AI frameworks to support precision medicine with human-centered AI principles.

DEDICATION

To my family.

ACKNOWLEDGMENTS

I would express my deep gratitude to the following great people who contributed to my academic and personal growth, whether through their guidance, encouragement, or friendship.

I would like to express my gratitude to my academic advisor, Dr. Xia (Ben) Hu, for his unwavering support, guidance, and encouragement throughout my doctoral journey. His expertise, mentorship, and invaluable feedback have been instrumental in shaping the direction of my research and fostering my intellectual growth. I am also equally grateful to my dissertation committee members, Dr. Xiaoning Qian, Dr. Theodora Chaspari, Dr. Ruihong Huang, for their insightful feedback on my dissertation and kind support during my PhD.

I would like to extend my appreciation to my close collaborators during the PhD journey. They are Dr. Na Zou, Dr. Xiaoqian Jiang, Dr. Yafen Liang and Dr. Kai Zhang, whose expertise, support, and collaboration have made my academic journey more fulfilling and rewarding. Also, I am particularly thankful for my colleagues in the lab for their friendship, encouragement, and willingness to lend a helping hand whenever needed. Their positive energy and enthusiasm have been a constant source of motivation, especially during challenging times. I am grateful for the countless discussions, brainstorming sessions, and problem-solving moments we shared together. Each interaction has been a source of inspiration and has played a crucial role in shaping my research ideas and methodologies.

I would like to express my deepest appreciation to my beloved parents Peifang Wu and Gang Ding for their unwavering love, encouragement, and understanding throughout my doctoral studies. Their unconditional support and sacrifices have been the cornerstone of my academic success, and I am forever grateful for their belief in me.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by dissertation committee consisting of Professor Xia Hu (Chair and advisor), Professor Theodora Chaspari and Professor Ruihong Huang from Department of Computer Science and Engineering and Professor Xiaoning Qian from Department of Electrical and Computer Engineering.

All the other parts of this dissertation were completed by the student independently.

Funding Sources

This work is, in part, supported by National Science Foundation (IIS-1939716, IIS-1900990 and IIS 2239257) and National Institutes of Health (1OT2OD032581-02-211). The views, opinions, and/or findings expressed in this work solely belong to the author(s) and should not be construed as representing the official views or policies of the National Institutes of Health and the U.S. Government.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Background and Related Work

Precision medicine [1] has become the focus of the biomedicine field, with more and more research and clinical efforts being invested in this area. The core of precision medicine is precisely identifying, diagnosing, treating each individual patient with their personal genetics, symptom and environment in consideration. With the development of techniques to record the healthcare and biomedicine data, large scale medical data becomes accessible, e.g., electronic health records (EHR) [2]. Artificial Intelligence (AI) has shown its promising performance when trained on large scale data and has non-trivial applications in the medical domain with its transformative power in prediction and analysis [3]. For example, surgical robots can assist the doctor with the surgery [4], AI powered medical chatbot can serve as the assistant for patients [5], etc. Before grounding the AI in medicine tasks, we need to make sure the medical AI system performs towards human good. However the ethical issue e.g., fairness [6], robustness [7, 8], etc is less investigated in medical AI compared to model performance like accuracy, precision, etc. So we need to develop human-centered AI (HAI) framework to support the precision medicine. We summarize the related works as follows.

Human-centered AI: Human-centered AI aims to design AI system that does human good and benefit the human beings [9]. The AI model needs to be unbiased, robust, and interpretable for human to apply it. In this dissertation, we will focus on the bias issue of AI model from the human perspective. The fairness of the AI requires the model outcome to be unbiased across different subgroups [10]. There are many initial efforts put into the fair AI from the ML community. For example, there are adversarial debiasing strategy [11], fairness regularization [12], etc. However, the fairness needs to be specially defined in many medical tasks and the debiasing strategies needs to be specially tailored for different medical scenarios [13].

Precision Medicine with AI: Precision medicine can be boosted by AI with its strong power in

the prediction and generation [14]. For example, AI can provide precise diagnosis of diseases [15], identify the phenotype from large group of patients [16], recommend the treatment for patients [17], etc. The AI system in medicine can support the doctors clinical decision rather than replacing them [18]. With more precise information and predictions on the hand, the doctor will make more informed decisions. However, direct usage of AI to support the precision medicine may cause some issues like bias, untrustworthy, etc [19]. Thus we need to take the consideration of ethical issue of AI in precision medicine [20].

To guide the development of medical HAI, we summarize three aspects corresponding to patients, doctors and developers respectively. (1). How to make sure the AI prediction is fair? (2). How does AI generate precise and fair ranking? (3). How to let AI take multi-modality data and perform multi-task like human? I put my research efforts around these three research questions with an emphasis on a wide range of medical applications including disease diagnosis, phenotyping, organ transplant, health event prediction, clinical decision support, etc.

1.2 Motivations and Challenges

For the first research question, we investigate the fairness and bias issue in the model prediction. The AI models could tend to be biased towards some underrepresented groups. How to make sure the predictions is fair and unbiased is our goal. However, there is a trade-off between the performance and fairness when we mitigate the bias of model predictions. So, we summarize the challenge (1) as: How to make a trade-off between performance and fairness. Because the performance and fairness are both very essential and the balance between them is a challenging task [21]. For the second research question, we find that the direct usage of AI predicted score in the clinical decision may incur the bias, even when the score is predicted with precision and fairness. The challenge (2) can be summarized as: How to generate precise and unbiased clinical decisions. From the prelim experiments, we show that the direct use of AI outputs in the organ transplant allocation will lead to the bias in patient prioritization. How to integrate the AI in the current clinical workflow to support the doctors clinical decision remains as a challenge [22]. For the third research question, we discuss the design of medical AI framework that is inspired

by human expertise. The human doctors will refer to multiple medical records with different modalities, e.g., radiology image, structured EHR, biomedical signal, etc [23]. Meanwhile, the doctor will usually have multiple tasks to do simultaneously, e.g., multiple potential diseases to diagnosis, treatment plan recommendation, etc [24]. Thus, we are inspired to design multi-task medical AI frameworks on multi-modality medical data. The challenge (3) is: How to leverage the inner and natural relations between different medical indicators. For example, the rejection and infection after the transplant are highly related [25], how to take advantage of this relation and precisely predict multiple indicators at the same time. The challenge (4) can be summarized as: How to learn from the multi-modality medical data. To tackle these 4 challenges, I have the contributions as follows.

1.3 Contributions

This dissertation aims to develop human-centered AI framework that can effectively support the precision medicine. We design novel methods to help the AI framework better collaborate the doctors in decision making and guarantee the precise as well as ethical outputs of AI model. As the prelim works, we conduct extensive experiments to benchmark the bias in health predictive models via the electronic phenotyping task. This benchmark comprehensively shows the non-trivial bias existing in the medical AI models. For the challenge (1), we design a two-step debiasing framework to precisely and fairly predict the medical outcomes in the context of organ transplant. The fair-aware knowledge distillation is elaborated to improve the prediction precision as well as constrain the bias in outputs. The bias exists in both knowledge distillation and end-to-end training stage, for which we should both apply the fairness regularization. For the challenge (2), we further illustrate the fairness in prediction score is not identical to the fairness in ranking, which is a very important task in the clinical decision. We propose a policy generation framework with reinforcement learning, which can directly generate precise and unbiased ranking policy to prioritize the patients for organ transplant. The fairness in ranking problem is defined from the computational perspective. In this work, we will consider both group and individual level fairness. The reinforcement learning can directly optimize the utility goal and the fairness goal which are non-differential

objectives. For the challenge (3), we design a multi-task learning framework to distill multiple tree-based model into the neural networks to learn the inner relations between post-transplant outcomes and predict them simultaneously. This framework take advantage of the superiority of tree model on tabular data and distill them into the heads part of multi-task learning framework. For the challenge (4), we design a framework to learn from multi-modality EHR data. To efficiently and effectively take advantage of large language model (LLM), we propose a cross-modality distillation framework to distill the knowledge from LLM into the predictive Transformer on structured EHR data. Meanwhile, we also model the patient similarity with contrastive learning. So the multi-modaltiy and patient similarity can be learnt within the same framework.

In summary, our contribution have five folds:

- A comprehensive benchmark to identify and mitigate the bias in electronic phenotyping. We evaluate the bias of baseline ML models on the phenotyping task and the effectiveness of commonly used debiasing strategies.
- The precise and fair medical outcome predictions achieved by a two-step debiasing framework and fair-ware knowledge distillation strategy. This framework is validated on the post-transplant outcome analysis in liver transplant.
- A clinical decision support framework that aims to optimize both utility and fairness goal with reinforcement learning. The results shows the generated policy can significantly reduce the bias in decision making.
- The novel tree-distilled multitask learning framework that can learn multiple related medical tasks on EHR data. The results further shows the power of AI in personalized medicine.
- A cross-modality framework is proposed to distill the knowledge from large language model (LLM) to boost the performance of predictive model on structured EHR data. We also model the patient similarity with the contrastive loss to further improve the performance on health event prediction.

1.4 Overview

The dissertation is organized as follows:

- **Chapter 2**: We will introduce the importance of electronic phenotyping and the bias in current electronic phenotyping methods. The debiasing strategies are evaluated on different baseline ML models for phenotyping.
- **Chapter 3**: We will introduce how to design a precise and fair predicting model. A two-step debiasing strategy and tree distillation method will be described with more details.
- **Chapter 4**: We will introduce the FairAlloc, a pipeline to directly generate precise and fair organ allocation policies by ranking the patients on waiting list.
- **Chapter 5**: We will introduce tree-based multitask learning framework that can process multiple tasks simultaneously. This chapter will also introduce a case study on the personalized treatment example.
- **Chapter 6**: We will introduce the CKLE framework, which distill the cross-modality knowledge from LLM and learn the patient similarity with contrastive loss. The effectiveness of CKLE is validated on cardiovascular health event predictions.
- **Chapter 7**: We summarize the dissertation by providing key insights and our contribution in this line of research. Meanwhile, we provide several promising directions in the future.

2. MEDICAL HAI WITH ELECTRONIC PHENOTYPING BIAS BENCHMARK

2.1 Overview

Electronic phenotyping is a fundamental task that identifies the special group of patients, which plays an important role in precision medicine in the era of digital health. Phenotyping provides real-world evidence for other related biomedical research and clinical tasks, e.g., disease diagnosis, drug development, and clinical trials, etc. With the development of electronic health records, the performance of electronic phenotyping has been significantly boosted by advanced machine learning techniques. In the healthcare domain, precision and fairness are both essential aspects that should be taken into consideration. However, most related efforts are put into designing phenotyping models with higher accuracy. Few attention is put on the fairness perspective of phenotyping. The neglection of bias in phenotyping leads to subgroups of patients being underrepresented which will further affect the following healthcare activities such as patient recruitment in clinical trials. In this work, we are motivated to bridge this gap through a comprehensive experimental study to identify the bias existing in electronic phenotyping models and evaluate the widely-used debiasing methods' performance on these models. We choose pneumonia and sepsis as our phenotyping target diseases. We benchmark 9 kinds of electronic phenotyping methods spanning from rule-based to data-driven methods. Meanwhile, we evaluate the performance of the 5 bias mitigation strategies covering pre-processing, in-processing, and post-processing. Through the extensive experiments, we summarize several insightful findings from the bias identified in the phenotyping and key points of the bias mitigation strategies in phenotyping.

2.2 Introduction

Phenotyping stands as a cornerstone in the realm of biomedical research, serving as the linchpin that enables medical practitioners to accurately pinpoint diseases [26, 27], facilitates the acceleration of drug development [28], and plays a pivotal role in the meticulous design of clinical trials [29]. Its foundational significance reverberates throughout the entire healthcare ecosystem,

fundamentally shaping the trajectory of patient care, research advancements, and medical innovation as illustrated in Figure 2.1.

Riding the wave of progress in electronic health records within the biomedical domain [30], the landscape of phenotyping has undergone a remarkable transformation, driven by the integration of cutting-edge computational methodologies, including advanced statistical analyses and artificial intelligence techniques [31]. As a result, electronic phenotyping methods have consistently demonstrated their prowess, exhibiting exceptional precision and efficiency across a multitude of scenarios [32]. This evolution heralds a new era in healthcare, one where data-driven insights are poised to revolutionize medical diagnosis and treatment.

However, bias is an inevitable factor in computational-based phenotyping methods, and its implications extend to various biomedical activities, including clinical trial design [29]. In this work, we will focus on the group bias which indicates the bias between different patient subgroups to make our contribution focused. For instance, when minority groups are underrepresented in the phenotyping process, this bias carries over to clinical trials during patient recruitment [33]. Addressing bias in electronic phenotyping poses a dual challenge for several reasons. Firstly, identifying bias from a computational standpoint is complex, as it often originates from two primary sources: data bias and model bias. Secondly, mitigating bias in electronic phenotyping and selecting appropriate debiasing techniques for different phenotype applications require careful consideration.

We are, therefore, highly motivated to embark on an extensive investigation aimed at identifying and mitigating biases within the realm of electronic phenotyping. This comprehensive study will encompass a meticulous review of prevalent electronic phenotyping methodologies, diligently scrutinizing the inherent biases within these approaches. Subsequently, we will delve into the computational aspects of bias identification and mitigation. Our research will encompass practical experiments designed to assess both the prevailing biases within existing electronic phenotyping algorithms and the efficacy of widely employed debiasing techniques. In addition, we present pivotal insights derived from our extensive experimentation. These findings encapsulate valuable

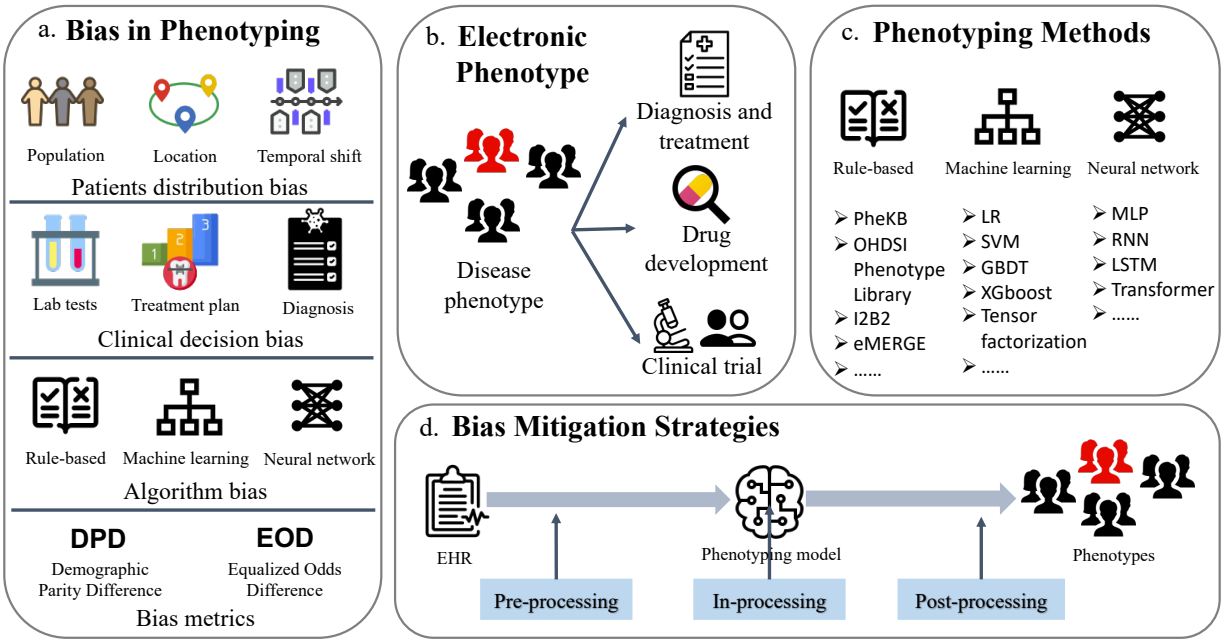


Figure 2.1: Overview of the identification and mitigation of the bias in electronic phenotyping.

knowledge and discoveries that shed light on the intricacies of electronic phenotyping and bias mitigation. By undertaking this multifaceted examination, we aim to pave the way for more equitable and unbiased electronic phenotyping practices. The contributions of this work can be succinctly summarized in three key aspects as follows:

- We benchmark and analyze the bias of 9 commonly used phenotyping models from the computational perspective.
- We evaluate 5 machine learning-based debiasing strategies for the phenotyping models. We analyze the advantages and disadvantages of each category debiasing strategy.
- We conduct extensive experiments to identify and mitigate the bias on pneumonia and sepsis phenotyping tasks and summarize insightful key findings from the experimental results.

2.3 Background

In this section, we embark on a comprehensive exploration of prevalent electronic phenotyping methods, classifying them for clarity and context. Subsequently, we delve into an insightful dis-

cussion on the latent biases that can emerge within these methods, dissecting them from both the data and model perspectives.

2.3.1 Electronic Phenotyping Methods

We categorize the electronic phenotyping methods into 4 categories, which are rule-based, traditional machine learning, neural network, and tensor factorization as shown in Figure 2.1(c).

Rule-based Method: Rule-based method is one of the most fundamental and widely applied phenotyping techniques [34, 35]. The core idea is to heuristically identify the phenotypes from electronic health records by the expert-defined rules. The widely adopted rule-based methods benefit from the characteristics of interpretability, simplicity, and ease of implementation. PheKB [34] is a public rule-based phenotyping algorithm that is widely used. However, due to the human-defined rules are usually limited to some specific scenarios, they are hard to adapt to different disease or patient distributions. For example, Kho at al. [36] designed a phenotyping method specialized for type II diabetes.

Traditional Machine Learning: The main kinds of traditional machine learning include logistic regression (LR) [37], tree-based methods [38, 39], and SVM [40]. These methods don't require large amounts of data in the training stage. Feature engineering [41] is an essential step for these methods to achieve competitive performance which will also require domain expertise. While traditional machine learning has found applications in various disease phenotyping tasks [42, 43], Li et al. introduced Xrare [44], which leverages Gradient Boosting Decision Trees (GBDT) for diagnosing rare diseases from genetics and phenotypic data. Nonetheless, these methods still have limitations that impact their performance and adaptability. For instance, SVM is tailored for binary classification, rendering it impractical for multiclass phenotyping. LR is sensitive to outlier data [45], which is very common in EHR [46], and the tree-based model needs laborious hyperparameter tuning for a stable performance [47].

Neural Network: With the increasing availability of electronic health records [48], neural networks have garnered significant attention in the healthcare domain due to their outstanding performance [49]. Their robust performance is primarily attributed to large-scale training data.

Furthermore, the diverse architectures of neural networks facilitate seamless adaptation to various tasks; for instance, RNN-based networks effectively process temporal EHR data [50], while Transformer-based models excel in clinical text analysis [51]. However, their inherent black-box nature [52] poses a challenge in real-world applications [53, 54]. Additionally, the scarcity of data in certain rare disease phenotyping tasks may render direct application of neural networks unfeasible [55].

Tensor factorization: Alongside traditional machine learning and neural network models, tensor factorization stands out as another prominent computational phenotyping method [56, 57]. Ho et al. [56] propose Limestone to generate patients' phenotypes without supervision. Afshar et al. designed a framework TASTE [57] for the temporal EHR data. Tensor factorization has the ability to break down high-dimensional patient data into more manageable low-dimensional vectors, which can then serve as phenotypes for various downstream tasks.

Given the strengths and weaknesses of these electronic phenotyping methods, the selection of the most suitable approach should be tailored to the specific task and application context.

2.3.2 Bias in Electronic Phenotyping

We conduct a comprehensive analysis of bias in electronic phenotyping, examining it from both data and model perspectives as presented in Figure 2.1(a). Bias in phenotyping becomes evident when we observe variations in the method's performance across different subgroups defined by sensitive attributes like gender, race, and other factors. The origins of potential bias and their impact on phenotype outcomes will be discussed in greater detail below.

Data-level bias: The EHR data encompasses a diverse range of sources, including lab tests, diagnosis codes, treatment codes, and more. Given that electronic phenotyping methods heavily rely on data, any bias within the data significantly impacts the phenotyping outcomes. We categorize data bias into two main types: human decision bias and patient distribution bias. Human decision bias arises from clinical judgments, where certain records, such as diagnoses [58] and treatments [59], may exhibit biases due to human clinical decisions. For instance, phenotype rules crafted by humans may inadvertently underrepresent certain subgroups [58]. On the other hand,

patient distribution bias indicates an imbalance in patient representation due to disparities in cohort selection procedures. This can occur when minority patient groups are underrepresented, possibly stemming from limited access to the healthcare system [60]. It's crucial to recognize that biases at the data level inevitably permeate into the models trained on such data [10].

Model-level bias: The bias in the phenotyping model will also affect the phenotype fairness. As described in Section 2.3.1, the bias can be summarized into two categories. The first one is the bias in human-defined rules, which usually exists in some heuristic phenotyping methods like the rule-based method [58]. The second one is algorithm bias which commonly exists in artificial intelligence methods. The artificial intelligence algorithm will be trained toward optimal prediction accuracy while sacrificing fairness [61, 62]. There will be prediction disparities between different subgroups, e.g., some subgroups will have more positive predictions, the accuracy will also be higher on some subgroups, etc [12].

The presence of bias in electronic phenotyping can lead to unfair treatment of certain patient subgroups. Moreover, the patient cohorts derived from phenotyping may introduce bias into subsequent processes, such as the recruitment of patients for clinical trials. Addressing bias in phenotyping, both at the data and model levels, represents an ongoing challenge and an area for further research.

2.3.3 Bias Mitigation Method

The methods for mitigating bias can be classified into three categories: pre-processing, in-processing, and post-processing [12]. These approaches are implemented at various stages within the electronic phenotyping pipeline as demonstrated in Figure 2.1(d).

Pre-processing method: Pre-processing method [63, 64] aims to remove the bias-related information in the input data. There are two kinds of input bias-related information. The first one is sensitive features (explicit bias information) such as gender and race for which we can directly remove them. The second one is implicit bias [65, 66]. For example, the zip code is not a sensitive attribute but may be related to the race population. Moreover, we can resample the subgroups in the training data or re-weight each sample to mitigate the bias in training [63].

Table 2.1: Statistical summary of MIMIC-III database

	# Patients	# Healthy	# Diagnoses	# ICD codes	# Medications
MIMIC-III	38699	7821	15692	5435	1339918
Pneumonia	1419	1419	\	1606	954
Sepsis	1096	1096	\	1513	892

In-processing method: In-processing method focuses on the model training part. The in-processing method will guide the model to be trained for unbiased predictions by adding fairness-related constraints or regularization. This kind of method is the most commonly used one in the machine learning community because of its flexibility and generalizability for different scenarios and settings. One kind of in-processing method is adding regularization, e.g., neural network local interpretation during the training stage [67, 68]. Another main category of the method is adversarial learning, which will train a model for prediction and another model for adversarial classification [66, 69].

Post-processing method: The post-processing method directly processes the model outputs to force the outputs to be less biased. This method can be widely applied to various kinds of methods but it needs the patients' sensitive attributes which may be unavailable due to the private issue [70, 71].

2.4 Datasets and Methods

2.4.1 Datasets and Tasks

Pneumonia and sepsis phenotype. We use the widely applied MIMIC-III [72, 73] as the dataset for the following experiments in this work. Based on the MIMIC-III, we choose pneumonia and sepsis as our target phenotyping diseases because of their significant importance [74, 75].

Cohort selection. We select the target patient cohort based on their "DIAGNOSIS" feature in the "ADMISSIONS" file. We filter out 1419 patients diagnosed with pneumonia and 1096 patients diagnosed with sepsis as shown in Table 2.1. For the negative patients, we randomly sampled the same number as the positive patients from the neonatal patients in MIMIC-III.

Data processing. We extract the diagnostic codes and drug names from the patients' histories. We follow the data preprocessing procedures in TASTE framework [57] to group these ICD-9 codes and medical names into higher-level categories to avoid the potential data leakage issue in the following model training. We will first convert the ICD-9 codes to the ICD-10 codes and respectively use the Clinical Classification Software (CCS) system and the Anatomical Therapeutic Chemical (ATC) classification system to transfer ICD codes and drug names into more general classifications. The large number of features from ICD codes and medications has been reduced discernibly to CCS and ATC codes. For pneumonia, we get 232 CCS codes and 285 ATC codes. For sepsis, we get 231 CCS codes and 270 ATC codes. As each patient may have multiple visits, we formulate the input containing both temporal features and static features. We formulate the input as 3D tensors [57] consisting of patients, hospital visits, and temporal attributes. For the sensitive attributes, we chose gender and race as the research targets.

2.4.2 Study Design

In this section, we will introduce the proposed study design to comprehensively investigate and mitigate the bias in electronic phenotyping. First, we will discuss how to quantitatively identify and measure the bias in phenotyping with two bias metrics. Then, we will investigate how to mitigate the bias from the computational perspective.

Identify bias in electronic phenotyping. To investigate the bias in electronic phenotyping comprehensively, we first benchmark 4 main categories of widely used phenotyping methods as described in Section 2.3.1. We include **9** electronic phenotyping methods in this work, which are the rule-based method, logistic regression (LR) [37], random forest (RF) [39], SVM [40], gradient boosting decision tree (GBDT) [38], MLP [76], RNN [50], and LSTM [77]. We use the ROCAUC as metrics to measure the phenotyping accuracy and demographic parity difference (DPD), and equality odds difference (EOD) as the bias metrics. We will introduce the details of the phenotyping methods and metrics in Section 2.4.3.1. We will analyze different methods' performance on the phenotyping tasks and the bias respectively.

Mitigate bias from the computational perspective. We evaluate three main categories of de-

biasing algorithms as introduced in Section 2.3.3 to mitigate the phenotyping bias. We choose **2** pre-processing debias method, **2** in-processing debias method and **1** post-processing debias method. All these representative debiasing methods will be tested on the phenotyping methods described above if applicable. We will use the bias and performance metrics to investigate the mitigation effectiveness of different debiasing methods on various phenotyping algorithms.

2.4.3 Methods

2.4.3.1 Bias measure metrics

We will introduce the details of two bias metrics and their clinical meaning in the electronic phenotyping as follows. We use $\hat{Y}$ to denote the prediction of the phenotyping model, Y to denote the true label, and S to denote the sensitive attribute of each patient, e.g., gender, race, etc.

Demographic Parity Difference (DPD): The DPD measures the disparities of positive model outputs between different subgroups as shown in the equation (2.1). In the context of phenotype, the positive outputs indicate the diagnosis of specific diseases. DPD implies the bias in the probability of diagnosis between different patient groups.

$$\mathbf{DPD} = |P(\hat{Y} = 1|S = 0) - P(\hat{Y} = 1|S = 1)| \tag{2.1}$$

Equalized Odds Difference (EOD): The EOD measures the disparities of true positive outcomes between different subgroups as presented in the equation (2.2). EOD measures the bias of correctly identifying patients with specific diseases or phenotypes.

$$\mathbf{EOD} = |P(\hat{Y} = 1|Y = 1, S = 0) - P(\hat{Y} = 1|Y = 1, S = 1)| \tag{2.2}$$

2.4.3.2 Eletronic phenotyping methods

We formulate four categories of phenotyping methods as follows.

Rule-based methods: Rule-based methods are usually human-defined $if...else...$ rules, whose inputs are selected features $X_{selected}$, e.g., ICD-codes, etc. Rule-based methods can be represented

as follows in general.

$$\hat{Y} = Rules(X_{selected}) \tag{2.3}$$

The rule-based method adopted in this work is based on the PheKB.

Traditional machine learning: Traditional machine learning methods consist of training and testing stages and require feature engineering on the raw patient data Raw_{train}, Raw_{test}. We formulate the traditional machine learning phenotype pipeline as follows:

$$X_{train}, X_{test} = FE(Raw_{train}, Raw_{test}) \tag{2.4}$$

$$\hat{Y}_{train} = ML(X_{train}) \tag{2.5}$$

$$\hat{Y}_{test} = ML(X_{test}), \tag{2.6}$$

where ML models can be LR, RF, GBDT, and SVM in this work.

Neural networks: Neural network needs to design the network architecture and train on large-scale data. We formulate the phenotyping method with the neural network as follows.

$$\hat{Y}_{train} = NN(X_{train}, \theta), Loss = l(\hat{Y}_{train}, Y_{train}) \tag{2.7}$$

$$\hat{Y}_{test} = NN(X_{test}, \theta), \tag{2.8}$$

where θ is the trainable parameters of the neural network and the loss function l can be binary cross entropy. We choose MLP, RNN, and LSTM these three representative models to instantiate NN in this work.

Tensor factorization: Tensor factorization algorithm decomposes the input data into latent factor matrices for all three dimensions. We use the latent factor matrix of patient dimension

for our phenotyping task and one machine learning algorithm as the classifier. This phenotyping method can be formulated as follows.

$$M^p, M^v, M^t = TF(X), Loss = l(M^p \cdot M^v \cdot M^t, X) \tag{2.9}$$

$$M^p_{train}, M^p_{test} = split(M^p) \tag{2.10}$$

$$\hat{Y}_{test} = ML(M^p_{test}), \tag{2.11}$$

where X represents the raw input and $split()$ is to separate the whole dataset into train and test sets. TF represents the tensor factorization process, and the loss function is based on the cross entropy between the product of the resulting three latent factor matrices and the raw input data. We choose PARAFAC [78] as the tensor factorization algorithm and LR as the classifier in this work.

2.4.3.3 Debiasing methods

We formulate three kinds of debiasing methods as follows.

Pre-processing debias: One kind of pre-processing debias method will be operated on the input data to remove the explicit and implicit bias features. Specifically, we utilize the Pearson Correlation Coefficient (PCC) to determine the level of correlation between the two variables and set a threshold to remove strongly correlated features that exceed this threshold. This process can be presented as follows.

$$X_{debias} = Remover(X, threshold) \tag{2.12}$$

$$\hat{Y} = Model(X_{debias}), \tag{2.13}$$

where $Remover()$ is the algorithm that removes the sensitive related features, for which we choose correlation remover in this work. The threshold is manually set for determining if the feature should be eliminated.

Another pre-processing debias method is resampling, which resample the ratio of different subgroups to make the balance of them. The process can be represented as follows.

$$X_{resample} = Resampler(X, S) \tag{2.14}$$

$$\hat{Y} = Model(X_{resample}), \tag{2.15}$$

In-processing debias: In-processing debias method performs during the model training stage. One method to guide the model to be trained toward fair predictions is by adding some fairness constraints, which is one kind of widely applied method. In our experiment, the classification reduction algorithm [79] is adopted for this guiding. Typically, the objective of this algorithm is to minimize the disparity in prediction between different groups during the training process. This process can be presented as follows.

$$\hat{Y} = Model(X), \tag{2.16}$$

$$Loss = l(\hat{Y}_{train}, Y_{train}) + fairness\_constraint, \tag{2.17}$$

where $fairness\_constraint$ is the regularization that ensures the prediction fairness, for which we use demographic parity constraint in this work.

Another mainstream of the in-processing debiasing method is adversarial learning, which will train a predictor model and an adversary model. The predictor model will be trained with conventional strategy as shown below.

$$\hat{Y}_{train} = Predictor(X_{train}, \theta_P), L_P = l(\hat{Y}_{train}, Y_{train}) \tag{2.18}$$

Meanwhile, an adversary model will be trained to predict the sensitive attributes based on the

predictions from the $Predictor$ model. This process can be formulated as follows:

$$\hat{S}_{train} = AdverseNN(\hat{Y}_{train}, \theta_A), L_A = l(\hat{S}_{train}, S_{train}) \tag{2.19}$$

The overall optimization goal is combining two losses of predictor network and adversary model as follows.

$$L = \alpha L_P + \beta L_A, \tag{2.20}$$

where α, β are hyper-parameters that control the ratio of two losses. In this work, we use the adversarial debiasing method proposed by Zhang et al. [11].

Post-processing debias: Post-processing debiasing method directly calibrates the model outputs, which can be formulated as:

$$\hat{Y} = Model(X), \tag{2.21}$$

$$\hat{Y}_{cal} = Calibrator(\hat{Y}), \tag{2.22}$$

In this work, a threshold-based post-processing technique is employed as a method of calibration, based on the principle of equality of opportunity in model predictions, as articulated by Hardt et al. [71].

2.4.4 Implementation Details

This section introduces the implementation details of different types of rule-based and machine learning models.

Algorithm implementation: In our experiment, all algorithmic implementations have been actualized within the Python 3.8 environment. We leverage the rule-based algorithms available

from the Phenotype Knowledgebase (PheKB) [34] community as part of our analytical framework on the MIMIC-III. Furthermore, we leverage the scikit-learn library to implement traditional machine learning methodologies, employing the tensorly library for tensor factorization algorithms and the PyTorch library for the development and deployment of our neural network models. For the critical task of debiasing methods, we call upon the 0.9.0 version of the fairlearn library for the traditional machine learning implementation. However, in instances where the fairlearn library does not provide any support, we undertake the development of our own debias procedures for our models.

Model and training detail: In pursuit of robust and reliable results during the training phase, we rigorously employ a 5-fold cross-validation methodology, thereby facilitating the robust estimation of our measuring metrics. In configuring the training hyperparameters, we set the maximum iterations for logistic regression (LR) and support vector machines (SVM) to 120, while opting for a total of 30 estimators for tree-based models. The hidden size of both LSTM and RNN models is set to 128. For neural network models, we deliberately define key parameters, specifying a learning rate of $1e-04$, a minibatch size of 256, and an epoch number of 40 to ensure convergence and effective training. Additionally, in the context of tensor factorization, we establish the rank of the latent factor matrix at 20.

2.5 Results

We analyze the experiment results from two perspectives. The first one is bias measurement in the phenotyping. The other is how the debiasing algorithms perform.

2.5.1 Bias Measurement in Phenotyping

We summarize several key findings from the bias measurement results in two diseases phenotyping as follows.

- **The electronic phenotyping bias across races is more significant than genders.** From Table 2.2, we can observe the race DPD bias is about 7% higher than the gender bias on average across different phenotyping methods without debiasing strategies. From Table 2.4,

Table 2.2: Debiasing results of pneumonia phenotyping (Demographic Parity Difference. Correlation remover and resample are the pre-processing methods. Reduction and adversarial mitigation are the in-processing methods. Threshold optimizer is a post-processing method.)

Disease phenotyping		**Rule Based**	**Machine Learning**				**Tensor Factorization**	**Deep Learning**		
Sensitive Attribute	**Debias Method**	**PheKB-ICD**	**LR**	**RF**	**SVM**	**GBC**	**PARAFAC+LR**	**MLP**	**RNN**	**LSTM**
Gender (Input include)	**Correlation Remover**	0.000	0.031±0.001	0.036±0.001	0.047±0.001	0.037±0.001	0.037±0.001	0.041±0.001	0.041±0.001	0.040±0.001
	Resample	0.010	0.036±0.001	0.040±0.001	0.043±0.001	0.038±0.001	0.037±0.001	0.038±0.001	0.035±0.002	0.037±0.001
	Reduction	/	0.039±0.001	0.036±0.001	0.039±0.001	0.037±0.001	0.047±0.001	0.036±0.001	0.036±0.001	0.035±0.001
	Threshold Optimizer	0.000	0.049±0.001	0.045±0.000	0.052±0.001	0.045±0.001	0.048±0.001	0.053±0.001	0.043±0.001	0.040±0.001
	Adversarial Mitigation	/	/	/	/	/	/	0.043±0.001	0.038±0.001	0.087±0.014
	w/o debias	0.000	0.031±0.001	0.036±0.001	0.047±0.001	0.037±0.001	0.037±0.001	0.041±0.001	0.064±0.001	0.040±0.001
Race (Input include)	**Correlation Remover**	0.006	0.127±0.002	0.141±0.001	0.039±0.000	0.142±0.001	0.131±0.001	0.146±0.001	0.150±0.001	0.148±0.001
	Resample	0.036	0.024±0.001	0.028±0.000	0.024±0.000	0.027±0.000	0.045±0.003	0.026±0.000	0.021±0.000	0.026±0.000
	Reduction	/	0.082±0.002	0.088±0.001	0.052±0.002	0.098±0.001	0.049±0.001	0.074±0.001	0.064±0.002	0.060±0.002
	Threshold Optimizer	0.001	0.034±0.001	0.034±0.001	0.026±0.000	0.036±0.002	0.029±0.001	0.030±0.000	0.023±0.000	0.028±0.001
	Adversarial Mitigation	/	/	/	/	/	/	0.141±0.001	0.169±0.002	0.146±0.001
	w/o debias	0.006	0.127±0.002	0.141±0.001	0.039±0.000	0.142±0.001	0.131±0.001	0.148±0.001	0.145±0.001	0.072±0.022
Gender (Input exclude)	**Correlation Remover**	0.000	0.037±0.001	0.037±0.001	0.037±0.000	0.037±0.001	0.050±0.001	0.036±0.001	0.043±0.001	0.043±0.001
	Resample	0.010	0.038±0.001	0.038±0.001	0.038±0.001	0.038±0.001	0.039±0.001	0.039±0.001	0.040±0.001	0.038±0.001
	Reduction	/	0.037±0.001	0.037±0.001	0.037±0.000	0.037±0.001	0.050±0.001	0.037±0.001	0.040±0.001	0.042±0.001
	Threshold Optimizer	0.000	0.047±0.001	0.037±0.001	0.047±0.001	0.043±0.001	0.060±0.001	0.041±0.001	0.044±0.001	0.040±0.001
	Adversarial Mitigation	/	/	/	/	/	/	0.041±0.001	0.035±0.001	0.035±0.000
	w/o debias	0.000	0.037±0.001	0.037±0.001	0.037±0.001	0.037±0.001	0.050±0.001	0.036±0.001	0.043±0.001	0.041±0.001
Race (Input exclude)	**Correlation Remover**	0.006	0.142±0.001	0.142±0.001	0.142±0.0009	0.142±0.001	0.103±0.000	0.141±0.001	0.138±0.001	0.141±0.000
	Resample	0.036	0.027±0.000	0.028±0.001	0.028±0.001	0.027±0.000	0.043±0.001	0.028±0.001	0.033±0.001	0.030±0.001
	Reduction	/	0.126±0.002	0.144±0.002	0.125±0.001	0.123±0.001	0.101±0.001	0.135±0.001	0.136±0.000	0.138±0.000
	Threshold Optimizer	0.001	0.025±0.000	0.027±0.001	0.043±0.000	0.041±0.000	0.019±0.000	0.033±0.000	0.044±0.000	0.035±0.001
	Adversarial Mitigation	/	/	/	/	/	/	0.125±0.001	0.135±0.001	0.110±0.004
	w/o debias	0.006	0.142±0.001	0.142±0.001	0.142±0.001	0.142±0.001	0.103±0.000	0.141±0.001	0.141±0.001	0.141±0.001

Table 2.3: Debiasing results of pneumonia phenotyping (Equalized Odds Difference)

Disease phenotyping		Rule Based	Machine Learning				Tensor Factorization	Deep Learning		
Sensitive Attribute	**Debias Method**	**PheKB-ICD**	**LR**	**RF**	**SVM**	**GBC**	**PARAFAC+LR**	**MLP**	**RNN**	**LSTM**
Gender (Input included)	**Correlation Remover**	0.007	0.030±0.000	0.008±0.000	0.055±0.002	0.000±0.000	0.024±0.000	0.007±0.000	0.010±0.000	0.006±0.000
	Resample	0.019	0.009±0.000	0.007±0.000	0.025±0.000	0.000±0.000	0.016±0.000	0.007±0.000	0.010±0.000	0.007±0.000
	Reduction	/	0.025±0.000	0.008±0.000	0.039±0.001	0.000±0.000	0.029±0.001	0.010±0.000	0.017±0.000	0.013±0.000
	Threshold Optimizer	0.005	0.039±0.001	0.061±0.001	0.052±0.002	0.055±0.001	0.037±0.000	0.050±0.001	0.045±0.001	0.043±0.000
	Adversarial Mitigation	/	/	/	/	/	/	0.015±0.000	0.010±0.000	0.146±0.084
	w/o debias	0.007	0.030±0.000	0.008±0.000	0.055±0.002	0.000±0.000	0.024±0.000	0.007±0.000	0.011±0.000	0.006±0.000
Race (Input included)	**Correlation Remover**	0.048	0.064±0.000	0.010±0.000	0.121±0.003	0.000±0.000	0.049±0.000	0.020±0.000	0.024±0.000	0.021±0.000
	Resample	0.072	0.053±0.003	0.009±0.000	0.074±0.004	0.000±0.000	0.107±0.003	0.021±0.000	0.028±0.000	0.043±0.001
	Reduction	/	0.102±0.001	0.093±0.001	0.116±0.002	0.074±0.001	0.140±0.002	0.110±0.001	0.125±0.003	0.135±0.003
	Threshold Optimizer	0.048	0.224±0.004	0.246±0.001	0.210±0.005	0.252±0.004	0.234±0.002	0.215±0.002	0.243±0.001	0.238±0.002
	Adversarial Mitigation	/	/	/	/	/	/	0.016±0.000	0.047±0.005	0.146±0.001
	w/o debias	0.048	0.064±0.000	0.010±0.000	0.121±0.003	0.000±0.000	0.049±0.000	0.018±0.000	0.021±0.000	0.024±0.000
Gender (Input excluded)	**Correlation Remover**	0.007	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.040±0.002	0.001±0.000	0.017±0.000	0.020±0.000
	Resample	0.019	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.050±0.000	0.002±0.000	0.015±0.000	0.006±0.000
	Reduction	/	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.040±0.002	0.000±0.000	0.019±0.000	0.020±0.001
	Threshold Optimizer	0.005	0.046±0.002	0.049±0.000	0.040±0.001	0.038±0.001	0.053±0.001	0.031±0.001	0.039±0.000	0.027±0.000
	Adversarial Mitigation	/	/	/	/	/	/	0.020±0.000	0.015±0.000	0.016±0.000
	w/o debias	0.007	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.040±0.002	0.001±0.000	0.020±0.000	0.017±0.000
Race (Input excluded)	**Correlation Remover**	0.048	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.028±0.001	0.001±0.000	0.010±0.000	0.013±0.000
	Resample	0.072	0.000±0.000	0.003±0.000	0.003±0.000	0.000±0.000	0.058±0.000	0.003±0.000	0.023±0.000	0.018±0.000
	Reduction	/	0.026±0.000	0.018±0.000	0.022±0.001	0.026±0.000	0.052±0.000	0.019±0.000	0.017±0.000	0.022±0.000
	Threshold Optimizer	0.048	0.247±0.001	0.257±0.001	0.241±0.002	0.236±0.003	0.125±0.000	0.216±0.001	0.230±0.004	0.195±0.001
	Adversarial Mitigation	/	/	/	/	/	/	0.036±0.001	0.017±0.000	0.010±0.000
	w/o debias	0.048	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.028±0.001	0.001±0.000	0.012±0.000	0.012±0.000

the race DPD bias is over 12% larger than gender bias. Similarly, when the bias metric is EOD, the race bias is 2%, 3% higher than gender bias on pneumonia and sepsis phenotyping respectively according to the Table 2.3 and Table 2.5.

- **Phenotyping bias varies across different phenotyping algorithms. Rule-based Phenotyping method shows significantly less bias.** From Table 2.2 and Table 2.4, we can find different phenotyping method presents various levels of bias under different settings. When sensitive attribute is included, in pneumonia phenotyping, RNN shows the highest gender bias, and MLP has the highest racial bias. For the sepsis phenotyping, SVM and MLP present the highest gender bias when gender is not included in the input. Meanwhile, MLP and RNN have the highest racial bias in sepsis phenotyping. When we exclude the sensitive

Table 2.4: Debiasing results of sepsis phenotyping (Demographic Parity Difference)

Disease phenotyping		Rule Based	Machine Learning				Tensor Factorization	Deel Learning		
Sensitive Attribute	**Debias Method**	**PheKB-ICD**	**LR**	**RF**	**SVM**	**GBC**	**PARAFAC+LR**	**MLP**	**RNN**	**LSTM**
Gender (Input include)	**Correlation Remover**	0.007	0.017±0.000	0.012±0.000	0.019±0.000	0.014±0.000	0.018±0.000	0.018±0.000	0.019±0.000	0.016±0.000
	Resample	0.001	0.041±0.001	0.032±0.001	0.042±0.001	0.033±0.000	0.034±0.000	0.033±0.000	0.037±0.000	0.035±0.001
	Reduction	/	0.021±0.000	0.012±0.000	0.025±0.000	0.014±0.000	0.042±0.001	0.019±0.000	0.016±0.000	0.019±0.000
	Threshold Optimizer	0.000	0.023±0.000	0.015±0.000	0.060±0.003	0.017±0.000	0.024±0.000	0.022±0.000	0.023±0.000	0.027±0.000
	Adversarial Mitigation	/	/	/	/	/	/	0.027±0.000	0.012±0.000	0.015±0.000
	w/o debias	0.007	0.017±0.000	0.012±0.000	0.019±0.000	0.014±0.000	0.018±0.000	0.019±0.000	0.016±0.000	0.016±0.000
Race (Input include)	**Correlation Remover**	0.140	0.122±0.001	0.149±0.001	0.051±0.002	0.148±0.001	0.135±0.002	0.163±0.001	0.163±0.001	0.160±0.001
	Resample	0.029	0.039±0.001	0.028±0.000	0.043±0.002	0.037±0.000	0.021±0.000	0.024±0.000	0.020±0.000	0.022±0.000
	Reduction	/	0.091±0.003	0.094±0.003	0.061±0.003	0.092±0.003	0.060±0.001	0.068±0.002	0.072±0.002	0.066±0.002
	Threshold Optimizer	0.004	0.041±0.001	0.033±0.000	0.028±0.001	0.053±0.001	0.029±0.000	0.035±0.001	0.047±0.001	0.047±0.002
	Adversarial Mitigation	/	/	/	/	/	/	0.158±0.001	0.156±0.002	0.166±0.001
	w/o debias	0.140	0.122±0.001	0.148±0.001	0.051±0.002	0.148±0.001	0.135±0.002	0.163±0.001	0.163±0.001	0.162±0.001
Gender (Input exclude)	**Correlation Remover**	0.007	0.015±0.000	0.014±0.000	0.015±0.000	0.015±0.000	0.063±0.002	0.014±0.000	0.017±0.000	0.019±0.000
	Resample	0.001	0.034±0.000	0.034±0.000	0.034±0.000	0.034±0.000	0.050±0.001	0.033±0.000	0.038±0.001	0.038±0.002
	Reduction	/	0.015±0.000	0.014±0.000	0.015±0.000	0.015±0.000	0.063±0.002	0.014±0.000	0.017±0.000	0.019±0.000
	Threshold Optimizer	0.000	0.026±0.000	0.025±0.000	0.020±0.000	0.019±0.000	0.050±0.001	0.016±0.000	0.026±0.000	0.020±0.000
	Adversarial Mitigation	/	/	/	/	/	/	0.013±0.000	0.010±0.000	0.023±0.000
	w/o debias	0.007	0.015±0.000	0.014±0.000	0.015±0.000	0.015±0.000	0.063±0.002	0.014±0.000	0.019±0.000	0.018±0.000
Race (Input exclude)	**Correlation Remover**	0.140	0.149±0.001	0.148±0.001	0.149±0.001	0.149±0.001	0.138±0.003	0.148±0.001	0.155±0.001	0.156±0.001
	Resample	0.029	0.035±0.000	0.034±0.000	0.034±0.000	0.035±0.000	0.058±0.001	0.035±0.000	0.026±0.000	0.026±0.000
	Reduction	/	0.127±0.000	0.147±0.001	0.138±0.000	0.125±0.002	0.127±0.001	0.149±0.001	0.156±0.002	0.154±0.001
	Threshold Optimizer	0.004	0.021±0.000	0.044±0.002	0.018±0.000	0.055±0.000	0.032±0.000	0.036±0.001	0.046±0.001	0.032±0.001
	Adversarial Mitigation	/	/	/	/	/	/	0.141±0.001	0.140±0.001	0.149±0.002
	w/o debias	0.140	0.149±0.001	0.148±0.001	0.149±0.001	0.149±0.001	0.138±0.003	0.148±0.001	0.155±0.001	0.154±0.001

attribute from input features, the RNN presents the largest gender bias while 4 ML models show the highest racial bias in both pneumonia and sepsis phenotyping. Moreover, we find the rule-based method presents significantly less bias compared to other methods. In the pneumonia phenotyping, rule-based method shows 4% lower gender bias and 11% lower race bias in terms of DPD. Noticeably, observed from Table 2.3 and Table 2.5, while the EOD gap in gender is 1% lower, the average race bias of rule-based algorithms is 1% higher.

- **Exclude the sensitive attributes from input data has a trivial effect on the bias.** Excluding the sensitive attributes is the most intuitive method to mitigate the bias from the observation in Table 2.2, 2.3, 2.4, and 2.5. However, we find that after excluding the sensitive attributes, the bias is still significant. In pneumonia phenotyping, the gender bias of LR,

Table 2.5: Debiasing results of sepsis phenotyping (Equalized Odds Difference)

Disease phenotyping		Rule Based	Machine Learning				Tensor Factorization	Deep Learning		
Sensitive Attribute	**Debias Method**	**PheKB-ICD**	**LR**	**RF**	**SVM**	**GBC**	**PARAFAC+LR**	**MLP**	**RNN**	**LSTM**
Gender (Input included)	**Correlation Remover**	0.005	0.037±0.000	0.010±0.000	0.085±0.001	0.004±0.000	0.027±0.000	0.011±0.000	0.011±0.000	0.009±0.000
	Resample	0.004	0.026±0.000	0.014±0.000	0.062±0.002	0.004±0.000	0.036±0.000	0.013±0.000	0.015±0.000	0.013±0.000
	Reduction	/	0.015±0.000	0.010±0.000	0.049±0.001	0.004±0.000	0.058±0.001	0.009±0.000	0.029±0.000	0.025±0.000
	Threshold Optimizer	0.010	0.029±0.000	0.014±0.000	0.121±0.002	0.012±0.000	0.027±0.000	0.019±0.000	0.020±0.000	0.020±0.000
	Adversarial Mitigation	/	/	/	/	/	/	0.046±0.002	0.016±0.000	0.007±0.000
	w/o debias	0.005	0.037±0.000	0.010±0.000	0.085±0.001	0.004±0.000	0.027±0.000	0.011±0.000	0.013±0.000	0.009±0.000
Race (Input included)	**Correlation Remover**	0.093	0.093±0.001	0.010±0.000	0.129±0.007	0.006±0.000	0.063±0.003	0.038±0.001	0.049±0.003	0.038±0.002
	Resample	0.062	0.056±0.004	0.010±0.000	0.087±0.004	0.007±0.000	0.111±0.003	0.048±0.000	0.063±0.001	0.063±0.001
	Reduction	/	0.100±0.002	0.091±0.001	0.090±0.002	0.096±0.002	0.150±0.001	0.141±0.001	0.129±0.001	0.142±0.001
	Threshold Optimizer	0.187	0.251±0.003	0.266±0.000	0.130±0.005	0.241±0.003	0.220±0.000	0.265±0.002	0.255±0.004	0.264±0.006
	Adversarial Mitigation	/	/	/	/	/	/	0.039±0.003	0.040±0.003	0.043±0.001
	w/o debias	0.093	0.093±0.001	0.010±0.000	0.129±0.007	0.006±0.000	0.063±0.003	0.038±0.001	0.042±0.002	0.042±0.002
Gender (Input excluded)	**Correlation Remover**	0.005	0.002±0.000	0.001±0.000	0.002±0.000	0.002±0.000	0.091±0.006	0.000±0.000	0.029±0.000	0.033±0.000
	Resample	0.004	0.002±0.000	0.006±0.000	0.002±0.000	0.002±0.000	0.093±0.000	0.004±0.000	0.027±0.000	0.031±0.000
	Reduction	/	0.002±0.000	0.001±0.000	0.002±0.000	0.002±0.000	0.091±0.006	0.000±0.000	0.028±0.000	0.032±0.000
	Threshold Optimizer	0.010	0.024±0.000	0.032±0.001	0.024±0.000	0.013±0.000	0.068±0.002	0.004±0.000	0.041±0.000	0.035±0.000
	Adversarial Mitigation	/	/	/	/	/	/	0.027±0.000	0.012±0.000	0.028±0.000
	w/o debias	0.005	0.002±0.000	0.001±0.000	0.002±0.000	0.002±0.000	0.091±0.006	0.000±0.000	0.034±0.000	0.034±0.000
Race (Input excluded)	**Correlation Remover**	0.093	0.004±0.000	0.001±0.000	0.004±0.000	0.004±0.000	0.090±0.004	0.002±0.000	0.039±0.001	0.045±0.001
	Resample	0.062	0.004±0.000	0.007±0.000	0.007±0.000	0.004±0.000	0.081±0.002	0.004±0.000	0.040±0.001	0.029±0.000
	Reduction	/	0.047±0.001	0.014±0.000	0.044±0.001	0.027±0.000	0.095±0.003	0.017±0.000	0.044±0.002	0.038±0.001
	Threshold Optimizer	0.187	0.229±0.003	0.253±0.005	0.225±0.001	0.280±0.005	0.134±0.001	0.264±0.001	0.232±0.002	0.255±0.001
	Adversarial Mitigation	/	/	/	/	/	/	0.015±0.000	0.022±0.000	0.032±0.001
	w/o debias	0.093	0.004±0.000	0.001±0.000	0.004±0.000	0.004±0.000	0.090±0.004	0.002±0.000	0.043±0.001	0.041±0.001

RF, GBC, and LSTM increases after the gender feature is excluded, and the racial bias of LR, RF, SVM, LSTM increases after the race is excluded. In sepsis phenotyping, RF, GBC, RNN, and LSTM's gender bias increases. SVM, GBC's racial bias increases after the race attribute is excluded.

2.5.2 Performance of Debiasing Algorithms

- **The debiasing strategies are more effective on racial bias than gender bias.** From the gender part in both Table 2.2 and Table 2.4, we can find that the gender bias decreases with a non-trivial level. The reason may be the bias across genders is relatively small and trivial. So the debiasing method couldn't effectively mitigate the gender bias. However, most debiasing methods can reduce race bias significantly. For example, the race bias of MLP in pneumonia

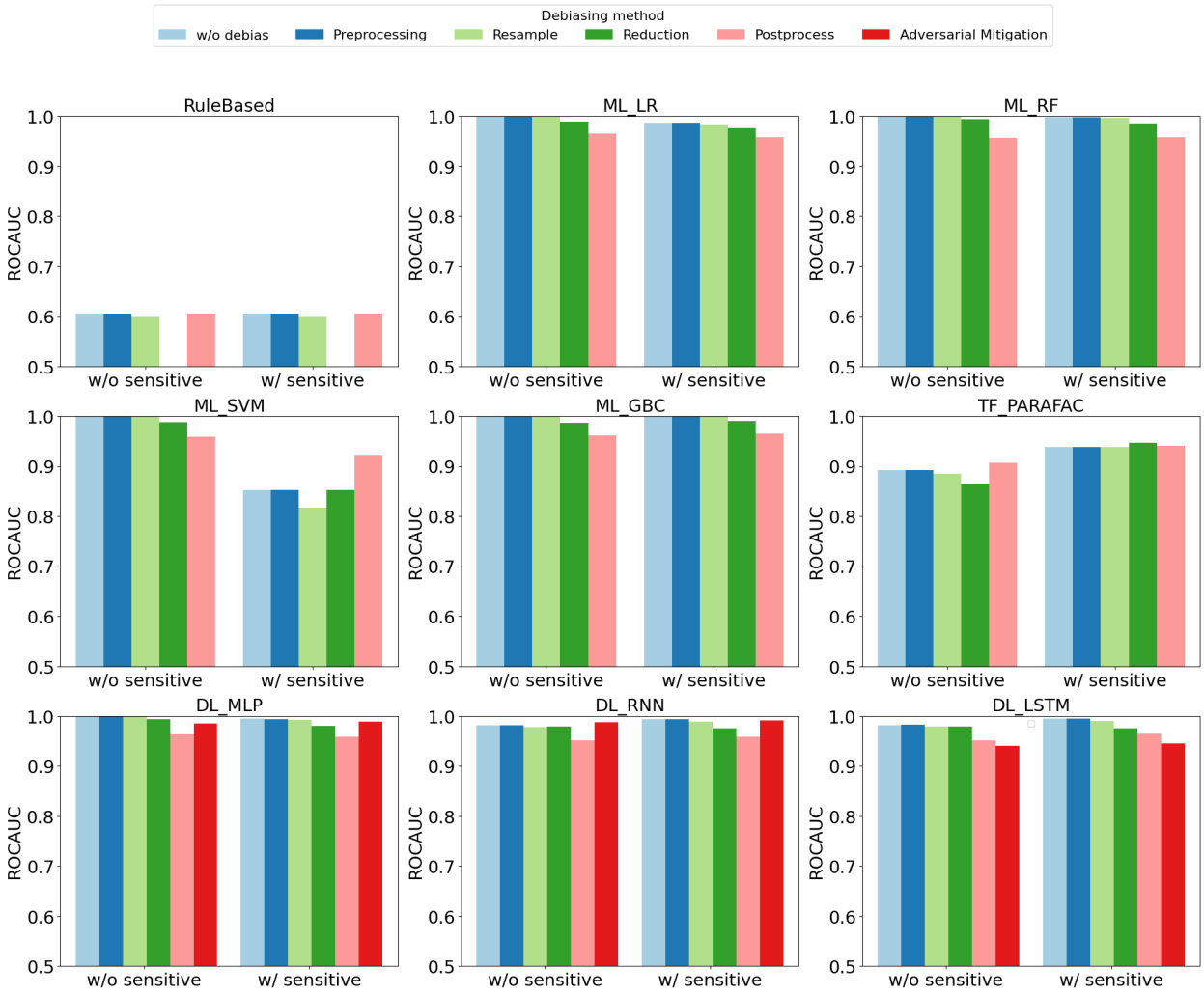


Figure 2.2: Pneumonia phenotype performance.

phenotyping is reduced by 12.2% with the resample debiasing method. The race bias of SVM can be further reduced by 1.5% with resample strategy in the sepsis phenotyping task.

- **Correlation removing method is not capable of mitigating the bias in phenotyping.** We can observe from Table 2.2 and Table 2.4, that removing the sensitive correlation from input features doesn't work for the sepsis and pneumonia phenotyping. For pneumonia phenotyping, the gender bias even increases a bit after the correlation removal in DL methods of **LSTM**. The race bias of **MLP** and **RNN** increases after the correlation removal. The situation is similar in the sepsis phenotype. This may be caused by the input feature containing

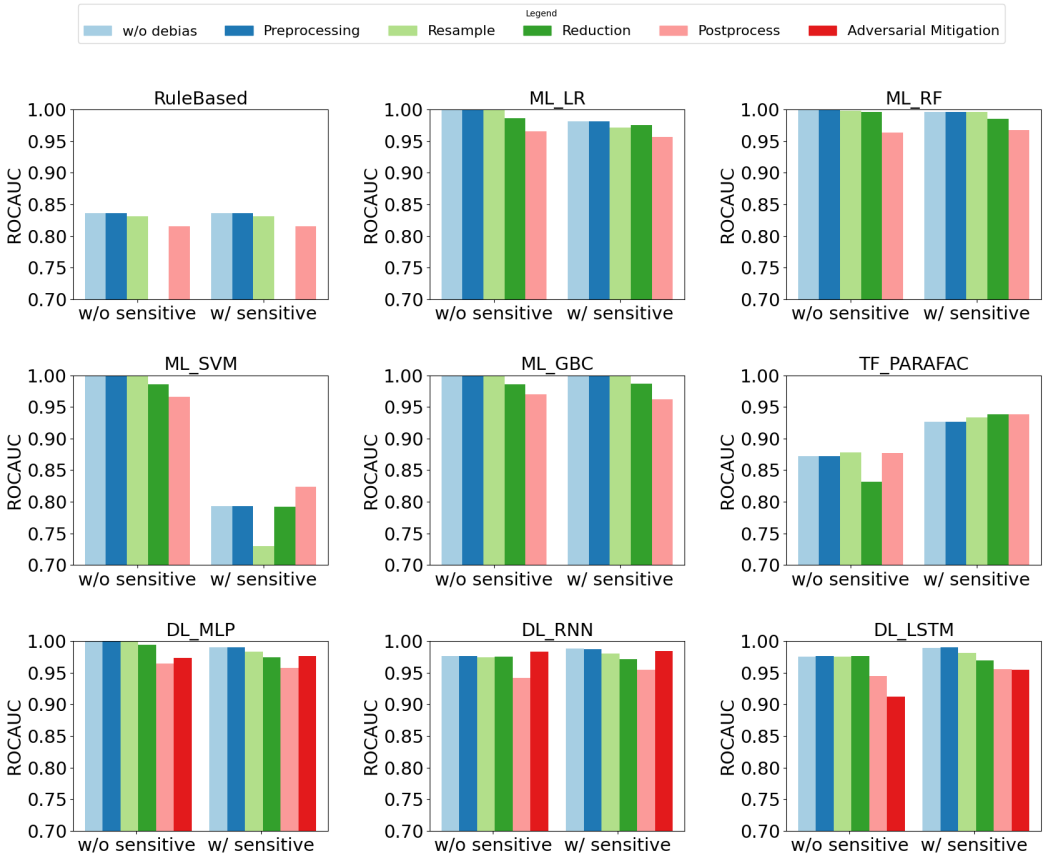


Figure 2.3: Sepsis phenotype performance.

little information related to the sensitive attributes.

- **Resample the patients' data and postprocess the outputs are two very simple yet effective debiasing methods.** From Table 2.2 and Table 2.4, we can find resampling the patients' data to make each subgroup size more balanced can significantly reduce the phenotype bias. The race bias in pneumonia phenotype has been reduced by 9% on average with either of resampling or postprocessing method. The highest gender bias can be reduced by 2.9% with resampling on RNN. For the sepsis phenotyping task, the highest race bias decrease by 14%. Nevertheless, the gender biases of our models are almost all below 2% and can hardly be further reduced even with resampling or postprocessing.

- **There is a trade-off between phenotyping accuracy and bias.** From Figure 2.2, we can find that most phenotyping models' phenotype accuracy will decrease when the debiasing method is applied. This phenomenon also appears in sepsis phenotype as shown in Figure 2.3. So when we develop and deploy the phenotyping method, we need to make a trade-off between accuracy and bias based on the real-world phenotyping requirement.

2.6 Discussion and Conclusion

From the experiment analysis on the main categories of phenotyping models and debiasing methods. We will discuss some limitations and future directions of this topic. We will also conclude this work with several takeaways and conclusions.

In this work, we choose two common diseases which are pneumonia and sepsis. However, there are some diseases that have specific characteristics. These specialties may make phenotyping bias on these diseases different from the findings we summarize in this work. For example, breast cancer is more commonly diagnosed among females compared to males [80]. The patients' data distribution across genders will be obviously different between females and males, which may cause significant gender bias in phenotyping. So for some specific diseases, we need to analyze their potential bias case by case.

We investigate the bias issue in phenotyping from a computational perspective. However, there is still a gap between the computational perspective and the clinical perspective. Addressing this gap represents one of the most promising and crucial directions for future research. In our future work, we will consider developing some methods that can clearly deliver computational fairness to the clinical practitioners and involve them to collaborate in the study. In this work, we mainly focus on the bias mitigation strategy in the data processing, model training, and output calibration steps. However, the data collection in healthcare is also very important. How to collect the data containing less bias remains a promising future direction.

To summarize, we comprehensively investigate the bias and the bias mitigation methods with pneumonia and sepsis phenotyping. From the perspective of phenotyping bias, we find that race bias is more obvious than gender bias and the rule-based phenotyping method demonstrates sig-

nificantly less bias than machine learning phenotyping methods. Simply excluding the sensitive attributes doesn't work well in bias mitigation. Moreover, from the perspective of bias mitigation, we find that resample and post-process these two methods are simple yet effective in bias mitigation. Moreover, if the fairness of the phenotyping model improves through mitigation, the phenotyping accuracy will be negatively affected to some extent. So the trade-off between fairness and accuracy needs to be considered when implementing and deploying the phenotyping model. The future work in this line of research can be derived in several directions. The first one is to develop more advanced debiasing methods for the phenotyping models according to the task specialties. The second is to bridge the gap of fairness between computation and clinical, which will help translate the computational debiasing methods into real-world clinical practice. The third direction is inspired by the findings from our experimental results that we can attach more importance to the healthcare data collection stage and improve the access of healthcare resources to the underrepresented groups.

3. MEDICAL HAI WITH FAIR-AWARE KNOWLEDGE DISTILLATION *

3.1 Overview

Liver transplant is an essential therapy performed for severe liver diseases. The fact of scarce liver resources makes the organ assigning crucial. Model for End-stage Liver Disease (MELD) score is a widely adopted criterion when making organ distribution decisions. However, it ignores post-transplant outcomes and organ/donor features. These limitations motivate the emergence of machine learning (ML) models. Unfortunately, ML models could be unfair and trigger bias against certain groups of people. To tackle this problem, this work proposes a fair machine learning framework targeting graft failure prediction in liver transplant. Specifically, knowledge distillation is employed to handle dense and sparse features by combining the advantages of tree models and neural networks. A two-step debiasing method is tailored for this framework to enhance fairness. Experiments are conducted to analyze unfairness issues in existing models and demonstrate the superiority of our method in both prediction and fairness performance.

3.2 Introduction

Liver transplant is an effective treatment option for end-stage liver diseases and acute liver failure such as hepatic failure. However, the transplant organ resources are scarce compared with the number of patients on the waiting list [81, 82]. Hence organ assignment becomes a crucial decision that demands careful consideration. A prevalently used assigning strategy is based on the Model for End-stage Liver Disease (MELD) score, which estimates the patient's current status based on three lab test results, including serum creatinine, total bilirubin, and INR of prothrombin time [83]. A higher MELD score indicates a worse situation of a patient, and thus a higher priority of the patient to receive organs. The new version MELD score also includes serum sodium for calculation [84]. For pediatric patients, the score definition is different, called Pediatric End-stage

*Reprinted with permission from "Fairly Predicting Graft Failure in Liver Transplant for Organ Assigning" by Sirui Ding, Ruixiang Tang, Daochen Zha, Na Zou, Kai Zhang, Xiaoqian Jiang, Xia Hu, 2022, AMIA Annual Symposium, Copyright by 2022 AMIA.

Liver Disease (PELD) score [85]. We do not differentiate those metrics in our study.

Despite its prevalence, MELD score has two main drawbacks. First, MELD score does not explicitly consider the post-transplant outcome [83, 86], which is an important metric for organ distributing decisions. Our experimental results show that MELD score only has a very weak correlation with graft failure rate (i.e., the likelihood of graft failure occurs) across genders and races with a Pearson correlation of only 0.36653 (see Table 3.2). Second, MELD score ignores the features of organs and donors [83, 84], which may lead to injudicious organ assigning decisions. (detailed in Section 3.6.1). As such, researchers are motivated to propose various substitute assignment strategies for liver transplant [87, 88].

Machine learning (ML) has provided data-driven solutions for the organ transplant task to better model post-transplant outcomes. The key idea is to train an ML model that takes the features of patients and donors as input, and outputs the predicted outcomes such as pre-transplant mortality, post-transplant mortality, etc. Then, the trained model is deployed to predict a score for each patient-donor pair, which can help clinicians make decisions of organ transplant. Recently, various ML models have been deployed and show promises in the organ transplant task [89, 90, 91]. For example, Byrd et al. [92] use logistic regression and gradient boosting models to predict mortality in liver transplant. Lau et al. apply neural network and random forest to predict graft failure after transplant [93]. Berrevoets et al. propose an interpretable method for real-time organ allocation [94].

Unfortunately, recent studies suggest that ML models could be unfair and show bias against certain groups of people in organ transplant. Several previous studies have discussed such fairness issues [95, 96, 97]. For example, Byrd et al. [92] show that the scores predicted by ML models underrate the mortality of the female group. Our preliminary experiments also show that the gap between GBDT's positive prediction rates across different race groups can be as large as 0.637 (see Table 3.3). The unfair predictions may cause unfair decisions towards specific race groups. Although some pioneer works point out the unfair issue, there exists no concrete solution that can tackle such unfairness problem to the best of our knowledge. Thus, we are motivated to study the

following research question: *can we develop an ML model that is both accurate and fair for the liver transplant task?*

While fairness problems in machine learning have been widely investigated recently [12, 10], there are few attempts to study the fairness problem in organ transplant tasks. Developing a fair ML system with competitive accuracy for organ transplant remains a challenging task due to two roadblocks. Firstly, organ transplant datasets contain both dense features (e.g., numerical lab test results) and sparse categorical features (e.g., blood type of recipients and donors). For sparse features, the existing studies simply use one-hot encoding for transformation [98]. However, one-hot encoding could lead to unsatisfactory performance when the feature cardinality is high due to the curse of dimensionality [99]. Secondly, it is challenging to incorporate fairness goals into the training process. Prior work mainly adopts tree-based models [100, 101] for organ transplant prediction due to its strong performance on handling dense inputs. However, existing bias mitigation algorithms mainly focus on the training process [12], including loss design and representation learning [102, 103]; neither of them can be directly applied to tree-based models because of the indifferentiable property.

To tackle these challenges, we propose a fair ML framework for liver transplant. Specifically, we focus on the prediction of liver* transplant graft failure which is one of the most important post-transplant outcomes. Motivated by the strong performance of DeepGBM [104] in recommendation tasks, we use an embedding layer to handle the sparse features and a distillation network with distilled knowledge from a tree-based model to handle the dense features. This design can not only combine the advantages of tree-based models and deep neural networks in handling the sparse and dense features, but also enable us to apply in-processing debiasing techniques to achieve fairness. In particular, we devise a two-step debiasing strategy that mitigates the fairness issues in both the knowledge distillation stage and the end-to-end training stage. We demonstrate the superiority of our framework through extensive experiments on the Standard Transplant Analysis and Research (STAR) dataset. Empirical results show that the proposed framework can precisely and fairly

*Liver and organ are considered exchangeable in this work when the context has no ambiguity.

predict graft failure across different races and genders.

3.3 Background of Fairness in Liver Transplant

In this section, we first describe fairness problems in liver transplant. Then we quantify the unfairness using the fairness metrics adopted in the ML community.

Fairness of liver transplant. Following the existing fairness research in medical fields [105, 106, 107], we study fairness in liver transplant at the group level and focus on race groups and gender groups. Specifically, a fair graft failure predictor should allow patients of different races and genders to have an equal chance of receiving compatible organs. However, fairness is a subjective term so that equal chance could have different interpretations. In this work, we consider fairness defined from two perspectives. On one hand, we expect the patients in different groups to have an equal percentage of being predicted as *graft failed*. In this sense, the patient in different groups will tend to equally receive an organ if allocating organs based on the predicted score. On the other hand, ML models are expected to provide an equal prediction quality for different groups, which can be quantified by true positive rates and false positive rates of the graft failure prediction.

Fairness metrics. The above two fairness definitions correspond to two commonly used fairness metrics for ML models: demographic parity and equalized odds, where the former demands different groups to have an equal percentage of a positive outcome, and the latter requires equal true positive and false positive rates. Specifically, we follow previous work and quantify the degrees of demographic parity and equalized odds with demographic parity difference (DPD) and equalized odds difference (EOD) [108], respectively. We put detailed mathematical definitions in Section 3.4.

3.4 Data and Problem Description

Dataset. The Standard Transplant Analysis and Research (STAR) organ transplant dataset is collected from patients registered on the Organ Procurement and Transplantation Network (OPTN) waiting list, de-identified by removing all the identifiers from data and randomly shifted dates under IRB protocol approval (HSC-MS-13-0549). It consists of the biomedical information of

both patients and organs/donors. The patients include the ones on the waiting list and the recipients who received organ transplants. The dataset also provides follow-up records of recipients' post-transplant outcomes. For the graft failure prediction task, we select 160360 recipients, where 41.8% of them suffer from graft failure. We manually choose 40 features of recipients and 40 features from organs/donors. The race and gender of each recipient are marked as sensitive attributes.

Notations. We denote scalars as lowercase alphabets (e.g., x), vectors as boldface lowercase alphabets (e.g., $\mathbf{x}$), matrices as boldface uppercase alphabets (e.g., $\mathbf{X}$). We represent the liver transplant dataset as $\mathcal{D} = \{(\mathbf{r}_i, \mathbf{s}_i, \mathbf{o}_i, y_i)\}_{i=1}^N$, where $\mathbf{r}_i \in \mathbb{R}^{M_r}$ denotes the features of the recipient (e.g., various kinds of lab test results, etc), $\mathbf{s}_i \in \mathbb{R}^{M_s}$ denotes the sensitive features of the recipient (e.g., the demographic information), $\mathbf{o}_i \in \mathbb{R}^{M_o}$ denotes the features of the organ, and $y \in \{0, 1\}$ denotes the post-transplant outcome describing whether the graft fails or not; here, M_r, M_s, and M_o are the corresponding feature dimensions, and N is the total number of data points.

Objective. The goal is to train a model that takes $\mathbf{r}_i$, and $\mathbf{o}_i$ as input, such that it can accurately predict y_i and is also fair w.r.t. the sensitive features $\mathbf{s}_i$ in terms of the fairness metrics. Previous studies have shown that, in most times, improving fairness can harm the model performance [109]. Thus, a desirable model is expected to achieve a good tradeoff and maximize prediction performance and fairness simultaneously.

Fairness metric definitions. We adpot two fairness metrics DPD and EOD [108] in our experiments, defined as follow:

$$\text{DPD} = \text{diff}_s P(\hat{y} = 1|s), \tag{3.1}$$

$$\text{EOD} = \max[\text{diff}_s P(\hat{y} = 1 \mid s, y = 1), \text{diff}_s P(\hat{y} = 1 \mid s, y = 0)], \tag{3.2}$$

where diff_s specifies the difference between the largest and the smallest value among the ones across all s, $\hat{y}$ is the model prediction, where $y = 1$ represents the positive outcome, e.g., graft failed. Specifically, DPD measures the performance gap between the positive outcomes across all groups, while EOD measures the gap between true positive rates or false positive rates based on

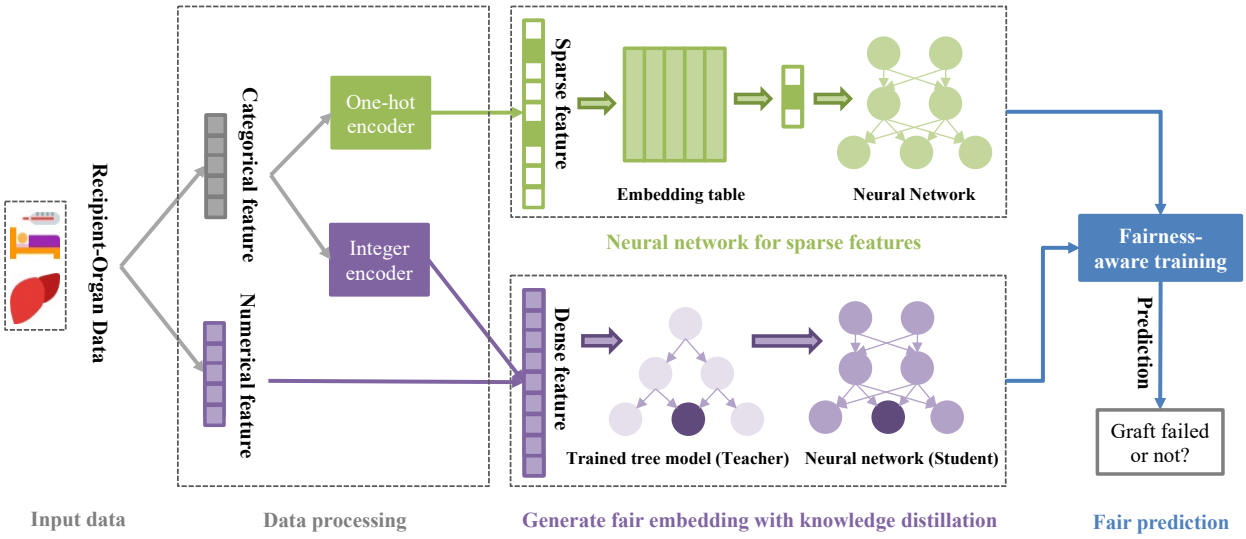


Figure 3.1: An overview of the workflow for graft failure prediction.

the confusion matrix across all groups.

3.5 Methodology

In this section, we propose our method for fairly predicting graft failure. Figure 5.1 shows the workflow, which consists of data processing, prediction model, and fairness-aware training. Firstly, we will introduce how we process the data to extract sparse and dense features from recipients and organs (Section 3.5.1). Then we introduce a tailored framework that takes the advantage of tree-based models and deep neural networks to make accurate predictions (Section 3.5.2). Finally, we present a two-step debiasing strategy to achieve fairness (Section 3.5.3).

3.5.1 Data Pre-processing

Following the data pre-processing practice in machine learning, we first impute the missing values. Specifically, we use zeros to replace the missing values for the numeric data. Then we identify the categorical features (i.e., the features that only have a fixed number of values) and numerical features from the recipient and organ features. For the categorical features, we employ two kinds of encoders, including a one-hot encoder that maps the raw features to one-hot sparse vectors, and an integer encoder which transforms the categorical features into numerical values,

where the latter are further concatenated with the original numerical features to serve as the final dense features.

3.5.2 Combining Deep Learning and Tree-based Model for Graft Failure Prediction

In previous works, tree-based methods such as random forest [93] have been adopted for graft failure prediction. However, the input space of graft failure prediction consists of both sparse categorical features and dense numerical ones. While tree-based methods often show strong performance on the dense features, they can hardly deal with the sparse features when the feature cardinality is high due to the curse of dimensionality [99]. In addition, it is quite difficult to incorporate fairness constraints into the tree-based methods. To tackle these challenges, we propose to combine deep learning and tree-based model for graft failure prediction. Our method is motivated by the success of DeepGBM [104] in recommendation tasks, where an embedding layer and a distillation network with distilled knowledge from a tree-based method are employed to handle the sparse and dense features, respectively. We will first elaborate on how we process the sparse and dense features, and then introduce the end-to-end training objective.

Sparse features. The sparse features from the recipient and the organ are combined and processed by a categorical neural network (CatNN) [110, 111], which is an embedding lookup layer that maps categorical indices to dense vectors, followed by feature interactions. Formally, given a recipient $\mathbf{r}$ and an organ $\mathbf{o}$, we denote the combined sparse features within $\mathbf{r}$ and $\mathbf{o}$ as $\mathbf{x}^s$. The embedding of a sparse feature can be denoted as

$$E_{\mathbf{V}_j}(x_j^s) = \text{embedding\_layer}(x_j^s, \mathbf{V}_j), \tag{3.3}$$

where x_j^s is the value of the j^{th} sparse feature of $\mathbf{x}^s$, $\mathbf{V}_j \in \mathbb{R}^{c \times d}$ stores all the trainable embedding vectors of the j^{th} sparse feature, and c and d are the cardinality and the dimension of the embedding table, respectively. Then a factorization machine (FM) is adopted to learn the first/second-order interactions of these features, denoted as $E_{\text{fm}}(\mathbf{x}^s)$, and a deep neural network is applied to learn the higher order interactions of these features, denoted as $E_{\text{deep}}(\mathbf{x}^s)$. For more details of FM and the

deep neural network, please refer to Eq. (2) and Eq. (3) in [104]. The output of FM and the neural network are summed to obtain the final sparse representations:

$$y_{\text{CatNN}}(\mathbf{x}^s) = E_{\text{fm}}(\mathbf{x}^s) + E_{\text{deep}}(\mathbf{x}^s) \tag{3.4}$$

Dense features. Similarly, we combine the dense features of recipient $\mathbf{r}$ and organ $\mathbf{o}$, denoted as $\mathbf{x}^d$. To take the advantage of the tree-based models in handling dense features, we train a neural network to distill the knowledge from a trained tree-based model [112]. This is not an easy task because the structures of the trees and neural networks are naturally different. Fortunately, Ke et al. [104] proposes an effective tree distillation strategy by distilling the clustering patterns of the leaf nodes. First, since tree-based methods often do not use all the features but instead greedily choose the useful features, we only select the used features of a tree to train the neural network. Formally, let NN_{dense} be the neural network for processing the dense features, $\mathbb{I}$ be the indices of the features that are used in the tree, and $\mathbf{x}^d[\mathbb{I}]$ denote the used dense features. Then NN_{dense} will take as input $\mathbf{x}^d[\mathbb{I}]$. Second, we train NN_{dense} by distilling the knowledge of how the tree partitions the data. Specifically, a tree-based model essentially partition the data into different clusters, where the data in the same leaf node belong to the same cluster. We train NN_{dense} to distill the knowledge from such tree structure by minimizing the following loss function:

$$L_{\text{KD}} = \sum_{i=1}^{N} \text{mse}(\text{NN}_{\text{dense}}(\mathbf{x}_i^d[\mathbb{I}]), \mathbf{c}_i), \tag{3.5}$$

where $\mathbf{c}_i$ is the one-hot encoded cluster of the i^{th} instance, cross-entropy$(\cdot, \cdot)$ is the cross-entropy loss. Due to the strong expressiveness of deep neural networks, NN_{dense} can well approximate the tree structure. Given NN_{dense}, the dense representations can be obtained by

$$y_{\text{KD}}(\mathbf{x}^d) = \text{NN}_{\text{dense}}(\mathbf{x}_i^d[\mathbb{I}]) \times \mathbf{q}, \tag{3.6}$$

where $\mathbf{q}$ is the leaf values of the tree. For multiple trees, we learn leaf embedding to reduce the

dimension of $\mathbf{c}_i$ and group the trees to reduce the number of neural networks following [104]. The leaf embeddings are trained independently based on the tree-based model and will be used as dense representations in the end-to-end training.

End-to-end training. The final output is obtained by combining sparse and dense representations, given as

$$\hat{y}(\mathbf{x}) = \sigma\left(w_1 \times y_{\text{KD}}(\mathbf{x}^d) + w_2 \times y_{\text{CatNN}}(\mathbf{x}^s)\right), \tag{3.7}$$

where w_1 and w_2 are trainable parameters to balance the two representations, $\mathbf{x}$ is combined sparse and dense features from $\mathbf{r}$ and $\mathbf{o}$, and $\sigma(\cdot)$ is the transformation function, such as Sigmoid. Finally, we can train the model in an end-to-end fashion with the following loss:

$$L = \sum_{i=1}^{N} \text{cross-entropy}(\hat{y}(\mathbf{x}), y). \tag{3.8}$$

3.5.3 Bias Mitigation

This subsection proposes a two-step debiasing strategy to mitigate the unfairness in the distillation stage and the final training stage, where the former focuses on the bias inherited from the tree-based model when performing knowledge distillation, and the latter aims to achieve fairness in the end-to-end training.

Fairness loss. Motivated by the successes of in-processing methods in debiasing machine learning models [102, 103], we use fairness loss to incorporate demographic parity in model training. Specifically, we propose the following loss:

$$\text{fairness-loss}(\hat{y}, s_{\text{maj}}) = (E[\hat{y}] - E[\hat{y}|s_{\text{maj}}])^2 \tag{3.9}$$

where $\hat{y}$ is the prediction, s_{maj} is the majority group, $E[\hat{y}]$ is the expected prediction regardless of the sensitive groups, $E[\hat{y}|s_{\text{maj}}]$ is the expected prediction of the majority group. The key idea is to enforce all the sensitive attributes to have similar prediction distributions like the majority group. In training, $E[\hat{y}]$ and $E[\hat{y}|s_{\text{maj}}]$ can be approximated with a batch of data. Thus, Eq.3.9 can be

naturally applied to the min-batch training of deep learning models.

Two-step debiasing. We propose to debias both the categorical neural network and the network for dense features. In the first step, we achieve fair knowledge distillation by plugging in Eq. 3.9 into Eq. 3.5:

$$L_{\text{KD}} = \sum_{i=1}^{N} \text{mse}(\text{NN}_{\text{dense}}(\mathbf{x}_i^d[\mathbb{I}]), \mathbf{c}_i) + \alpha_{\text{KG}} \times \text{fairness-loss}(y_{\text{KD}}(\mathbf{x}^d), s_{\text{maj}}), \tag{3.10}$$

where α_{KG} is a hyperparameter to balance prediction performance and fairness. In the second step, we incorporate the fairness constraint into the end-to-end training. Specifically, we similarly debias Eq. 3.8 with

$$L = \sum_{i=1}^{N} \text{cross-entropy}(\hat{y}(\mathbf{x}), y) + \alpha \times \text{fairness-loss}(\hat{y}(\mathbf{x}), s_{\text{maj}}), \tag{3.11}$$

where α is a balancing hyperparameter. These two debiasing steps complement each other towards fair final predictions, where the first step focuses on the dense representations which serve as the input of the end-to-end training, and the second step debiases the CatNN and the embedding tables.

3.6 Experiments

In this section, we perform analysis on the datasets and conduct experiments to evaluate the proposed framework. We mainly focus on the following research questions:

- **RQ1:** Does MELD score align with the post-transplant outcomes for different races and genders (Section 3.6.1)?
- **RQ2:** Can the proposed framework makes accurate and fair predictions of the graft failure (Section 3.6.2)?
- **RQ3:** How does each stage of debiasing contribute to the fair predictions (Section 3.6.3)?

Race	MELD-score		Number of people		Receiving rate		Graft failure rate	
	Male	Female	Male	Female	Male	Female	Male	Female
I	20.05852	20.36856	89700	49815	0.56405	0.49941	0.32300	0.29592
II	21.56156	22.74271	10209	8131	0.60251	0.57004	0.36482	0.34067
III	21.14621	21.49130	18282	12074	0.51400	0.47176	0.27754	0.26194
IV	17.82069	19.35089	5878	3095	0.53215	0.53312	0.24616	0.25818
V	22.13557	23.38609	686	676	0.54082	0.44822	0.28032	0.28713
VI	22.20161	19.42105	248	152	0.50000	0.55263	0.26613	0.29762
VII	19.80120	20.59470	664	491	0.63253	0.60285	0.26905	0.27703

Table 3.1: Statistical information from liver transplant dataset

3.6.1 Statistical Analysis of the Liver Transplant Dataset

For statistical analysis, we select the patients from 7 main races and 2 genders with recorded MELD scores. There are 14 subgroups intersected by races and genders. The average MELD score and the total number of people of each divided subgroup are calculated in Table 3.1. We can observe there are obvious gaps between each subgroups' MELD score. The minimum MELD score is only 76.2% of the maximum MELD score. Additionally, the size of majority races is much larger than minority races.

Due to the variety existing in each subgroup's MELD score and group size, we take two perspectives that correspond to the organ receiving rate and graft failure rate to better investigate the liver transplant task.

- **Organ receiving rate (ORR)** represents the chance of a group of patients on the waiting list to receive organs. We use the accumulated samples on the waiting list recorded receiving liver transplants based on the MELD score as the number of receiving patients for each subgroup, denoted as n_r. The receiving rate is calculated by dividing them by the total number of people in this group registered on the waiting list, denoted as n_w.

- **Graft failure rate(GFR)** reflects the percentage of graft failed for a group of patients who have received the transplant liver. We count the recorded graft failure samples, denoted as n_f, and divide it by the number of patients who already received organs, denoted as

n_r. These two metrics provide an intuitive measure to explore organ assigning and post-transplant outcomes, which are the most essential stages in the liver transplant task.

Formally, these two metrics can be denoted as:

$$ORR = \frac{n_r}{n_w}; GFR = \frac{n_f}{n_r}. \tag{3.12}$$

For the organ receiving rate, we can observe from the organ receiving rate column in Table 3.1 that obvious gaps exist between organ receiving rate of different subgroups. The highest receiving rate is 0.63253 of subgroup interacted by race VII and male, while the lowest receiving rate is 0.44822 of subgroup interacted by race V and female. However, the latter subgroup's average MELD score is significantly higher than the former subgroup. This means the latter subgroup should have higher priority on the waiting list, which contradicts our findings from the observed data. This phenomenon indicates the MELD score does not align with organ receiving rate. As presented in Table 3.2, the Pearson correlation between organ receiving rate and MELD score is -0.32376. This means the MELD score has no close relation with the organ receiving rate from the group-level analysis.

For the graft failure rate, we can observe that notable gaps exist between graft failure rates across different subgroups as shown in the graft failure rate column in Table 3.1. The subgroup with the highest graft failure rate is the male race II subgroup with a 0.36482 graft failure rate. The lowest graft failure rate exists in race IV male groups, which is 0.24616. The MELD score of the latter subgroup is smaller than the former subgroup. It suggests better pre-transplant medical condition, which may explain the lower graft failure rate. To quantify and further look into the relations between MELD score and graft failure rate, we calculate the Pearson correlation between them. The Pearson correlation is still very weak as shown in Table 3.2. It implies the MELD-score cannot indicate group-level graft failure rate at the post-transplant stage.

To summarize, we analyze two main components of organ transplant statistically, the organ assignment and post-transplant outcome. The results show remarkable gaps across subgroups in

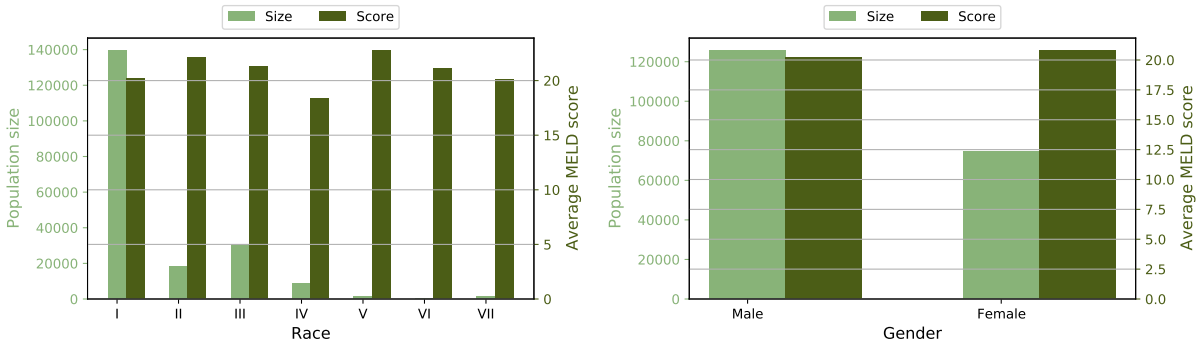


Figure 3.2: Population size and average MELD score across races and genders.

	MELD-score	Population size
Organ receiving rate	-0.32376	-0.02243
Graft failure rate	0.36653	0.33444

Table 3.2: Pearson correlation between demographic information and liver transplant metrics

both two components, which indicates a strong bias existing in organ transplant systems.

3.6.2 Results of Prediction and Fairness Performance

We conduct experiments to compare the prediction and fairness performance of the proposed method with multiple baseline methods (Table 3.3). The key observation is that the proposed model can provide competitive prediction performance with less bias across subgroups.

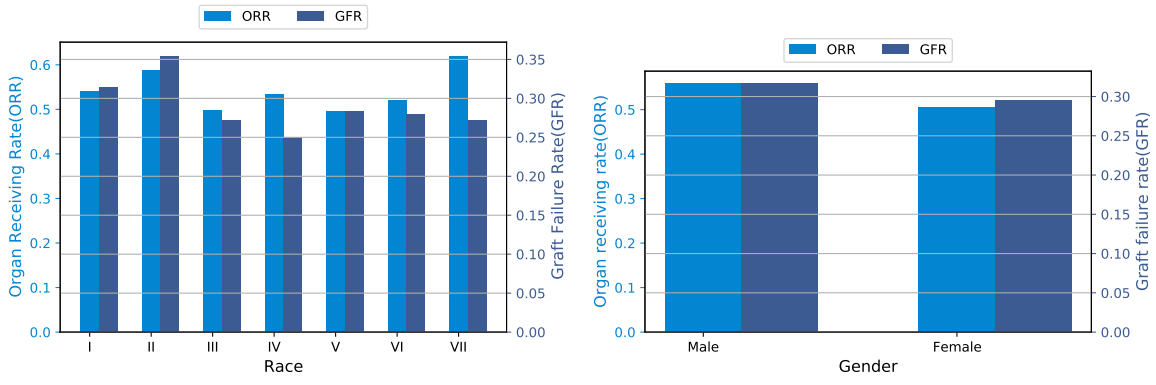


Figure 3.3: Average organ receiving rate and graft failure rate across races and genders.

Model	Sensitive attribute: Race			Sensitive attribute: Gender		
	ROC AUC	DPD	EOD	ROC AUC	DPD	EOD
MELD-score	0.505±0.000	—	—	0.505±0.000	—	—
Logistic Regression	0.777±0.000	0.648±0.017	0.834±0.007	0.777±0.000	0.021±0.000	0.033±0.001
Random forest	0.804±0.000	0.630±0.030	0.703±0.047	0.804±0.000	0.020±0.001	0.036±0.001
GBDT	0.809±0.000	0.637±0.027	0.713±0.033	0.809±0.000	0.017±0.000	0.031±0.001
W/o first-step	0.793±0.000	**0.596±0.022**	0.687±0.038	0.792±0.000	0.016±0.002	0.027±0.002
W/o second-step	0.793±0.001	0.616±0.041	0.745±0.076	0.793±0.001	0.014±0.007	0.026±0.009
Ours	0.792±0.000	**0.597±0.015**	**0.662±0.029**	0.793±0.001	**0.011±0.001**	**0.022±0.003**

Table 3.3: Comparison of prediction and fairness performance on graft failure prediction

Compared with the MELD score, we observe that machine learning models show much stronger prediction capability of graft failure. The poor graft failure prediction performance of MELD score aligns with the weak correlations between MELD score and graft failure rate from statistic analysis in Table 3.2. The tree model has better and less biased prediction performance than linear model. This may be caused by the tree model's internal selection of features, which could implicitly omit some features with bias.

Compared with baseline machine learning methods, when the sensitive attribute is race, the proposed method can significantly debias the prediction with only 2.1% decreases of ROC AUC, while the two fairness metrics decrease by 5.5% averagely. As for gender, the ROC AUC decreases only 2.0%, however, the two fairness metrics decrease by 32.2% on average. Recall that the parity loss we applied is based on the demographic parity. In Table 3.3, we observe improvement not only on DPD but also on EOD. This can validate the effectiveness of our debiasing method, which can generally mitigate the unfairness issues.

3.6.3 Ablation Study

To validate the effectiveness of our two-step debiasing strategy, we conduct ablation study to investigate the contribution of each component. From Table 3.3, we observe that by only adding the debiasing method in knowledge distillation step (first step), the proposed model can only improve the DPD metrics. When only debiasing the end-to-end training step, both fairness metrics improve to some extent. The model achieves the best debiasing performance when the two debiasing steps

are combined. This is because the knowledge-distilled embedding and end-to-end training are interleaved, which verifies the necessity of the two-step debiasing strategy.

3.7 Conclusion

This work aims at fair graft failure prediction for developing unbiased organ assigning strategy. A two-step knowledge distillation framework is built to encourage fair prediction towards different groups while preserving competitive performance. The fair and competitive prediction performance of the whole framework has been experimentally signified on graft failure prediction dataset. In the future, we will investigate and identify more fairness issues such as intersection fairness problem. Furthermore, we will continue designing debiasing methods for liver transplant tasks, fairness problem discovered from the liver transplant task can also inspire research on other organ transplant systems.

4. MEDICAL HAI WITH REINFORCEMENT LEARNING BASED FAIR RANKING

4.1 Overview

Liver transplant is a widely adopted solution for many end-stage liver diseases. In practice, it is important to match liver organs with patients fairly. Our goal is to train an agent to rank the patients for each organ to optimize three objectives: (1). post-transplant metrics (e.g., graft failure rate and survival rate), (2). fairness across sensitive groups (e.g., gender and race), and (3). fairness across individuals. We formulate organ allocation as a ranking problem and define two fairness metrics from group and individual perspectives. Our approach FairAlloc learns an agent to jointly optimize the ranking and fairness metrics. We train FairAlloc and compare it with six baselines on the organ-patient data collected from Organ Procurement and Transplantation Network (OPTN). FairAlloc significantly improves group and individual fairness metrics by up to 37.9% and 39.9%, respectively, while achieving a competitive ranking performance based on post-transplant metrics.

4.2 Introduction

Liver transplant has become a widely used and the only life-saving therapy for patients who suffer from end-stage liver disease. Organ allocation is a critical decision for this treatment to match organs with patients: given an available organ from a donor, how should we rank the patients on the waiting list? From the guidelines of the U.S. Department of Health & Human Services [113], utility and justice are the major principles to be balanced in organ allocation, where the former refers to the maximization of the important medical factors, e.g., patient survival time, and the latter refers to the fairness of the allocation for the patients on the waiting list, which has been rarely studied in the literature. In organ transplants, fairness is synonymous with justice to a large extent. Fairness means the patients are treated with equity regardless of their social class, race, gender, etc. Meanwhile, the current allocation strategy used in liver transplants is based on the MELD score, calculated based on the lab test results of patients on the waiting list. The higher the score, the higher the priority of the patient to receive an organ. However, the MELD score does not

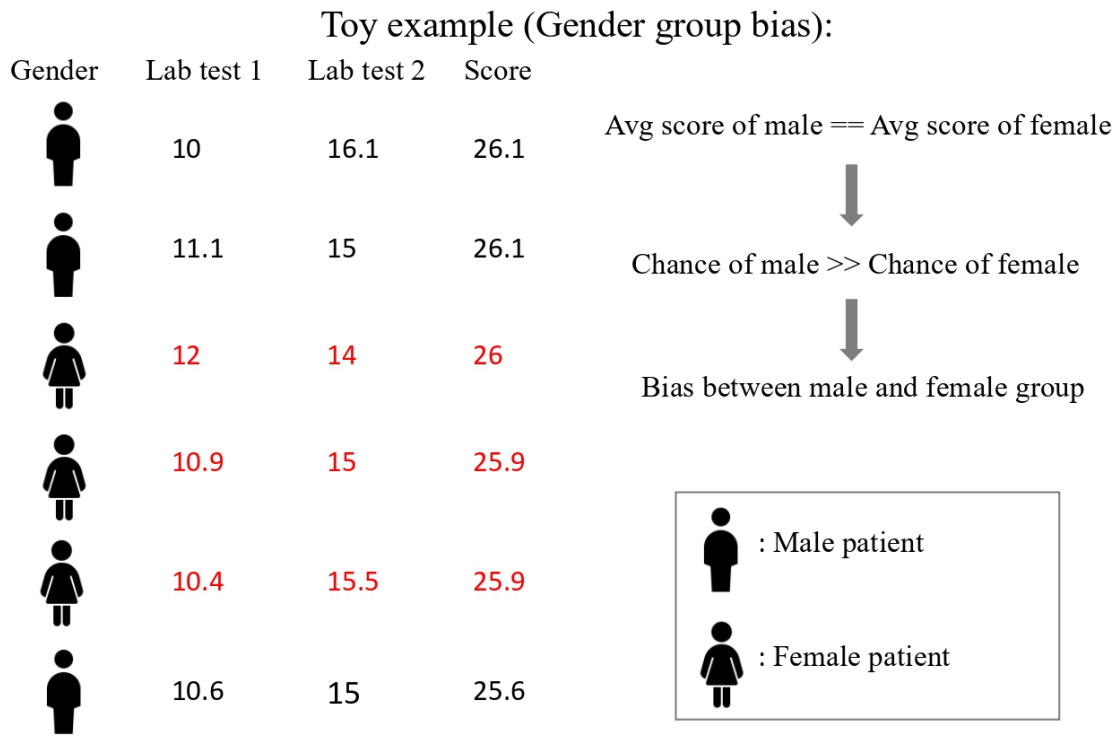


Figure 4.1: An illustrative example of unfairness in organ allocation (the numerical value in this figure is just for illustration with no real-world meanings).

explicitly consider justice in its calculation. It is an open challenge to allocate the organs to achieve both utility and justice since there are conflicts between them under many circumstances [114], which makes the balance between them very important.

Organ allocation aims to rank the patients on the waiting list, which is currently based on the MELD score. The list of the ranked patients serves as an important criterion for the allocation. In practice, the top-ranked patient is not guaranteed a donor's liver; some other factors can also affect doctors decisions [115], such as the age of the patient, the distance between donor and recipient, etc.

We now use an illustrative example to discuss the fairness issue in the patient ranking problem. In Figure 4.1, the average score of the male group is the same as that of the female group. Whereas the top 2 patients are all males. As a result, the male group has a significantly higher chance of being selected because the doctors may only focus on the top-ranked patients. Beyond gender,

the fairness issue may also exist in other sensitive groups (e.g., races) and even individuals (i.e., a patient with a slightly lower score could be ranked significantly lower). However, enabling a fair ranking is challenging because (1). the ranking procedure is non-differentiable, which means many existing gradient-based debiasing methods can not be applied [12, 102], and (2). It is hard to balance fairness and utility and achieve fairness at both group and individual levels. In the following, we present FairAlloc to address these challenges.

In this work, we consider organ allocation fairness from the group perspective (i.e., certain groups of people with the same characteristics, such as gender and race) and the individual perspective (i.e., each pair of individual patients). Then we present FairAlloc, a general framework for fair organ allocation based on deep reinforcement learning. FairAlloc trains a neural network with policy gradients to predict the score of each patient to optimize the ranking reward defined by utility, group fairness, and individual fairness. We apply FairAlloc to the organ allocation datasets from Organ Procurement and Transplantation Network (OPTN) [116] to provide potential suggestions for doctors to make decisions.

4.3 Background and Significance

4.3.1 Machine Learning for Organ Transplant

Compared to traditional clinical analysis methods, which are mainly based on statistics, machine learning (ML) models are trained on a large volume of data and can provide a more precise prediction of the medical indicators in organ transplant, e.g., mortality, transplant successful rate, etc. This can potentially increase the success rate of the transplant surgery since the ML model could better pair patients with organs. ML has recently shown promising results in organ allocation. Some previous works [117, 118, 119] employ ML for precise prediction of various factors such as post-transplant living rate, graft failure rate, etc. For instance, Rhu et al. [118] use the cox regression model with medical indicators such as total bilirubin to predict graft failure after a liver transplant. Liu et al. [117] adopt random forests to predict short-term survival rates based on blood test results. Additionally, some initial studies [90, 91] leverage ML to directly allocate organs.

Xu et al. [120] propose to learn the matching representations for organ allocation. Berrevoets et al. [121] design the allocation algorithm by applying queuing theory and synthetic control. However, existing works [122] mainly focus on maximizing utility, and the fairness issue caused by ML in the allocation was rarely studied in the literature.

4.3.2 Fairness in ML

When ML is applied in the organ transplant scenario, the prediction of ML will serve as an important indicator in clinical decisions. So, we need to ensure the predictions are fair across different groups. Most existing ML debiasing techniques incorporate fairness objectives as part of the loss functions [11] so that the fairness objectives can be optimized with gradient descent. Unfortunately, the fairness objective in the patient ranking problem is non-differentiable, so these methods can not be directly applied. Some recent studies have investigated fairness in ranking problems. Singh et al. [123] approach fairness in ranking from the exposure allocation perspective and propose a reinforcement learning framework [124] to learn fair ranking policies. However, unlike the standard ranking problem, patient ranking needs to consider lots of factors from both organs and patients. Moreover, the post-transplant metrics are delayed feedback that could only be available years after the transplant, and thus they are hard to predict.

4.3.3 Fairness in Organ Transplantation

Fairness in organ transplantation means the patients are allocated organs based on their medical conditions rather than their social attributes such as gender, race, social class, etc. The bias in organ transplantation occurs when the patients do not have their deserved chance of receiving the organ based on their medical condition. This bias can be caused by the patients being discriminated against by their non-medical attributes, e.g., race, gender, etc. Note that we are pursuing equity rather than equality in organ transplantation. The former suggests that patients are given organs based on their medical condition. The latter indicates that the patients are given organs equally regardless of their conditions, which is not desirable in organ transplants. The fairness issue in organ transplants has gained considerable attention from the healthcare community [95, 96]. Parent

et al. [96] identify the geographic bias of liver transplant allocation and propose a new region districting model to alleviate this bias. However, mitigating the bias in organ allocation has rarely been studied for ML models. Our previous work [61] proposes a fair ML prediction framework for organ transplants. Whereas this framework only predicts the graft status after transplant, which are only intermediate results before ranking. In practice, doctors may only look into the top-ranked patients in their decision so that the higher-ranked patients will have a significant advantage. In this work, we aim to develop a framework to directly rank the patients fairly.

4.4 Proposed Methods

This section first introduces the motivation and challenges of fair ranking for organ allocation. Then we mathematically formulate the ranking problem, followed by an introduction of utility and fairness metrics. Further, we elaborate on the FairAlloc framework.

4.4.1 Problem Formulation and Notations

Organ allocation can be formulated as a ranking problem. Given an organ O and n patients $P = \langle p_1, p_2, \ldots, p_n \rangle$, where p_i denotes a patient ($i \in \{1, 2, \ldots, n\}$), an ML model is expected to output a permutation $\pi = \langle \pi(p_1), \pi(p_2), \ldots, \pi(p_n) \rangle$, where $\pi(p_i)$ denotes the rank of patient p_i in the permutation. Given a training set $X_{\text{train}} = \{\langle O_j, P_j \rangle\}_{j=1}^{m}$ of size m, the goal of organ allocation is to learn an ML permutation policy π such that it can optimize utility and fairness ranking metrics on an unseen testing set X_{test}.

4.4.2 Ranking Metrics in Organ Allocation

The ranking metrics are defined based on a given scoring criterion that quantifies the medical conditions. In previous work, some common criteria have been used, such as CTP score [125], MELD score [126], and post-transplant scores (e.g., graft failure rate and survival rate) [127]. In our work, the scores are predicted by a neural network with the donor and patient features as the input (we will elaborate in the later sections). Based on the scores, we first introduce the exposure rate, which defines the chance of receiving an organ for a specific ranking position. Let r be a ranking position, where $1 \leq r \leq n$. Following [124], we define the exposure rate of position r as

$$Ex(r) = \frac{1}{\log_2(r+1)} \tag{4.1}$$

Intuitively, the highly ranked patients will have a higher chance of being selected. A doctor will often go through the list starting from the top-ranked one until a suitable patient is found. If a patient is ranked high, he/she will have a higher chance of being assigned an organ since he/she is more likely to be seen by the doctor. Clinically, the exposure rate is used to measure a patients chance of receiving an organ in the liver transplant scenario. Now we introduce a utility metric and two fairness metrics.

Utility metric. We adopt Normalized Discounted Cumulative Gain (NDCG) as the utility metric, which has been commonly used in ranking tasks [128]. We first define Discounted Cumulative Gain (DCG) as:

$$DCG = \sum_{i=1}^{n} \frac{y_i}{\log_2(r_i+1)} \tag{4.2}$$

where y_i is the groundtruth score of the patient that is ranked r_i. Ideally, if the patients are perfectly ranked (i.e., a patient with a higher score will always be ranked higher), the DCG metric will be maximized, which is defined as Ideal DCG (IDCG). In practice, we often adopt Normalized DCG, i.e., NDCG, which is defined as follows:

$$NDCG = \frac{DCG}{IDCG} \tag{4.3}$$

Note that NDCG $\in [0, 1]$, where a larger value indicates a better utility.

Group-level fairness metric. At the group level [129], we demand that the patients exposure rates from different groups (e.g., gender and race) align with their scores. For example, if the female group has a higher average score, then we expect the female group should also have a higher average exposure rate. Formally, we define the group bias as follows:

$$Bias_{grp} = \frac{\text{Range}_s(\overline{\text{Ex}_s(r)})}{\text{Range}_s(\overline{y_s})}, \tag{4.4}$$

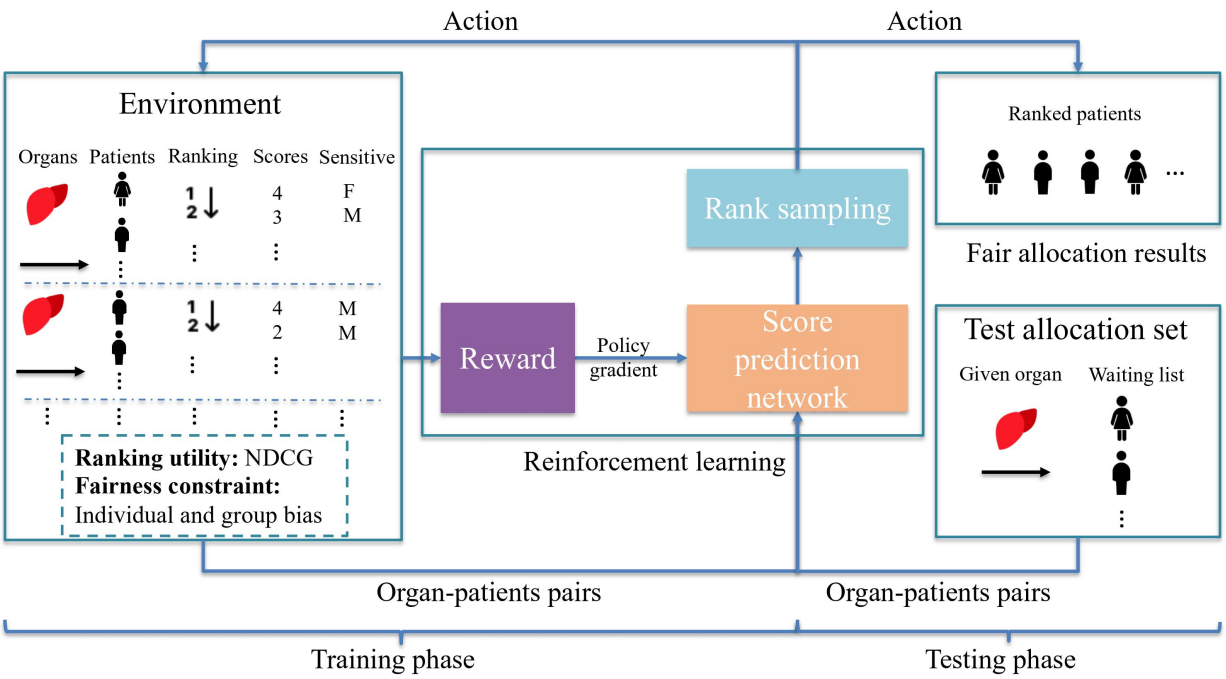


Figure 4.2: An overview of FairAlloc. In the training phase, we sample patients rankings with the score prediction network and update the network based on rewards. In the testing phase, we apply the trained network to rank patients given an unseen organ.

where $\overline{Ex_s}(r)$ is the average exposure rate for the patient in group s, $\overline{y_s}$ is the average score for the patients in group s, and $\text{Range}_s(\cdot)$ is the difference between the maximum value and the minimum value across groups. A larger Bias_{grp} suggests a higher group bias. Our ranking algorithm aims to minimize Bias_{grp}.

Individual-level fairness metric. The individual-level fairness [130] demands that the exposure rate of each patient align with his/her score. For instance, if two patients have very similar scores, they should also have similar exposure rates. Formally we define individual bias as

$$Bias_{ind} = \frac{\text{Range}_s(\text{Ex}_s(r))}{\text{Range}_s(y_s)}, \tag{4.5}$$

where $E(x_i(r))$ is the exposure rate for each individual patient, y_i is the score for each individual patient, and $\text{Range}_i(\cdot)$ is the difference between the maximum value and the minimum value across all patients. Similarly, we aim to minimize $Bias_{\text{ind}}$.

4.4.3 Learning Fair Allocation Policy Framework

It is challenging to optimize the above three metrics because the ranking procedure is non-differentiable. To tackle the challenge, we design FairAlloc to optimize the three utility/fairness metrics with deep reinforcement learning. FairAlloc consists of three modules: (1). a score prediction module that predicts scores for unseen organ-patient pairs, (2). a rank-sampling module that samples permutations of the patients, and (3). a policy gradient module that trains the score prediction network based on rewards. An overview of FairAlloc is presented in Figure 4.2.

Score prediction with neural networks. For unseen organ-patient pairs, we use a neural network to estimate their scores, which will be used for ranking. Specifically, we pre-train a multilayer perceptron (MLP) network using the training data and apply this network to the testing data. The input of the MLP is the concatenated features of the organs and patients. The score prediction is naturally a regression task. Thus, we train the MLP with the mean squared error (MSE) loss:

$$L_{\text{score}} = \sum_{j=1}^{m} \left(y_j - \text{MLP}(x_j)\right)^2 \tag{4.6}$$

where y_j is the true score, and x_j is the concatenated feature. In practice, we use mini-batch training [131]. More details are provided in the appendices. Once trained, the MLP can be applied to unseen organ-patient pairs to predict the score:

$$y' = \text{MLP}(x) \tag{4.7}$$

where x is the organ-patient feature, y' is the predicted score. Note that the initial score prediction network may not achieve the best-ranking performance. Thus, the network will be fine-tuned later with the end-to-end ranking reward.

Rank sampling. Inspired by [124], we use a Plackett-Luce [132] based method to generate rankings from the predicted scores. Starting from the highest-ranked position, we sample the remaining patients with probabilities proportional to their predicted scores. We repeat the sampling

process and insert patients one by one into the list until there is no patient left. Formally, for ranking position r, the probability of a patient p_i being sampled is

$$P(\pi(p_i) = r) = \frac{e^{y'_i}}{\sum_{j=1}^{n} I(p_j) e^{y'_j}}, \tag{4.8}$$

where $I(p_j)$ is a binary indicator suggesting whether p_j is already in the list. Specifically, $I(p_j) = 1$ if p_j is not in the list, and $I(p_j) = 0$ otherwise. This sampling strategy converts scores to rankings. The probability for a specific ranking perturbation can be obtained by

$$P(\pi) = P(\pi(p_1) = r_1, \ldots, \pi(p_n) = r_n) = \prod_{i=1}^{n} P(\pi(p_i) = r_i). \tag{4.9}$$

Reward optimization with policy gradient. Now we aim to adjust the ranking to optimize the utility/fairness metrics. To achieve this, we apply a policy gradient [133] to fine-tune the score prediction network to maximize the ranking reward. The reward is a combination of the three ranking metrics:

$$R = \alpha \cdot NDCG - \beta \cdot Bias_{ind} - \gamma \cdot Bias_{grp}, \tag{4.10}$$

where α, β, γ are hyperparameters. NDCG, $Bias_{ind}$, and $Bias_{grp}$ are the three utility/fairness metrics. Let θ be the parameters of the score prediction network MLP.

Our objective is to maximize the expected reward:

$$J = E[R]. \tag{4.11}$$

Following the policy gradient theorem [133], we can calculate the gradient of J with respect to θ:

$$\nabla_\theta J = \nabla_\theta E[R] = E[\nabla_\theta \log P(\pi) R], \tag{4.12}$$

where $P(\pi)$ is defined in the rank sampling module. Note that the gradient of $P(\pi)$ can be

	Sensitive attribute: Gender			Sensitive attribute: Race		
Model	**Ind bias**	**Grp bias**	**NDCG**	**Ind bias**	**Grp bias**	**NDCG**
MLP	0.475±0.052	0.063±0.009	0.899±0.045	0.475±0.052	0.153±0.016	0.899±0.045
GBDT	0.426±0.041	0.061±0.011	0.922±0.038	0.426±0.041	0.136±0.015	0.922±0.038
LR	0.428±0.040	0.061±0.008	0.923±0.037	0.428±0.040	0.141±0.009	0.923±0.037
RankNet	0.467±0.063	0.061±0.008	0.901±0.049	0.467±0.063	0.143±0.019	0.901±0.049
LambdaRank	0.469±0.057	0.063±0.010	0.901±0.049	0.469±0.057	0.151±0.014	0.901±0.049
DebiasedMLP	0.479±0.068	0.066±0.008	0.897±0.055	0.477±0.077	0.145±0.020	0.900±0.058
Ours	0.288±0.038	0.041±0.009	0.905±0.047	0.288±0.036	0.102±0.018	0.905±0.047

Table 4.1: Overall performance with graft status as score criteria.

backpropagated to the score prediction network so that MLP can be updated. However, this update can be unstable because of high variance. Thus, we further use a baseline to reduce variance [134]:

$$\nabla_\theta J = E[\nabla_\theta \log P(\pi)(R - B)], \tag{4.13}$$

where B is the average reward for the current matching organ.

4.5 Experiments

We design experiments to answer the following research questions:

- **RQ1:** Can FairAlloc mitigate the bias in organ allocation while maintaining comparable ranking performance?
- **RQ2:** How will each of the proposed fairness-related rewards contribute to the performance of FairAlloc?
- **RQ3:** How will FairAlloc perform under different hyper-parameters settings?

4.5.1 Performance Comparison

To answer RQ1, we compare FairAlloc with the baselines. Quantitatively, Tables 4.1 and 4.2 report the overall performances using the graft status and the survival time as the scoring criterion, respectively. We make several observations:

	Sensitive attribute: Gender			Sensitive attribute: Race		
Model	**Ind bias**	**Grp bias**	**NDCG**	**Ind bias**	**Grp bias**	**NDCG**
MLP	0.468±0.075	0.049±0.017	0.766±0.031	0.468±0.075	0.129±0.034	0.766±0.031
GBDT	0.421±0.077	0.048±0.017	0.782±0.037	0.421±0.077	0.116±0.034	0.782±0.037
LR	0.421±0.077	0.047±0.015	0.784±0.044	0.421±0.077	0.116±0.026	0.784±0.044
RankNet	0.448±0.070	0.048±0.017	0.771±0.038	0.448±0.070	0.120±0.027	0.771±0.038
LambdaRank	0.449±0.064	0.048±0.010	0.774±0.039	0.449±0.064	0.127±0.035	0.774±0.039
DebiasedMLP	0.459±0.088	0.048±0.018	0.761±0.041	0.458±0.080	0.120±0.031	0.767±0.052
Ours	0.341±0.024	0.037±0.006	0.775±0.040	0.339±0.024	0.099±0.011	0.774±0.040

Table 4.2: Overall performance with survival time as score criteria.

The rankings generated by the standard ML methods have very high individual and group biases. For the survival time criterion, while NDCG can reach up to 0.784, there are significant group biases (up to 0.049 for the gender group and 0.129 for the race group) and individual bias (up to 0.468). Similarly, for the graft status, the individual bias can be as large as 0.479, while the gender and race group biases are up to 0.066 and 0.153, respectively. The results suggest that the standard ML methods suffer from a significant unfairness issue.

FairAlloc achieves lower group and individual biases and competitive NDCG. (1). The gender (race) group bias has been reduced by 37.9% (33.3%) and 24.5% (23.3%) for the graft status and the survival time, respectively. (2). The individual bias is reduced by 39.9% for graft status and 27.6% for survival time. 3) Meanwhile, FairAlloc has only 0.017 and 0.009 NDCG drops for graft status and survival time, respectively, which is greatly outweighed by the fairness improvement it provides. Overall, FairAlloc significantly reduces group and individual biases and retains the ranking performance.

Simply debiasing the scores can not reduce the individual bias and group bias compared to the standard MLP model. For the graft status criterion, although the Debiased MLP decreases the race group bias from 0.153 to 0.145, the individual bias increases from 0.475 to 0.479, and the gender group bias increases from 0.063 to 0.066. For the survival time, while the individual bias and group bias are reduced, the NDCG is negatively and significantly affected, decreasing from 0.766 to 0.761. The unsatisfactory performance of the Debiased MLP suggests a gap between scoring and ranking, i.e., simply debiasing the prediction score may not necessarily lead to a fair

Number of ranking patients	Score gap	Ranking gap of baseline	Ranking gap of FairAlloc	NDCG of baseline	NDCG of FairAlloc
30	0.050	7.05	0.000	0.938	0.909
14	0.133	5.13	0.467	0.912	0.891
12	0.100	2.40	0.600	0.836	0.778
30	0.056	5.83	0.556	0.770	0.821

Table 4.3: Ranking patients for the case study.

ranking. In contrast, our FairAlloc optimizes the ranking with reinforcement learning, which leads to much better performance.

The bias across races is larger than the bias across gender. When the graft status is used as scoring criteria, the race group bias of MLP is 0.153, which is 2.4 times larger than the gender group bias. For the survival time, the race group bias of MLP is 0.129, which is 2.6 times compared to the gender group bias. A possible explanation is that some race groups are very rare, while the populations of the male and female groups are relatively more balanced. Some rare race groups could have only very few patients, so they are significantly underrepresented. This observation suggests that we may need to pay more attention to race unfairness in the future [135].

Qualitatively, we analyze the ranked patients of different organs by comparing FairAlloc with MLP on gender groups. We are particularly interested in how the algorithms rank the patients when different gender groups have similar scores. Table 5.5.4 lists the patient ranking of four organs, where the score gaps are very small. We make the following observations:

The standard ML methods can cause a significant ranking bias even when the score gap is small. The ranking gap between the male and female groups of the baseline is small for all four cases, while the ranking gap is large. In particular, for the first organ, the ranking gap reaches 7.05, while the score gap is 0.050. This suggests that the scoring bias does not align with the ranking bias, which explains why simply devising the score does not work.

FairAlloc significantly reduces the ranking gap. Compared with the baseline, FairAlloc reduces the ranking gap by at least 4X for all four organs. Notably, for the first organ, the ranking gap is reduced from 7.05 to 0.00.

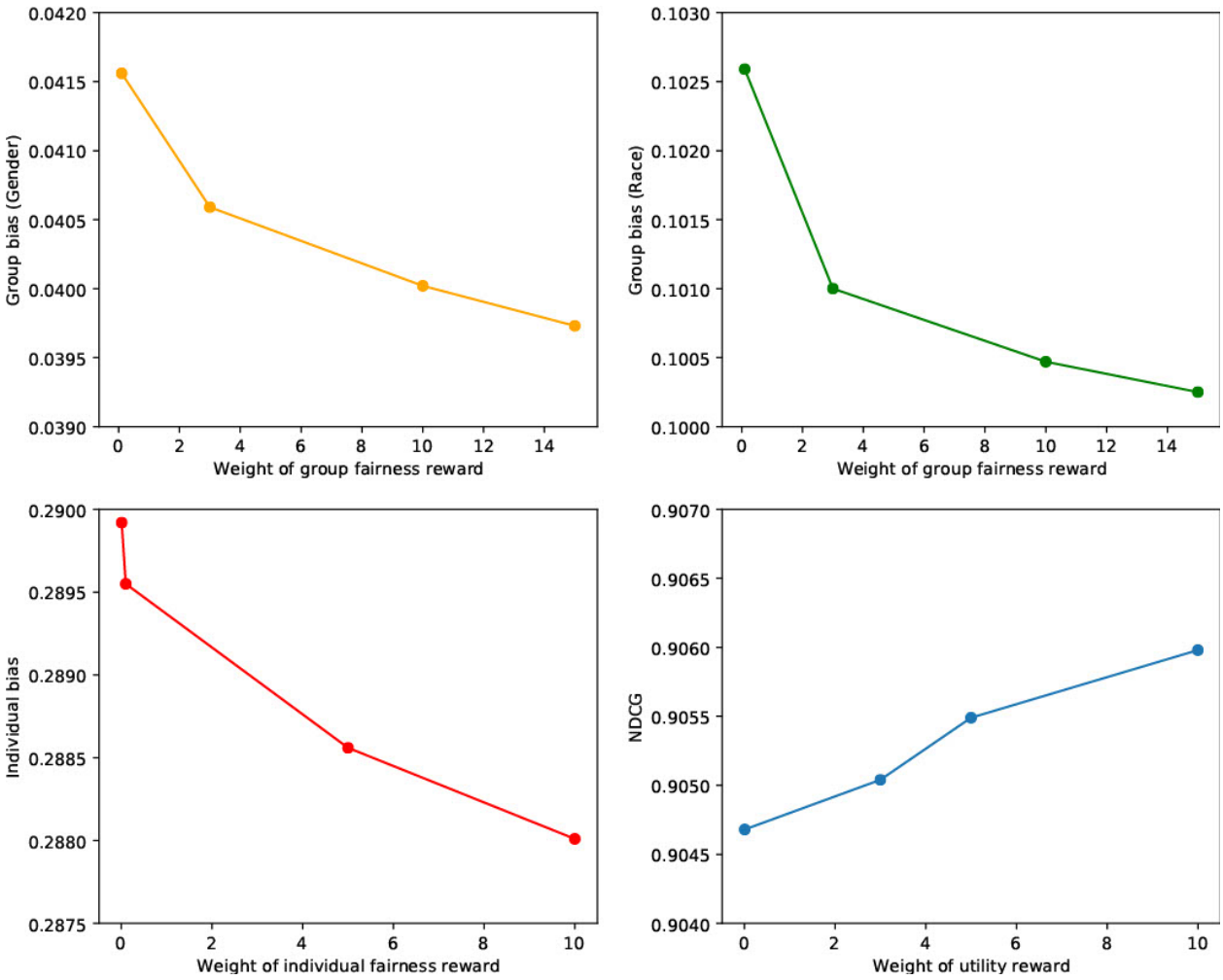


Figure 4.3: Utility and fairness results with graft status as score criteria under different hyper-parameters.

FairAlloc has a competitive ranking performance. The NDCG gap between FairAlloc and the baseline is at most 0.058. Interestingly, for the fourth organ, FairAlloc even improves the ranking performance by 0.051. The improvement could be attributed to fine-tuning the scores with policy gradients to optimize NDCG.

4.5.2 Hyper-parameters Sensitivity Analysis

To study RQ3, we analyze how the key hyper-parameters impact the performance of FairAlloc. We consider three hyper-parameters: the weight of group fairness reward, the weight of individual fairness reward, and the weight of utility reward. Figure 4.3 visualizes their impacts. We make

three interesting observations: (1). The group fairness across different genders and races (yellow and green lines) can be further improved when the weight of group fairness reward increases. (2). Similarly, the individual bias (red line) can be reduced with a larger weight of individual fairness reward. (3). NDCG (blue line) increases as the weight of the utility reward increases. The results suggest that the three hyper-parameters can control and balance the importance of the three ranking metrics. They can be specified based on their needs.

4.6 Conclusion

The equity issue in organ transplantation has gained increasing attention. In this work, we design a reinforcement learning framework, FairAlloc, for fairly ranking the patients given an organ. By considering both recipient and organ features, FairAlloc can fairly assign the organ to patients on the waiting list. The reinforcement learning objective can directly optimize the utility metrics, such as NDCG and the fairness metrics. Thus, the policy generated by FairAlloc can not only achieve a high ranking performance but also be fair across individual patients and groups of different races and genders. Extensive experiments demonstrate that FairAlloc significantly reduces individual and group bias while maintaining a competitive ranking performance, ensuring equity in organ allocation. Beyond the comparison with the baselines, we have conducted comprehensive analytical experiments to understand how FairAlloc performs under different hyper-parameters settings and reward functions. We also take an insightful look at the allocation strategy with a case study, showing that FairAlloc can significantly reduce the bias. Overall, our work could pave the way for real-world fair organ allocation systems and inspire the fair allocation of various medical resources.

5. MEDICAL HAI WITH TREE-BASED MULTITASK LEARNING *

5.1 Overview

Organ transplant is the essential treatment method for some end-stage diseases, such as liver failure. Analyzing the post-transplant cause of death (CoD) after organ transplant provides a powerful tool for clinical decision making, including personalized treatment and organ allocation. However, traditional methods like Model for End-stage Liver Disease (MELD) score and conventional machine learning (ML) methods are limited in CoD analysis due to two major data and model-related challenges. To address this, we propose a novel framework called CoD-MTL leveraging multi-task learning to model the semantic relationships between various CoD prediction tasks jointly. Specifically, we develop a novel tree distillation strategy for multi-task learning, which combines the strength of both the tree model and multi-task learning. Experimental results are presented to show the precise and reliable CoD predictions of our framework. A case study is conducted to demonstrate the clinical importance of our method in the liver transplant.

5.2 Introduction

Organ transplant is a crucial therapeutic option for individuals with end-stage diseases, e.g., kidney failure [136], liver failure [137], liver cancer [138], etc. However, due to the complex surgical procedures and high risk of graft failure [139], how to allocate organs properly remains an important yet challenging problem. To increase allocation precision and effectiveness, doctors often need to consider a series of post-transplant factors, especially the cause of death (CoD) analysis [140], such as rejection, infection, cancer, and recurrent disease [141]. Accurately predicting and analyzing these CoDs before the transplant can aid doctors in making better clinical decisions regarding organ allocation [142] and precise treatment after the surgery [143]. In this work, we focus on liver transplant as a case study. Currently, the MELD score [83] is widely used as the

*Reprinted with permission from "Multi-Task Learning for Post-transplant Cause of Death Analysis: A Case Study on Liver Transplant" by Sirui Ding, Qiaoyu Tan, Chia-yuan Chang, Na Zou, Kai Zhang, Nathan R. Hoot, Xiaoqian Jiang, Xia Hu, 2023, AMIA Annual Symposium, Copyright by 2023 AMIA.

standard medical indicator to aid doctors in making better clinical decisions. Nevertheless, MELD cannot provide a granular analysis of the aforementioned CoDs factors, since it was originally designed for the 3-month mortality prediction of liver-related diseases. While some statistical methods have been proposed, they are either intended for a limited number of predictors [144] or make strong assumptions about the input features and outcomes, such as linear relations and feature independence [119]. These limitations hinder the accurate prediction of post-transplant CoDs, necessitating the development of more advanced computational methods to support precise clinical decision-making in liver transplant.

Machine learning (ML) has recently received remarkable success in predicting transplant-related medical outcomes [145]. For example, Lau et al. employed neural networks and random forest to predict post-transplant graft failure [93]. Ding et al. developed a prediction framework based on knowledge distillation for the graft status prediction with consideration of fairness issues [146]. Despite their success, the complex nature of liver transplant makes it infeasible to apply previous ML methods directly for post-transplant CoD prediction. We identify two significant challenges from the data and model-related aspects as follows.

First, from a data perspective, a patient usually has multiple CoDs which makes the analysis a multi-label learning task. In addition, recorded CoDs (positive samples) are scarce compared to negative samples, i.e., successful transplantation or unrecorded data. As a result, there is an imbalance problem in the data, making it difficult for machine learning models to accurately predict the positive class [147]. This is because we do not have enough data to learn ML models for different CoD tasks independently.

Therefore, it is infeasible to directly apply traditional multi-class learning methods [148] and existing ML methods for organ transplant [145] in the post-transplant CoDs analysis.

Second, from a modeling perspective, tree-based models like GBDT [38, 149] tend to perform better than neural network (NN) based approaches [150] in the healthcare field, since the majority of organ transplantation records are EHR/tabular data [151].

We also verified this in our preliminary experiments, as shown in Table 5.1. Despite the relative

advantages of tree-based models, they are still limited in tackling our CoD tasks, since they cannot capture the complementary correlations among different CoD tasks (a.k.a. multiple labels) [152]. Thus, there is an urgent need to devise more advanced tree-aware models that can simultaneously handle multiple prediction targets.

To tackle the above challenges, we propose a tree-distillation multitask learning framework, called **CoD-MTL**, for post-transplant CoD analysis. In this paper, we focus on the prediction of rejection and infection since they are the most common post-transplant CoDs [140]. Specifically, for challenge (1), instead of modeling the rejection and infection independently, we develop a multitask learning model [153] with a shared network layer under the CoD-MTL framework to capture their semantic correlations, since they are intrinsically associated with each other in the organ transplantation field. The shared neural networks will take advantage of the various related tasks to alleviate the unbalanced data problem in CoD analysis. For challenge (2), we design a novel tree distillation strategy in CoD-MTL to effectively transfer the advances of tree-based models into neural networks for different CoD tasks. As a result, a principled approach is obtained to integrate the capacity of multitask learning in capturing complementary information across various tasks and the power of tree-based models in modeling tabular data in an end-to-end fashion.

We validate the effectiveness of our framework on the real-world liver transplant dataset. Experiment results show the CoD-MTL can accurately predict the post-transplant CoDs. The case study demonstrates the clinical importance of CoD-MTL to help doctors in organ transplant clinical decisions.

5.3 Data and Problem Description

Data preparation. In this work, we use a patient cohort obtained from the patients registered on the liver transplant waiting list of the Organ Procurement and Transplantation Network (OPTN) [116], consisting of a total of $8,922$ patients who underwent liver transplantation. Out of these patients, $4,160$ died due to rejection (including both acute and chronic rejection), and 3627 died due to infection after the transplant. In addition, we also randomly selected 2000 patients as negative samples who had no documented death after transplantation. In this study, we con-

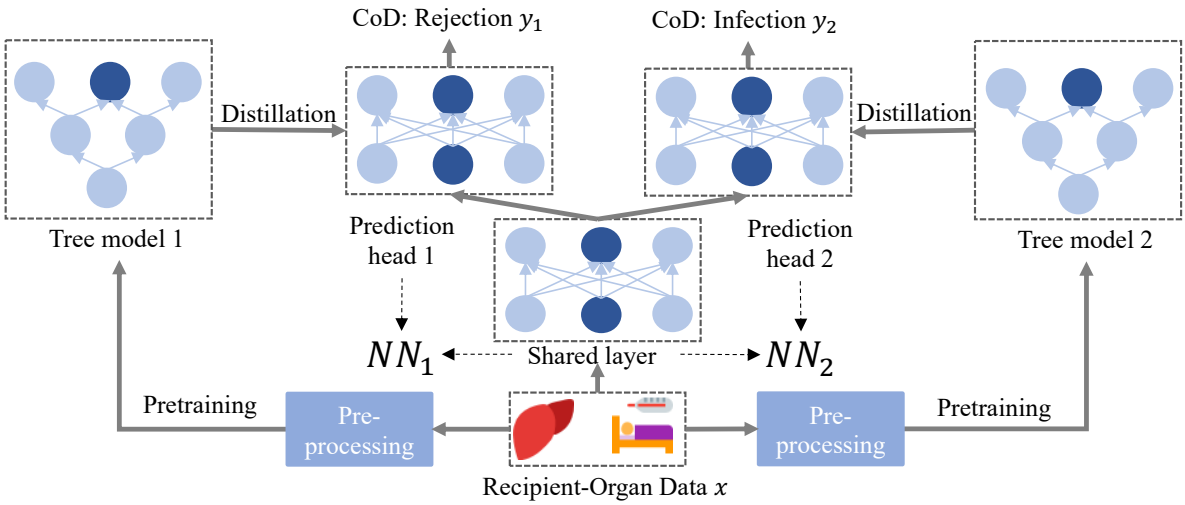


Figure 5.1: An overview of the CoD-MTL workflow for multiple CoDs prediction.

sider 102 features from both the donor organs and recipients, excluding sensitive attributes such as gender and race. The donor/organ features are divided into three categories: the donor's basic information, the donor's history of diseases, and information on the donor's death. Similarly, the patient/recipient features are categorized as the patient's basic information, history of diseases, and transplant-related laboratory tests.

Problem formulation. We are given a dataset $\{p_i, o_i, Y_i\}_{i=1}^N$ consisting of N patient-organ pairs. Each patient p_i (or organ o_i) is associated with a L_p (or L_o) dimensional feature vector $\mathbf{x}_i^p \in \mathbb{R}^{L_p}$ (or $\mathbf{x}_i^o \in \mathbb{R}^{L_o}$). For each patient-organ pair (p_i, o_i), there are M possible causes of death (CoDs), denoted as $Y_i = \{y_j \in \{0,1\}\}_{j=1}^M$, where $y_j = 1$ if the j-th CoD causes the death of the patient and $y_j = 0$ otherwise. The goal is to train a machine learning model that can predict the probability of each CoD for a given patient-organ pair (p_i, o_i) based on their input features $\mathbf{x}_i = \langle \mathbf{x}_i^p, \mathbf{x}_i^o \rangle$. The model should learn to predict multiple CoDs simultaneously, and the learning objective is to minimize the cross-entropy loss between the predicted probabilities and the ground-truth labels.

5.4 Methodology

This section will introduce the proposed post-transplant CoD prediction framework (CoD-MTL) in detail. Firstly, we will describe the pre-processing procedure for input data (Section 5.4.1). Then we introduce the multi-task learning framework for post-transplant CoDs prediction (Section 5.4.2). Finally, the proposed tree-distillation strategy for multi-task learning will be elaborated (Section 5.4.3).

5.4.1 Data Pre-processing

To effectively learn from original liver transplant EHR data, we use an encoder to transform categorical features into numerical values following the standard ways of processing raw data. These numerical features are then concatenated with the original numerical features. To address any missing values, we impute all features with zero. This processed data is used as input for both the tree and multi-task learning models in the CoD-MTL framework. Additionally, to ensure a robust evaluation, the data samples are shuffled during the K-fold cross-validation stage.

5.4.2 Multi-task Learning for Multiple CoDs Prediction

Immunosuppressive drugs that patients take to prevent rejection after liver transplant surgery can weaken their immune system and increase their susceptibility to infections [154]. To investigate the clinical relationship between rejection and infection prediction tasks, we adopt a multi-task learning approach that uses a shared deep learning module and customized prediction heads for different CoDs. The CoD-MTL framework is designed based on the multi-task learning paradigm [24], as illustrated in Figure 5.1. To predict the j-th CoD, we formulate the output as follows:

$$y_j = \text{Head}_j(\text{SharedLayer}(x_i)), \tag{5.1}$$

where $\text{Head}_j(\cdot)$ refers to the prediction head part for the j-th task, and $\text{SharedLayer}(\cdot)$ denotes the shared layer of multiple tasks in the multi-task learning pipeline. We will provide further details

about the Head and SharedLayer modules in the following subsection.

5.4.3 Tree-distillation Boosted Multi-task Learning

In this subsection, we will elaborate on the proposed tree-distillation strategy in the multi-task learning framework. Firstly, we will present the process of integrating the tree model into neural networks using knowledge distillation. Next, we will introduce a new approach to integrate the tree models into the multi-task learning framework. Then we will describe the learning process of the whole CoD-MTL framework.

Tree distillation in the neural network. Tree-based models like GBDT have shown great success across various healthcare scenarios and tabular data [155, 156]. Recently, DeepGBM [104] has been developed to combine the merits of GBDT and deep neural networks by distilling the knowledge of GBDT to deep neural networks. Despite its effectiveness, DeepGBM is designed for a single learning task and cannot model the correlations between multiple learning tasks, as shown in CoD analysis. Inspired by this, we propose to upgrade DeepGBM for multi-task learning, i.e., distilling multiple task-specific GBDT models into a multi-task deep neural network. Assume $V^{t,i}$ denotes the sparse leaf index that corresponds to the i-th patient of the training data in the t-th tree of T, we first transform the leaf outputs of one GBDT model T into a dense embedding as below:

$$\mathbf{E}^i = \text{Emb}(||_{t \in T}(V^{t,i}); \theta), \tag{5.2}$$

where $\mathbf{E}^i$ represents the dense embedding table obtained from the embedding model $\text{Emb}(\cdot)$ with trainable parameter θ, where $\text{Emb}(\cdot)$ is a fully connected neural network. The notation $||_{t \in T}(V^{t,i})$ indicates the concatenated sparse representation across multiple trees in GBDT. To learn the embedding model, we optimize the objective function as:

$$min \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}'(\mathbf{W} \times \text{Emb}(||_{t \in T}(V^{t,i}); \theta) + \mathbf{b}, q^i), \tag{5.3}$$

where $\mathbf{W}$ and $\mathbf{b}$ are the parameters that map the dense embedding into the final prediction, and q^i is the corresponding leaf prediction of the i-th sample. The loss function $\mathcal{L}'$ can be chosen as

Model type	Model	CoD: Rejection		CoD: Infection	
		AUROC	AUPRC	AUROC	AUPRC
Traditional ML (single task)	Logistic Regression	0.551±0.008	0.482±0.005	0.569±0.013	0.471±0.013
	GBDT	0.588±0.008	0.497±0.010	0.611±0.011	0.499±0.014
	Random Forest	0.583±0.016	0.504±0.009	0.608±0.009	0.506±0.020
Neural Networks (single task)	MLP	0.571±0.012	0.493±0.008	0.592±0.003	0.483±0.011
Multitask learning model	Multitask Learning	0.595±0.021	0.517±0.015	0.614±0.019	0.515±0.028
The proposed method	CoD-MTL	0.640±0.012	0.557±0.012	0.646±0.007	0.553±0.018

Table 5.1: Performance comparison on Two CoD Prediction Tasks

the cross-entropy loss function, which is commonly used in classification tasks.

After the embedding of sparse representations from tree models' leaves, we can use this dense embedding E^i as the distillation target to further distill the tree structures into a neural network. The distilled neural network can approximate the tree model by optimizing the following objective:

$$\mathcal{L}_{distill} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\text{NN}(\mathbf{x}_i^p[\mathbb{I}^T]);\theta_{NN}), \mathbf{E}^i), \tag{5.4}$$

where $\text{NN}(\cdot)$ represents the distilled neural network with trainable parameters θ_{NN}, and $\mathbf{x}_i^p$ is the input feature for the i-th patient. $\mathbb{I}$ denotes the indices of the features selected from the tree model.

Integration of tree model in multi-task learning. When it comes to predicting multiple post-transplant CoDs, we propose a multi-task tree-distillation paradigm to achieve this. First, we train a GBDT model for each CoD prediction task. For the j-th CoD, we have GBDT model T_j and the distilled network $\text{NN}_j(\cdot)$ with trainable parameters θ_{NN_j}. We then develop the distilled neural network for multiple CoD tasks, as shown in Figure 5.1. Specifically, the distilled model NN_j for each CoD task includes a shared layer for representation learning and a task-specific prediction head for each CoD task, as shown in Formula 5.1. The prediction head for the j-th CoD is a simple neural network as follows.

$$y_j(x^i) = \mathbf{W}_j \times NN_j(\mathbf{x}_i[\mathbb{I}^{T_j}]);\theta_{NN_j}) + \mathbf{b}_j, \tag{5.5}$$

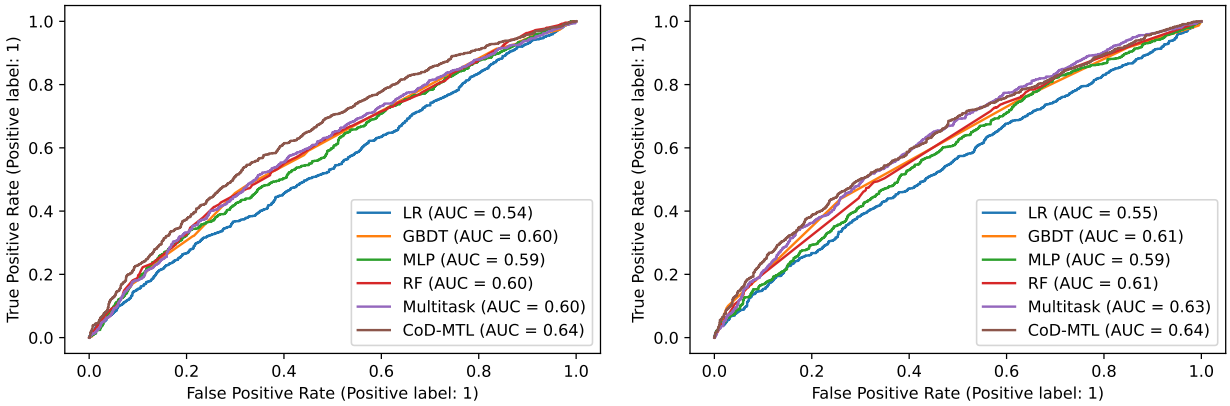


Figure 5.2: ROC curves for rejection and infection CoDs (From left to right).

where $\mathbf{W}_j, \mathbf{b}_j$ are associated parameters to transfer the dense embedding to final predictions for the j-th task.

Learning process of CoD-MTL. To train our model, we optimize the parameters of CoD-MTL according to the following multi-task loss function.

$$\begin{aligned} \mathcal{L}_j &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\mathrm{NN}_j(\mathbf{x}_i[\mathbb{I}^{T_j}]);\theta_{NN_j}),\mathbf{E}_j^i), \\ \mathcal{L}_{multi} &= \sum_{j=1}^{M}\alpha_j(\beta_j\mathcal{L}'(y_j,y_j')+\gamma_j\mathcal{L}_j). \end{aligned} \tag{5.6}$$

$\mathcal{L}_j$ is the knowledge distillation loss function for the j-th CoD task, i.e, NN_j. $\mathcal{L}_{multi}$ is the overall multi-task loss function for M CoD tasks, where α_j, β_j, and γ_j are trade-off parameters to control the importance of different terms.

5.5 Experiment

In this section, we provide a comprehensive evaluation of CoD-MTL from the computational and clinical perspectives by answering the following research questions (**RQ**).

- **RQ1:** Can the CoD-MTL accurately predict the rejection and infection as the CoDs? (Section 5.5.2)
- **RQ2:** To what extent can the Cod-MTL be considered trustworthy for predicting CoDs?

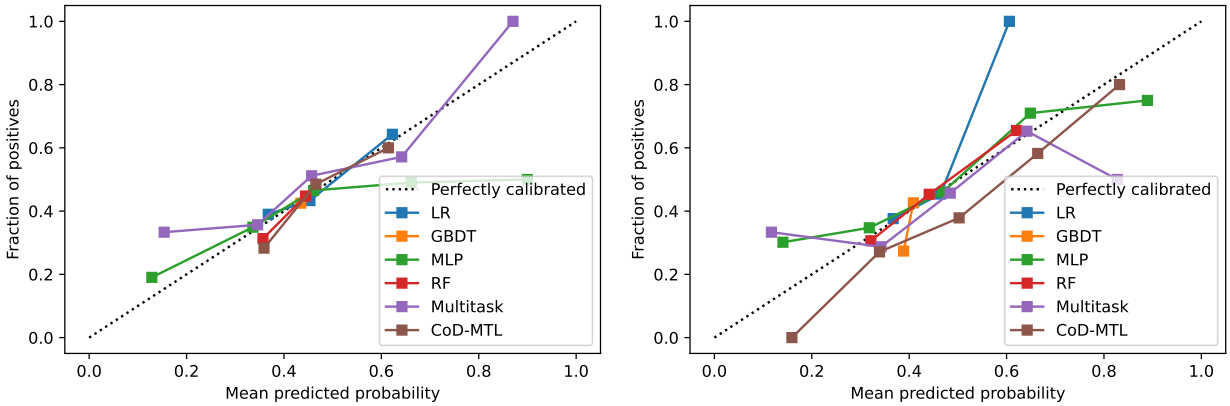


Figure 5.3: Calibration curves for rejection and infection CoDs (From left to right)

(Section 5.5.3)

- **RQ3:** How could the CoD-MTL help the doctor make the clinical decision in liver transplant? (Section 5.5.4)

5.5.1 Experimental Settings

Baseline methods. We choose the baseline methods from three categories which are traditional ML, neural network, and multitask learning model respectively. For traditional ML, we select three commonly used methods as baselines, which are Logistic Regression (LR) [37], Gradient Boosting Decision Tree (GBDT), and Random Forest (RF) [39]. For the neural network, we use a multi-layer perceptron (MLP) [157] as the neural network baseline model. For the multitask learning model, we use the hard parameter sharing multitask learning frameworks as the baseline method.

Evaluation metrics. To ensure a fair comparison, we adopt the K-fold cross-validation strategy to evaluate the baseline and proposed methods. The AUROC and AUPRC metrics will be computed by averaging across multiple folds to assess the prediction accuracy. Additionally, we calculate the standard deviation (STD) of AUROC/AUPRC across different folds to evaluate the model uncertainty. To evaluate the clinical significance of CoD-MTL, we will engage a clinical expert to assist us in the case study.

Implementation details. We implemented the baseline machine learning methods using scikit-

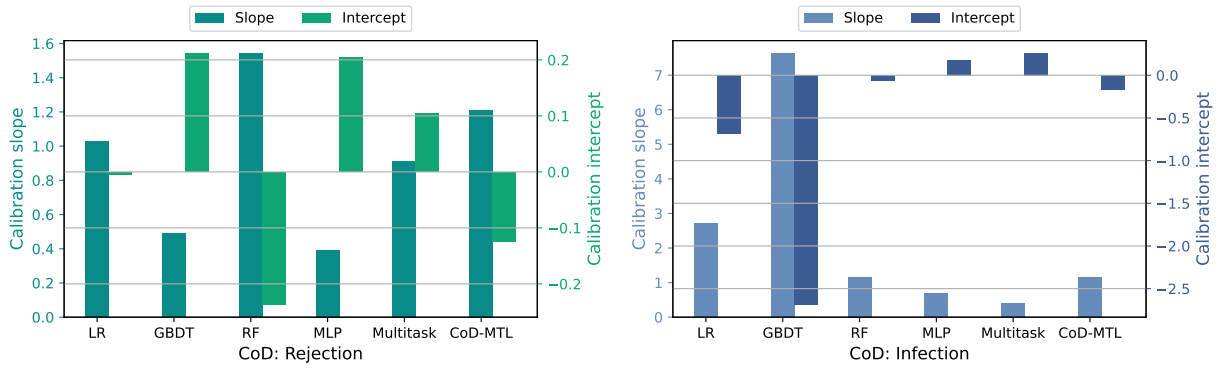


Figure 5.4: Calibration performance on rejection and infection prediction tasks.

learn [158] and PyTorch. The CoD-MTL framework was implemented using LightGBM [112] and PyTorch. We trained the CoD-MTL for 100 epochs using AdamW as the optimizer with a learning rate of 0.001. All the experiments were conducted on a server equipped with NVIDIA V100 GPUs and Intel Xeon CPUs. We set K to 4 for cross-validation.

5.5.2 Prediction Performance on Rejection and Infection as CoDs

We present the superior performance of CoD-MTL compared to the baseline machine learning methods as shown in Table 5.1. Several observations can be summarized as follows:

Firstly, the tree model outperforms MLP method on the CoD prediction task. For the rejection prediction, we observe that GBDT can achieve higher AUROC and AUPRC by 3.0% and 0.81% compared to MLP. For the infection prediction task, the AUROC and AUPRC of GBDT are higher than MLP by 3.2% and 3.3% respectively. This may be due to the ability of GBDT to identify important features from the EHR and eliminate irrelevant features that are less related.

Secondly, the multitask learning model improves performance compared to the single MLP. For the rejection prediction, the multitask learning baseline outperforms the single MLP by 4.2% and 4.9% on AUROC and AUPRC, respectively. For the infection prediction, the multitask learning baseline achieves 3.7% and 6.6% higher AUROC and AUPRC than the single MLP model. Our findings demonstrate that combining two highly related tasks in the multitask learning model can boost the performance of each single task. The shared model parameters can help the model learn

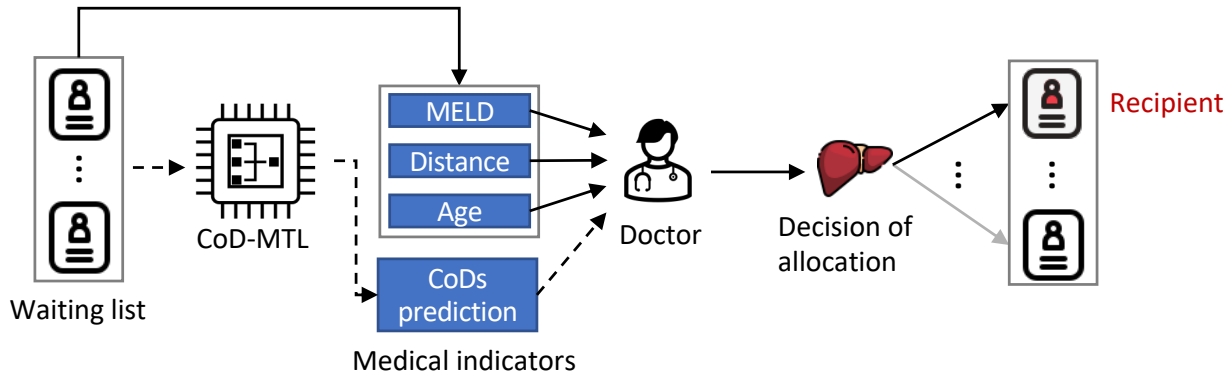


Figure 5.5: Illustration of how can CoD-MTL help the doctor make the clinical decisions in liver transplant.

common knowledge for both tasks and make more precise predictions for each CoD of the patients.

Thirdly, our results demonstrate that CoD-MTL outperforms the other baseline methods by a significant margin. Specifically, we observe a maximum improvement of 16.1% and 15.6% in terms of AUROC for rejection and infection prediction, respectively. Similarly, the maximum improvement in AUPRC is 15.6% and 17.4% for rejection and infection prediction, respectively. These results provide strong evidence of the effectiveness of CoD-MTL in leveraging the advantages of both tree models and multitask learning. By utilizing highly related features and common knowledge between the two tasks, CoD-MTL achieves superior performance on both CoD prediction tasks.

Moreover, we performed the sensitivity analysis using the ROC curve for a single fold of data. Figure 5.2 shows the ROC curves for rejection and infection prediction tasks. As seen in the figure, CoD-MTL exhibits a steeper slope than other baseline methods for both tasks. This indicates that CoD-MTL has a higher sensitivity, which is crucial for accurately identifying patients at high risk of rejection and infection. Early detection of rejection and infection is critical for preventing organ failure or loss and timely medical intervention. Therefore, the superior sensitivity of CoD-MTL makes it a promising approach for liver transplant outcome prediction.

5.5.3 Model Calibration Analysis

The proposed ML model can produce reliable predictions with well-calibrated probabilities, which is crucial for clinical applications [159]. To further investigate the model uncertainty on rejection and infection prediction, we plot the calibration curves on one fold of data, as shown in Figure 5.3. The calibration curve of CoD-MTL in both the left and right parts of Figure 5.3 is close to the diagonal line, indicating that the predicted probabilities correspond to the observed fractions well. To quantitatively measure the calibration performance of the models, we calculate the calibration slope and intercept of the calibration curve in Figure 5.3, as shown in Figure 5.4. The calibration slope of CoD-MTL is close to 1 on both tasks, indicating that the predicted probabilities are well-calibrated with the true probabilities. Although LR's calibration slope is closer to 1 on the rejection prediction task, it is not well-calibrated on the infection task. Similarly, the proposed model achieves a calibration intercept near 0 on both tasks, indicating that the predicted probabilities are well-centered around the observed fractions. These observations suggest that the proposed CoD-MTL model can output reliable predictions with rather small uncertainty across different tasks, making it a promising tool for organ transplant outcome prediction.

5.5.4 Case Study

Our proposed model represents a significant improvement in clinical decision support for liver transplantation as shown in Figure 5.5. It takes into account a variety of factors that can affect patient outcomes, such as the likelihood of rejection or infection, to provide a more nuanced analysis of each patient's individual medical situation.

For instance, in cases where two patients from the same transplant center appear to be very similar, our model may reveal that they have different probabilities of dying from rejection or infection. We present two pairs of patients as shown in Figure 5.6. Patients A and B come from the same transplant center, and patients C and D come from another transplant center. We can observe that patients A and B have the same MELD score which is 20, the same age which is 38, and nearly the same distance from the donor which is 30 and 29 respectively. Our model predicts patient A

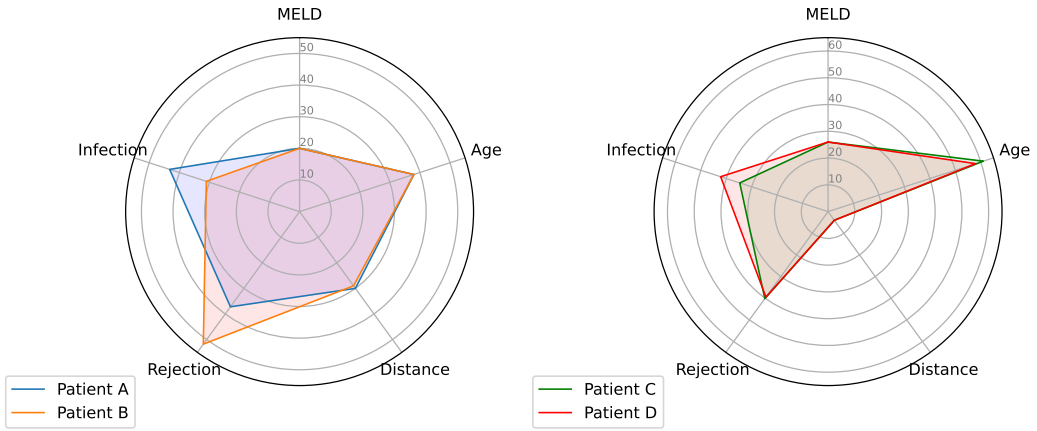


Figure 5.6: Two pairs of patients with similar features from the same transplant centers.

with a higher probability of dying from infection and patient B with a higher probability of dying from rejection. The situation is similar for patients C and D, who share very similar characteristics related to allocation. Patient D is predicted with a higher probability of infection.

This kind of detailed analysis can be invaluable for clinicians who are looking to make more informed decisions about patient care [140]. With this level of information, doctors can develop more personalized treatment plans that are tailored to the specific needs of each patient. For example, they may choose to administer more aggressive immunosuppressant therapy to a patient who is at a higher risk of rejection, while opting for a more cautious approach for a patient who is at a lower risk. By providing doctors with more detailed and accurate information about patient outcomes, our model can help to improve the overall efficiency and effectiveness of liver transplantation. This, in turn, can lead to better outcomes for patients and more efficient use of healthcare resources.

5.6 Discussion of Limitation

Additionally, we need to address some limitations and identify corresponding solutions for future improvement of our CoD-MTL framework. A critical constraint of our current model is its lack of interpretability, as the inability to explain predictions may impede its deployment in organ transplant scenarios. To address this gap, we will incorporate explainable AI techniques [160]

into our future work to provide a human-understandable interpretation of the predictions. Another limitation is the failure to consider fairness in our current framework. Equity is an ethical goal that clinical decision support systems should aim to achieve [161, 162]. Therefore, our future work will place a strong emphasis on integrating fairness constraints within our prediction framework to ensure it is suitable for multiple outcomes in organ transplants. The proposed framework has the potential to extend beyond our focus on organ transplants and can be applied to other medical fields that use multi-task learning. For instance, the prediction of complications and the length of stay in the ICU [77, 27] may benefit from our CoD-MTL framework, subject to future refinement.

5.7 Conclusion

In this work, we propose a novel multi-task learning framework named CoD-MTL for the cause of death prediction in organ transplant. The key innovation lies in designing a tree-distillation strategy in multi-task learning, which serves as a bridge to combine the merits of tree-based models and multi-task deep neural networks for more accurate prediction of the transplant EHR data. Empirical results on the liver transplant cohort show the output of CoD-MTL to be accurate and reliable for the precise liver transplant. The clinical case study further demonstrates our framework can be a promising clinical decision support tool for physicians in organ transplantation-related allocation and treatment procedures. We will attach more emphasis on the explainability and fairness of the framework as a future direction.

6. MEDICAL HAI WITH CROSS-MODALITY DISTILLATION FROM LARGE LANGUAGE MODEL

6.1 Overview

Health event prediction is empowered by the rapid and wide application of electronic health records (EHR). In the Intensive Care Unit (ICU), precisely predicting the health related events in advance is essential for providing treatment and intervention to improve the patients outcomes. EHR is a kind of multi-modal data containing clinical text, time series, structured data, etc. Most health event prediction works focus on a single modality, e.g., text or tabular EHR. How to effectively learn from the multi-modal EHR for health event prediction remains a challenge. Inspired by the strong capability in text processing of large language model (LLM), we propose the framework **CKLE** for health event prediction by distilling the knowledge from LLM and learning from multi-modal EHR. There are two challenges of applying LLM in the health event prediction, the first one is most LLM can only handle text data rather than other modalities, e.g., structured data. The second challenge is the privacy issue of health applications requires the LLM to be locally deployed, which may be limited by the computational resource. **CKLE** solves the challenges of LLM scalability and portability in the healthcare domain by distilling the cross-modality knowledge from LLM into the health event predictive model. To fully take advantage of the strong power of LLM, the raw clinical text is refined and augmented with prompt learning. The embedding of clinical text are generated by LLM. To effectively distill the knowledge of LLM into the predictive model, we design a cross-modality knowledge distillation (KD) method. A specially designed training objective will be used for the KD process with the consideration of multiple modality and patient similarity. The KD loss function consists of two parts. The first one is cross-modality contrastive loss function, which models the correlation of different modalities from the same patient. The second one is patient similarity learning loss function to model the correlations between similar patients. The cross-modality knowledge distillation can distill the rich information in clinical text

and the knowledge of LLM into the predictive model on structured EHR data. To demonstrate the effectiveness of **CKLE**, we evaluate **CKLE** on two health event prediction tasks in the field of cardiology, heart failure prediction and hypertension prediction. We select the 7125 patients from MIMIC-III dataset and split them into train/validation/test sets. We can achieve a maximum 4.48% improvement in accuracy compared to state-of-the-art predictive model designed for health event prediction. The results demonstrate **CKLE** can surpass the baseline prediction models significantly on both normal and limited label settings. We also conduct the case study on cardiology disease analysis in the heart failure and hypertension prediction. Through the feature importance calculation, we analyse the salient features related to the cardiology disease which corresponds to the medical domain knowledge. The superior performance and interpretability of **CKLE** pave a promising way to leverage the power and knowledge of LLM in the health event prediction in real-world clinical settings.

6.2 Introduction

The rapid adoption of Electronic Health Records (EHR) [163] has transformed healthcare, offering vast repositories of patient information. In the Intensive Care Unit (ICU), the ability to predict health-related events [164, 165] in advance is paramount for optimizing treatment strategies and improving patient outcomes. EHR is multi-modality data [166] containing clinical text [167] (e.g., diagnosis notes) and time series data [168] (e.g., ECG, EEG), and structured data [169] (e.g., lab tests). However, existing health event prediction models often focus on a singular modality, such as text [170] or tabular EHR [171], presenting a significant challenge in effectively harnessing the entirety of multi-modal EHR data [172]. Health event prediction [173] is an essential task in the field of medicine. It is the foundation for precision medicine [1], personalized treatment [174], etc. With the rapid development of electronic health records (EHR), the data in healthcare becomes more accessible for training the ML models. In the realm of digital health [175], we can design ML models to precisely predict health event in advance from the EHR data.

Heart failure [176], a multifaceted clinical syndrome marked by the heart's compromised ability to pump blood effectively, stands as a formidable challenge within healthcare systems globally.

The unpredictable nature of heart failure exacerbations necessitates predictive models that can anticipate events, enabling clinicians to intervene proactively [177]. Hospitalizations and adverse outcomes associated with heart failure place a considerable burden on both patients and healthcare resources [178]. Accurate prediction models offer the potential to enhance patient care, reduce hospitalizations, and optimize treatment strategies [179]. Hypertension, often referred to as the "silent killer," remains a prevalent cardiovascular condition characterized by elevated blood pressure levels [180]. The insidious nature of hypertension makes it imperative to identify and predict impending events, such as severe complications like strokes and heart attacks [181]. Timely interventions based on accurate predictions can mitigate risks and improve long-term outcomes for individuals living with hypertension [182]. Predictive models tailored to the dynamic nature of blood pressure fluctuations and patient-specific factors are instrumental in shaping personalized care plans. While traditional predictive models have made strides in these domains, the integration of multi-modal data and advanced processing techniques, such as those offered by LLMs [183], opens new avenues for refined predictions. Predicting events in heart failure and hypertension introduces specific challenges that necessitate a targeted approach. These challenges include the need to assimilate and interpret diverse data modalities within EHRs, ranging from clinical narratives to structured data and temporal trends. Additionally, the intricate interplay of factors contributing to heart failure and hypertension requires models that can capture the complexity of patient health trajectories.

Efforts are put into building health event predictive model on EHR. Some works [184] use the structured EHR to build the predictive model. Others [185] use clinical text to predict health events. There are some works [186, 187] that use both structured and text EHR data. Some of them simply use the clinical text as auxiliary information [186]. Others generate the embedding from clinical text and fuse the multi-modal representations to make final predictions. With the widely application of large language model (LLM), it provides a transformative way to build predictive model on multi-modality EHR data [188]. Some previous works are put into how to apply LLM in health event prediction [189]. However, there are still challenges that hinder the landing of

LLM applied to health event prediction. Compared to the traditional deep learning models for text processing, the special characteristics of LLM pose several challenges in the healthcare application. We summarized the challenges from the model and data perspectives as follows.

The size of LLM is not scalable and portable for real-world health predictive applications [190]. As we know, directly using the online LLM for inference has privacy issues [191] and is very expensive [192]. The local model is needed in many real-world clinical scenarios, e.g., hospitals, medical centers, etc [193]. However, the large size of LLM limits its local deployment and the efficiency of inference didn't meet the real-time requirement of AI healthcare algorithms [194]. **Learning from both clinical text and structured data remains a challenge.** LLM mainly handles the text data, which is only one modality in EHR data. There are other modalities like structured EHR data which could be learned with predictive models, e.g., Transformer. There is a need to effectively learn both modalities in one framework and adapt LLM in the end-to-end training pipeline [195, 196]. Meanwhile, the clinical text usually contains much noise [197], which will mislead the model learning if directly embedded [198]. **How to model the patient similarity in multi-modality learning.** Previous multi-modal methods fuse the embeddings of multi-modal data and cannot mine the latent relations between patients [199]. Learning the patient similarity is inspired by the doctor's clinical practice which will refer to the past and related patients' history.

We are motivated to mitigate these challenges by proposing the **CKLE** framework. For the first and second challenge, the **CKLE** framework distills the knowledge from LLM on the cross-modality EHR data. The cross-modality distilling can integrate the LLM's knowledge into the prediction model without increasing the model complexity. To fully exploit and utilize the knowledge from LLM, we refine and augment the raw clinical text with prompt learning on LLM which can effectively remove the noise. The augmented clinical text from LLM contains less noise and more general textual information with augmentation. To simultaneously mine the patient similarities and the latent cross-modality relations, we design a contrastive loss to model the pairs of patients and each patient's text-visit pairs. The contributions of this work can be summarized as follows:

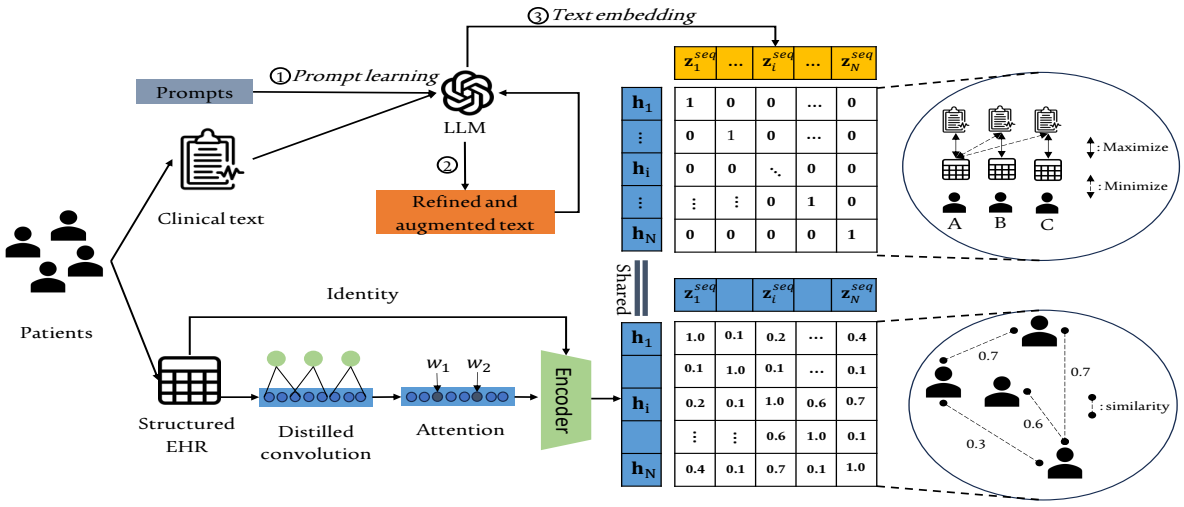


Figure 6.1: Overview of the CKLE framework.

- We distill the cross-modal knowledge from LLM to boost the health event prediction and fully exploit the LLM capability by prompting for the text augmentation.
- We design a contrastive distillation loss to learn the multi-modality knowledge from teacher model and similarities between patients at the same time.
- Extensive experiments are conducted on two representative health event prediction tasks to validate the effectiveness of the **CKLE** framework. **CKLE** can achieve competitive prediction performance on the real-world text-rich EHR data.

6.3 Method

6.3.1 Problem Formulation

For each patient, there will be multiple visits V_i, where $i \in [1, m]$ indicates the m visits to the hospital. Each visit can be represented by the ICD codes demonstrating the diagnosis and treatment in the i-th visit as $C_i = \{c_1, c_2, ..., c_k\}$, where k is a constant number of the ICD codes. Additionally, there are attached clinical notes from the doctors for each visit denoted as N_i. The

dataset and problem can be formulated as follows.

Patient visits dataset: The input dataset can be denoted as $D = P_1, P_2, ..., P_n$ containing n patients. For the ith patient, $P_i = (V_i, N_i)$ where each patient has both visits data and text data.

Health event prediction: Given the i-th patient features of previous $t-1$ visits P_i^{t-1}, the goal is to train the prediction model $Q(\theta)$ with learnable θ parameters, which takes P_i^{t-1} as input and precisely predict the targeted health event y_i at the t-th visit of the patient. For diagnosis prediction, y_i is a multi-class target. For heart failure prediction and ventilator intervention prediction, $y_i \in 0, 1$ is a binary-class target.

6.3.2 CKLE Framework Overview

6.3.2.1 Representation learning from visits data

Long-short term feature modeling. We adopt the dilated convolution to learn the long term and short term information from the multiple visits features inspired by [200]. The long-short term feature extraction can be achieved by setting different dilation rate d. The dilation convolution layers $\mathbf{dconv}$ with dilation rate d can be represented as follows.

$$\mathbf{dconv^d}(V_i^q) = \sum_{j=0}^{K} V_i^q[q + d \cdot j] \cdot f(j), \tag{6.1}$$

where V_i^q is the q-th dimension feature corresponds to i-th patient input V_i. The convolution filer with filter size K is denoted as $f(j), j \in [0, K]$. This illustrates how to learn the representations with a given reception length. Modeling the hidden features at different scale requires multiple convolution with various dilated rate which can be represented as follows.

$$c_i = \mathbf{concat}[dconv^{d_1}(V_i), ..., dconv^{d_r}(V_i)], \tag{6.2}$$

where c_i is the convolution embedding of the i-th patient by combining multiple dilated convolution representations with dilated rate from d_1 to d_r. r is the number of different dilated convolution.

We also employed the feature recalibration module proposed in [200] to attach suitable attention to different features. The feature recalibration module can be formulated as follows.

$$\mathbf{recal}(c_i) = \sigma_1(W_1\sigma_2(W_2c_i))), \tag{6.3}$$

The c_i is hidden representation learnt by dilated convolution. W_1, W_2 are trainable parameters that serves as the features learnable weights. σ_1, σ_2 are activation function which are *Sigmoid, ReLU* respectively. The $\mathbf{recal}$ weights are then applied to c_i with element-wise multiplying.

$$c_i' = \mathbf{recal}(c_i) \odot c_i \tag{6.4}$$

To further improve the representation learning performance, a residual module is applied to remain the original information from patients.

$$V_i' = \mathbf{recal}(V_i) \odot V_i, \tag{6.5}$$

$$E_i = \mathbf{concat}[c_i', V_i'] \tag{6.6}$$

The E_i is the concatenated embedding after the convolution and feature recalibration. The temporal embedding $\mathrm{z_i}$ will be generated by feeding E_i into a temporal model $\mathbf{temp}$, e.g., RNN, GRU.

$$\mathrm{z_i} = \mathbf{temp}(E_i). \tag{6.7}$$

Patient similarity modeling. When the doctors are making diagnosis and clinical decisions, they will usually refer to the history of similar patients. Inspired by this process in traditional clinical workflow, we design a contrastive learning module to model the patient similarity which could take advantage of the features from similar patients for health event prediction.

In the conventional contrastive learning setting [201], the embedding of each patient should

be the closet as their own embedding (positive pairs) and farthest to other patients embedding (negative pairs). The target of the contrastive loss is an identity matrix. However, this could not learn the similarity between different patients. So we design a soft target for the contrastive loss to model patient similarity.

We use the ICD codes of each patient as the semantic label $\mathbf{I_p}$ of patients. The soft target can be represented as:

$$s_p = \frac{I_p \cdot I_p^\top}{||I_p||^2}, \tag{6.8}$$

where s_p denotes the similarity between each patient pairs. The logits y_i between patients with batch size $|bs|$ are obtained through:

$$y_i = \frac{\exp(\mathbf{dis}(\mathrm{z_i}^\top, \mathrm{z_i})/\tau))}{\sum_{j=1}^{|bs|} \exp(\mathbf{dis}(\mathrm{z_i}^\top, \mathrm{z_i})/\tau))}, \tag{6.9}$$

$\mathbf{dis}$ can be distance computation method which is cosine similarity in this work and τ indicates the temperature hyper-parameters. Similarly, the target of patient-patient pairs can be calculated with s_p as:

$$y_i' = \frac{\exp(s_p)}{\sum_{j=1}^{|bs|} \exp(s_p)))}, \tag{6.10}$$

Thus the loss function $\mathcal{L}_{\text{psim}}$ for patient similarity modeling can be represented as:

$$\mathcal{L}_{\text{psim}} = -\frac{1}{|bs|} \sum_{i=1}^{|bs|} y_i' \log y_i, \tag{6.11}$$

6.3.2.2 Exploit medical knowledge from LLM

Text augmentation with prompt learning. The raw clinical text contains unignorable noise and redundant information. To remove the unrelated information and increase the generalizability of text, we propose to achieve the text augmentation with prompt learning on LLM. For each patient i, we have raw notes N_i. The LLM with frozen knowledgeable parameters are exploited to augment and polish the raw text with prompt learning. The prompt E input to the LLM are

designed as *"Refine the following clinical text without changing its meanings:"*. Then the prompt will be attached with each patient's clinical notes N_i. The augmented and refined clinical text can be generated as follows.

$$N_i' = \mathbf{LLM}(E, N_i), \tag{6.12}$$

where N_i' is the augmented and refined notes.

Cross-modality distillation from LLM. Each patient have clinical text and tabular data, we obtain the embedding $\mathrm{z_i}$ of tabular data through the representation learning on multiple visits. To further take advantage of the strong power of LLM, we generate the embedding h_i of clinical text from the LLM as follows.

$$\mathrm{h}_i = \mathbf{LLM_{emb}}(N_i'). \tag{6.13}$$

To exploit the rich information from multi-modality, a cross-modality distillation strategy is designed to transfer the knowledge from LLM to the health prediction model. The $\mathbf{LLM}$ serves as the teacher model with frozen parameters and the prediction model $\mathbf{Q}$ is the student model with parameters θ. The $\mathrm{z}_i, \mathrm{h}_i$ are structured EHR and clinical text embedding generated by $\mathbf{LLM}$ and $\mathbf{Q}$ respectively.

As each patient has structured EHR and clinical text, the z_i, h_i from the same patient P_i are highly related. On the other hand, the EHR and clinical text from different patients shares few common information for which the embeddings should have larger distance. Inspired by this domain knowledge, the distillation objective $\mathcal{L}_{\text{cmkd}}$ is designed to learn the contrastive relations [202] between EHR-text pairs formulated as:

$$\mathcal{L}_{\text{cmkd}} = -\frac{1}{bs}\sum_{i=1}^{|bs|}\log\frac{\exp(\mathbf{dis}(\mathrm{z}_i, \mathrm{h}_i)/\gamma))}{\sum_{j=1}^{|bs|}\exp(\mathbf{dis}(\mathrm{z}_i, \mathrm{h}_j)/\gamma))}, \tag{6.14}$$

where γ is the temperature hyper-parameters.

6.3.2.3 *Training for CKLE*

The final prediction of the health event can be represented as:

$$\hat{y}_t = \sigma(W\mathrm{h_i} + b), \tag{6.15}$$

where W, b are learnable parameters and σ is the activation function, like *sigmoid*. The overall loss function $\mathcal{L}$ of **CKLE** can be represented by combining these objectives together.

$$\mathcal{L} = \alpha l(\hat{y}_t, y) + \beta \mathcal{L}_{\text{psim}} + \eta \mathcal{L}_{\text{cmkd}}, \tag{6.16}$$

where l is the conventional prediction loss function, e.g., cross entropy. y is the ground truth label. α, β, η are hyper-parameters to control the ratio of different objectives.

6.3.3 Experimental Setup

6.3.3.1 *Dataset and tasks*

The well-known medical EHR dataset MIMIC-III is used for the experiments. We filter out to get patients with both clinical text and the corresponding structured data. Inspired by [186], the patients with multiple visits will be used for the health event prediction. Each patient's previous visits will be used to to predict the last visit. Two tasks related to cardiology diseases are selected as the representative health event prediction tasks in this work. Details of each task will be introduced as follows.

- Hypertension prediction: This task is a binary classification to predict the hypertension of the patient's next hospital visit.
- Heart failure prediction: This task is a binary classification, which predicts whether the heart failure will happen to the patient in the next hospital visit.

6.3.3.2 *Baselines*

- RETAIN: A widely used interpretable healthcare prediction framework with the reverse attention module proposed by Choid et al [203].

- AdaCare: Ma et al. [200] designed the AdaCare framework for representation learning on EHR data by modeling the short and long term features and provide explainability with competitive performance.

- Dipole: An interpretable prediction framework based on bi-directional RNN is proposed by Ma et al. [204]. Dipole can memorize the long-term history information and provide clinical meaningful interpretation.

- CGL: This is a speciallized health event prediction framework [186] with text-rich EHR data. CGL use graph learning to learn the patient similarities for event prediction.

- Chet: It is proposed by Lu et al [184] to use dynamic disease graph to learn the temporal variation of diseases for each patient.

- EHR+LLM: This is the baseline multi-modal method that use BlueBERT to generate clinical text embeddings. The fusing embeddings of text and structured EHR will be used to make predictions.

6.3.3.3 Implementation details

We adopt the data pre-processing method from CGL framework [186]. The clinical notes "Discharge summary" from the patient will be filtered out for its high correlation with the prediction targets. The train-valid-test set are splited with the ratio of $6000/125/1000$. We use PyTorch to implement all the baseline and train/test all models on the Nvidia Tesla V100 GPU. F1-score, AUROC, and AUPRC are used as metrics to evaluate the performance of prediction.

6.4 Results

In this section, the results of experiments aim to answer several research questions (RQ) as follows:

- **RQ1:** How does the **CKLE** framework performed in health event prediction compared to SOTA baselines?

Models	Hypertension			Heart Failure		
	F1 (%)	AUROC (%)	AUPRC (%)	F1 (%)	AUROC (%)	AUPRC (%)
RETAIN	76.02(0.33)	75.04(0.77)	80.24(0.78)	68.70(0.10)	84.53(0.06)	76.90(0.10)
AdaCare	78.89(0.42)	75.80(0.01)	79.41(0.00)	70.67(1.01)	84.87(0.12)	77.33(0.38)
Dipole	78.13(0.98)	73.28(2.20)	77.82(0.68)	70.37(0.31)	84.40(0.62)	76.13(0.15)
CGL	72.14(1.90)	70.52(0.15)	73.01(0.46)	71.18(0.30)	84.79(0.20)	73.88(1.33)
Chet	77.35(0.74)	77.77(0.11)	80.56(0.18)	71.06(0.67)	84.88(0.19)	77.88(0.54)
EHR+LLM	77.60(0.35)	74.79(0.29)	76.40(1.39)	67.47(0.76)	84.40(0.26)	74.03(0.91)
CKLE	78.28(0.67)	78.14(0.86)	79.71(1.14)	70.77(0.15)	85.30(0.10)	78.00(0.00)

Table 6.1: Prediction performance on cardiovascular diseases.

- **RQ2:** When the labeled data is limited, can the **CKLE** still shows competitive performance?
- **RQ3:** What are the contribution of each core part in the **CKLE** framework?
- **RQ4:** What are the visualization results of the model embedding? Can it explain the learning mechanism behind **CKLE**?

6.4.1 CKLE Precisely Predicts the Health Event on Multi-modal EHR Data

From the Table 6.1, the proposed framework **CKLE** are compared with baseline methods on cardiovascular prediction tasks (Hypertension prediction and heart failure prediction). Several key findings are summarized from the results.

The proposed CKLE framework can achieve competitive performance compared to state-of-the-art baselines on health event prediction tasks. For the hypertension prediction, we observe the **CKLE** achieves the best performance measured by AUROC. Meanwhile, **CKLE** achieves best performance on AUROC and AUPRC on the heart failure prediction. The high performance on these two cardiology related event prediction demonstrates the potential of applying CKLE for health event prediction in advance, especially for emergency medicine.

Modeling the patient similarity can improve the health event prediction performance. From two tasks presented in Table 6.1, we can observe the graph based prediction method e.g., **Chet** and the **CKLE** framework can outperforms other categories of baselines including CNN-based method (Adacare), RNN-based (Diploe). The **CKLE** framework outperforms the RETAIN

by 2.52% and Adacare by 1.80% on average. The reason behind this phenomenon is probably due to either graph based method or our proposed method take the patient similarity into the modeling process. We take advantage of the patient similarity to help improve the health event prediction of a particular patient. With more related information as the input, the event prediction accuracy can be reasonably increased.

Directly combining the text features generated from clinical text is infeasible for the performance improvement. The baseline LLM method for health event predictions directly leverage the embedded clinical text generated from LLM as the additional features doesn't work efficiently to enhance the prediction accuracy to some extent. For example, compared to the backbone model, the naive LLM boosted method improves 2.08% F1 score on the hypertension prediction task. Meanwhile the other performance metrics on hypertension and heart failure prediction drops, which means the direct use of text features cannot improve the performance. There are two potential causes of this unsatisfying performance with additional LLM generated features. Firstly, the direct encoding of clinical text with LLM will inevitably include noise and redundant information which will affect the performance of the model. The second reason is the cross-modality knowledge distillation from LLM is more effective than naive LLM usage.

Cross-modality distillation from LLM is more effective than directly concatenating generated features. For the hypertension prediction, the CKLE improves the prediction performance by 3.61% on average compared to single modality model and 4.48% on average compared to directly using LLM generated text features. For the heart failure prediction, the average improvement is 0.709% compared to single modality model and 1.07% compared to directly using LLM. We take a further step to investigate the superiority of the cross-modality distillation strategy. The first reason is directly concatenating the features increase the dimension of the input data, which will suffer from the *curse of high dimension*. The second reason is cross-modality distillation with the proposed contrastive loss can learn the inner correlations between different modality features.

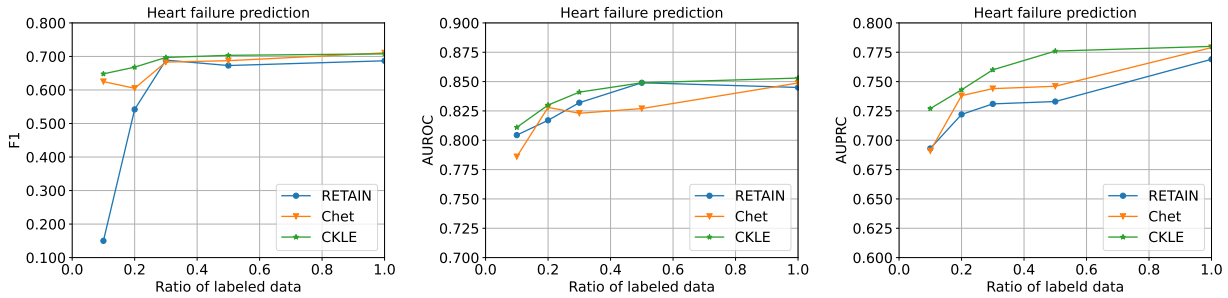


Figure 6.2: Performance comparison with limited labeled training data.

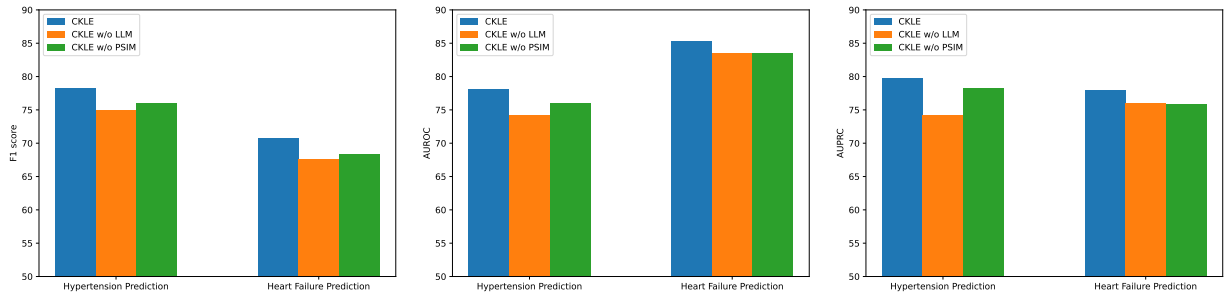


Figure 6.3: Results of ablation study.

6.4.2 CKLE has Competitive Performance with Limited Labeled Data

To evaluate the effectiveness of **CKLE** under the limited label settings which is a common scenario in the medical application, we conduct the experiments by reducing the ratio of labeled training data on the heart failure prediction. From Figure 6.2, we have two observations as follows.

Increase the number of labeled data can increase the overall prediction performance, but the marginal effect exists. We can observe the performance of these models can gain non-trivial improvement when the ratio of labeled data is increased from 0.1 to 0.5. But the performance on whole dataset doesn't have obvious improvement compared to half of the dataset, which indicates increasing the training data has marginal performance improve when reaches a threshold of enough labeled data.

The CKLE framework can outperform baseline methods with limited labeled data. As presented in Figure 6.2, the **CKLE** framework can still surpass the baselines under different ratio of the training data. When we only use 0.1 labeled data to train the **CKLE**, we can still gain 3.18%

improvement compared to the Chet model. Similarly, the **CKLE** can achieve the competitive performance with only 0.5 training data compared to the best baseline model trained on full data.

6.4.3 Ablation Study

We conduct ablation study to evaluate the contribution of each part in our framework. The two key designs in the **CKLE** framework is cross-modality distillation from LLM and patient similarity modeling with contrastive loss (PSIM). The ablation study is conducted on heart failure prediction and the results are presented in Figure 6.3. We can observe each part has significant contribution to the performance improvement. If the knowledge is not distilled from the LLM to the predictive model, the performance is not competitive compared to the **CKLE** because the rich knowledge from LLM is powerful and helpful for various downstream health predictive tasks. Additionally, the PSIM part which leverages the patient similarity can further improve the predictive performance.

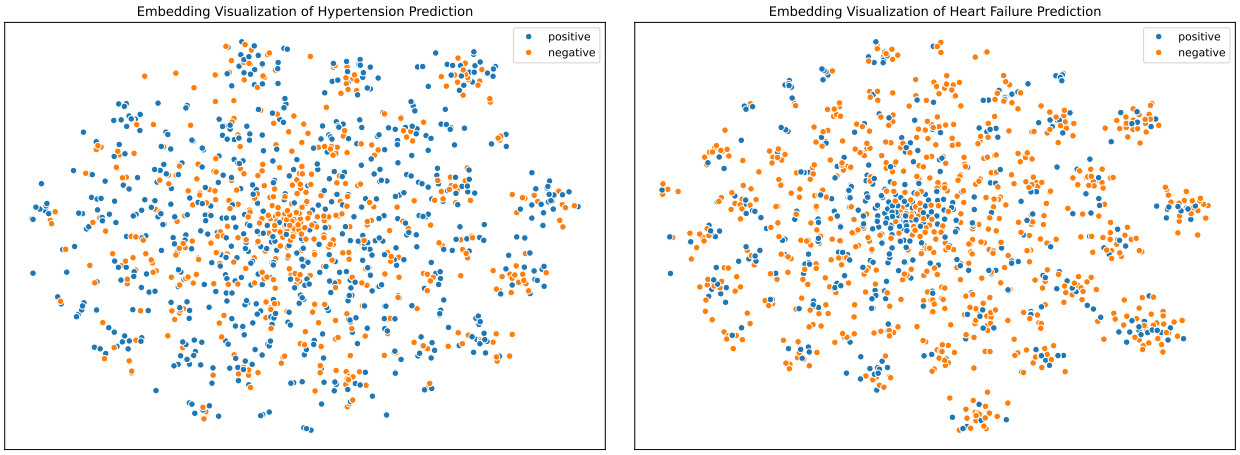
6.4.4 Embedding Visualization

To illustrate the effectiveness of the representation learning ability of the **CKLE** framework, we plot the embedding of each patients in hypertension and heart failure prediction via t-SNE. As shown in Figure 6.4, we compare the embeddings generated from the baseline method and the **CKLE** framework. For hypertension prediction task, the embedding visualized in Figure 6.4b have better clusters of negative and positive patient samples compared with the visualized embedding of the baseline method RETAIN in Figure 6.4a. Similarly, **CKLE** can produce better clusters of embedding on the heart failure prediction tasks.

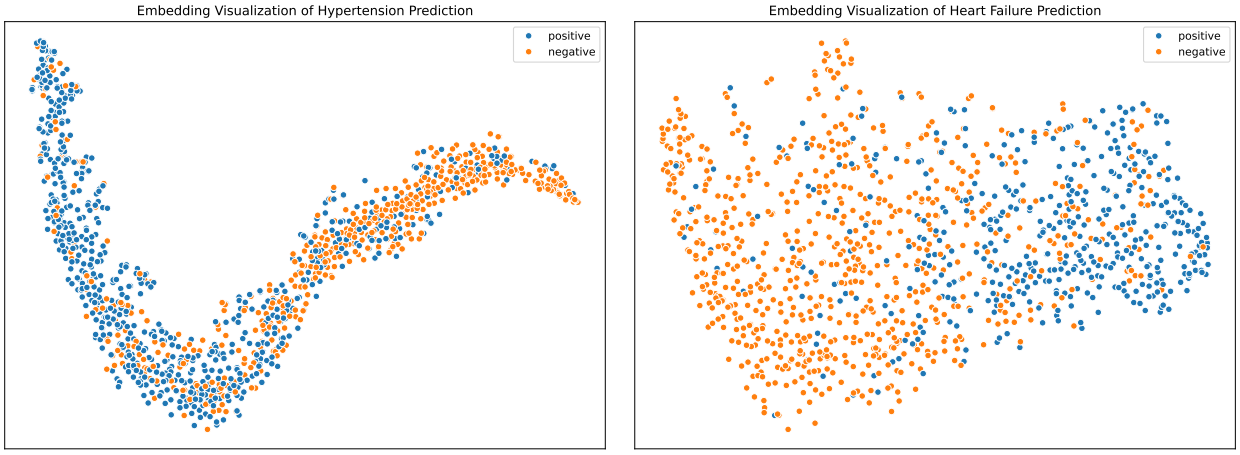
6.4.5 Case Studies on Model Interpretation

6.4.5.1 Case study I: Important Features of Hypertension Prediction

As shown in the left part of the Figure 6.5, we present the 20 most important features for the hypertension prediction. Each feature is assigned a score that signifies its relative importance in the prediction model. Higher scores imply a stronger relationship with the occurrence of hypertension. The most influential feature is coded 401.9, corresponding to unspecified essential hypertension,



(a) RETAIN



(b) CKLE

Figure 6.4: Embedding visualizations (t-SNE) on hypertension and heart failure prediction by RETAIN and CKLE.

which is intuitive as it directly relates to the condition being predicted. Subsequent features include a mix of codes representing both related conditions and general health indicators. We have three observations as follows: (1). The hypertension in previous visits is an important indicator for the occurrence of hypertension the future visits. This makes sense because patients with a history of hypertension tend to be more have hypertension in future hospital visits. (2). The renal disease is highly related to the hypertension. This is a kind of complex and bidirectional relationship. The renal disease can cause and exacerbate the hypertension and vice versa. This finding also corresponds to the medical knowledge in this field [205]. (3). There are also several important features related to newborn infants, which may suggest a correlation between the circumstances of birth and the likelihood of developing hypertension later in life. Compared to hypertension in the other groups of patients, newborn hypertension is relatively rare [206]. From the salient features we observed in the hypertension prediction, the infection of infant(V29.0), respiratory problem in infant(769), feeding issue of the infant(V50.2, 779.3), preterm infants (765.19) are risk factors with high probabilities.

6.4.5.2 Case study II: Important Features of Heart Failure Prediction

In the heart failure prediction, the 20 most important features are presented on the right part of the Figure 6.5. There are several types of the risk factors observed from the important input features. (1). Previous cardiovascular conditions (428.0, 403.91, 428.33, 410.71) of the patients play important role in the heart failure in the future visits. (2). Heart function can also be impacted by metabolic factors (250.00, 272.4, 272.0, 276.5, 244.9, 276.1), e.g., diabetes, hyperlipidemia, and thyroid disorders. (3). Infections and postoperative complications (599.0, 995.92, 998.11, 038.9, 995.91) can exacerbate heart failure and contribute to its development or progression as well. (4). There are also other risk factors related to different organ have significant relations with the heart failure. For example, there are renal and fluid disorders (584.9, 511.9), neurological and seizure disorders (780.39), gastrointestinal disorders (530.81). Interestingly, urinary tract infection and mechanical ventilation associated pneumonia are also included, which may reflect the complex interactions between infections, treatment interventions, and heart failure risk. Hypertensive heart

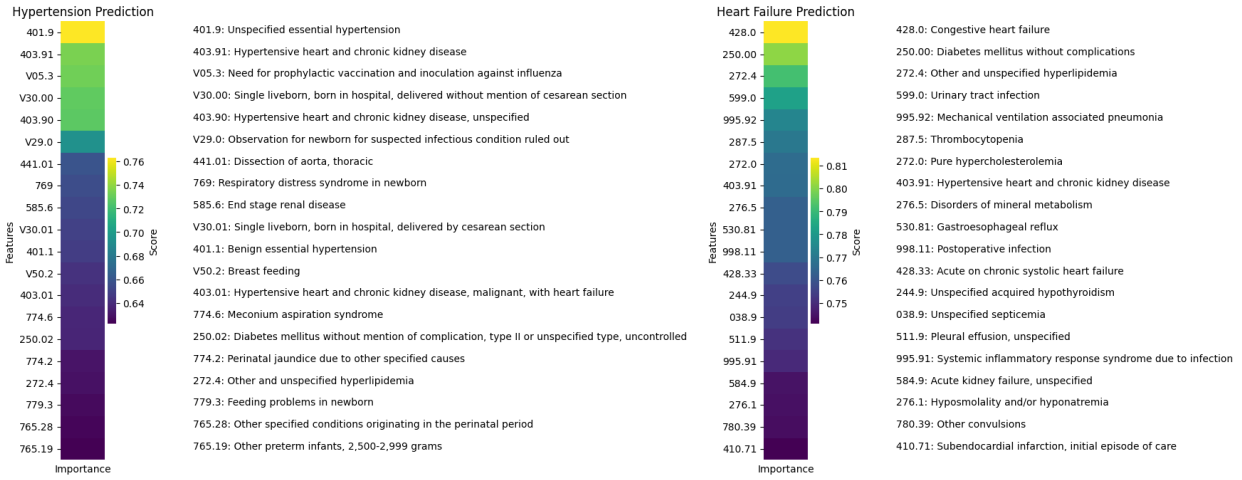


Figure 6.5: Feature importance heatmap for hypertension and heart failure prediction.

and chronic kidney disease and acute on chronic systolic heart failure are directly related to heart function and therefore understandably have high importance scores. Moreover, some features attract less attention from the medical field for heart failure analysis, e.g., disorders of mineral metabolism (276.5), systemic inflammatory response syndrome due to infection (995.91), post-operative infection (998.11) etc.The presence of features for subendocardial infarction and initial episode of care, underscores the multifaceted nature of heart failure risk factors and highlights the potential for machine learning models to discern complex patterns in clinical data for predictive purposes.

6.5 Discussion

Multi-modality learning has been widely discussed and attracted lots of attention for health-care data. The data in the healthcare domain has different characteristics compared with data in the other domains. From the standpoint of data-centric AI, three healthcare data challenges are summarized as follows: (1). The noise in the healthcare data is prevalent and unignorable. This can be caused by the device noise, human bias, noise in recording process, etc. (2). The clinical text is usually obscure. There are lots of professional terms in the medical domain, which requires prior knowledge. (3). There is usually a privacy issue of the health data to request the model deployed locally. Most hospitals will not put their data on the cloud server or use the online models

to help with their clinical workflows. These data level challenges put out the requests to design novel and suitable AI models with an emphasis on the precision, robustness and privacy.

Previous works mainly advance the technique and models to learn the embedding of different modalities and combine them in an efficient way. In the realm of LLM, the representation learning ability from text data has been transformativly boosted. How to efficiently leverage the LLM in the multi-modality healthcare data remains as an open research question. We distinguish the **CKLE** framework from the related work from four aspects. (1) The knowledge of LLM is effectively learned by the health predictive model with knowledge distillation. The knowledge from LLM is powerful which leads to large-scale parameters of the LLM that has efficiency issue. (2). We explore a novel method of patient similarity learning with contrastive loss function. The patient similarity can be learned by taking advantage of the contrastive loss, which can be used to learn postive and negative pairs. We design the soft labels for the contrastive loss function to learn the similarities between patients with more granularity. This contrastive loss for patient similarity learning can be easily adapted to other predictive model by inserting into the loss function. (3). Besides competitive prediction accuracy, the **CKLE** framework can learn better representations validated by embedding visualization. From the observations in the Table 6.1, the increase of performance metrics indicates the effectiveness of the **CKLE** to improve the predict accuracy. However, we cannot observe the representation effectiveness through the numerical results, which are also very important to evaluate the model. In the t-SNE embedding visualization experiments, we can observe a more clear discrimination between two categories predicted by the **CKLE** compared to the baseline method. (4). **CKLE** predictive model preserves the global model interpretability, which can provide the feature importance by the attention score. The interpretability is a very essential aspect when we build the medical AI models. In this paper, we distill the knowledge from LLM into the predictive model, which is a type of Transformer. The interpretability of Transformer can be represented as the attention score for each input features. The feature importance can show which feature plays an important role in the predictions. From the model interpretation analysis, we study two cases on hypertension and heart failure prediction. The top 20 important features

we get corresponds to the medical knowledge with the domain expert. So our model can produce precise as well as interpretable predictions on the health events.

From the collaboration with domain expert in the cardiology diseases, we can validate some already known medical knowledge and discover some new features which lacks enough attention previously. The **CKLE** can not only precisely predict health events but also can discover some medical findings.

7. CONCLUSION AND FUTURE DIRECTIONS

This dissertation presents the research efforts I devoted to the human-centered AI for precision medicine. The aim of my research is to design and develop AI framework that can improve and support the precision medicine with the human-centered principles, e.g., fairness. I achieved this goal by combining advanced AI methods such as knowledge distillation, fairness ML, multitask learning, multi-modality learning. First, we benchmark the bias in electronic phenotyping and the commonly used debiasing strategies. Then, to address the challenge of balancing performance and fairness trade-off, we design human-centered AI framework to support the precision medicine through fair-aware knowledge distillation to make fair predictions of medical outcomes. Furthermore, to generate fair and precise clinical decisions, we design a reinforcement learning based fair ranking framework to generate precise and unbiased organ allocation policy, which can directly support the doctor's decision. To simultaneously predict multiple medical outcomes, we integrate the tree model into the multi-task learning framework for post-transplant cause-of-death analysis. Moreover, we propose a cross-modality distillation framework to distill the knowledge from LLM for health event prediction on structured EHR data.

In summary, this dissertation investigate and substantiate the potential of AI in precision medicine with a special focus on the human good. Then I will discuss some potential future directions in this line of research.

- **Training Foundation Model from Unlabeled and Multimodal Medical Data:** There are large amounts of unlabeled and multimodal data in the medical domain. One potential direction is focusing on how to leverage the self-supervised learning to train a foundation model on them. The foundation model can be used on various downstream tasks like diagnosis, phenotyping, treatment, etc., by fine-tuning or prompt learning.

- **Improve Fairness and Robustness of Medical Foundation Model:** Another direction will be investigating and evaluating the potential fairness and robustness issue in current medical

foundation model and propose novel pre-training method to train less biased and more robust medical foundation model.

REFERENCES

[1] M. R. Kosorok and E. B. Laber, "Precision medicine," *Annual review of statistics and its application*, vol. 6, pp. 263–286, 2019.

[2] T. Heart, O. Ben-Assuli, and I. Shabtai, "A review of phr, emr and ehr integration: A more personalized healthcare and public health policy," *Health Policy and Technology*, vol. 6, no. 1, pp. 20–25, 2017.

[3] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology*, vol. 2, no. 4, 2017.

[4] S. Panesar, Y. Cagle, D. Chander, J. Morey, J. Fernandez-Miranda, and M. Kliot, "Artificial intelligence and the future of surgical robotics," *Annals of surgery*, vol. 270, no. 2, pp. 223–226, 2019.

[5] S. Divya, V. Indumathi, S. Ishwarya, M. Priyasankari, and S. K. Devi, "A self-diagnosis medical chatbot using artificial intelligence," *Journal of Web Development and Web Designing*, vol. 3, no. 1, pp. 1–7, 2018.

[6] M. A. Ahmad, A. Patel, C. Eckert, V. Kumar, and A. Teredesai, "Fairness in machine learning for healthcare," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3529–3530, 2020.

[7] R. Hamon, H. Junklewitz, I. Sanchez, *et al.*, "Robustness and explainability of artificial intelligence," *Publications Office of the European Union*, vol. 207, 2020.

[8] A. Barberis, H. J. Aerts, and F. M. Buffa, "Robustness and reproducibility for ai learning in biomedical sciences: Renoir," *Scientific Reports*, vol. 14, no. 1, p. 1933, 2024.

[9] B. Shneiderman, *Human-centered AI*. Oxford University Press, 2022.

[10] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[11] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

[12] M. Du, F. Yang, N. Zou, and X. Hu, "Fairness in deep learning: A computational perspective," *IEEE Intelligent Systems*, vol. 36, no. 4, pp. 25–34, 2020.

[13] R. J. Chen, T. Y. Chen, J. Lipkova, J. J. Wang, D. F. Williamson, M. Y. Lu, S. Sahai, and F. Mahmood, "Algorithm fairness in ai for medicine and healthcare," *arXiv preprint arXiv:2110.00603*, 2021.

[14] K. B. Johnson, W.-Q. Wei, D. Weeraratne, M. E. Frisse, K. Misulis, K. Rhee, J. Zhao, and J. L. Snowdon, "Precision medicine, ai, and the future of personalized health care," *Clinical and translational science*, vol. 14, no. 1, pp. 86–93, 2021.

[15] J. Shen, C. J. Zhang, B. Jiang, J. Chen, J. Song, Z. Liu, Z. He, S. Y. Wong, P.-H. Fang, W.-K. Ming, *et al.*, "Artificial intelligence versus clinicians in disease diagnosis: systematic review," *JMIR medical informatics*, vol. 7, no. 3, p. e10010, 2019.

[16] A. Movaghar, D. Page, D. Scholze, J. Hong, L. S. DaWalt, F. Kuusisto, R. Stewart, M. Brilliant, and M. Mailick, "Artificial intelligence–assisted phenotype discovery of fragile x syndrome in a population-based sample," *Genetics in Medicine*, vol. 23, no. 7, pp. 1273–1280, 2021.

[17] C. Wang, X. Zhu, J. C. Hong, and D. Zheng, "Artificial intelligence in radiotherapy treatment planning: present and future," *Technology in cancer research & treatment*, vol. 18, p. 1533033819873922, 2019.

[18] F. Magrabi, E. Ammenwerth, J. B. McNair, N. F. De Keizer, H. Hyppönen, P. Nykänen, M. Rigby, P. J. Scott, T. Vehko, Z. S.-Y. Wong, *et al.*, "Artificial intelligence in clinical deci-

sion support: challenges for evaluating ai and practical implications," *Yearbook of medical informatics*, vol. 28, no. 01, pp. 128–134, 2019.

[19] S. Dalton-Brown, "The ethics of medical ai and the physician-patient relationship," *Cambridge Quarterly of Healthcare Ethics*, vol. 29, no. 1, pp. 115–121, 2020.

[20] J. Morley, C. C. Machado, C. Burr, J. Cowls, I. Joshi, M. Taddeo, and L. Floridi, "The ethics of ai in health care: a mapping review," *Social Science & Medicine*, vol. 260, p. 113172, 2020.

[21] J. S. Kim, J. Chen, and A. Talwalkar, "Fact: A diagnostic for group fairness trade-offs," in *International Conference on Machine Learning*, pp. 5264–5274, PMLR, 2020.

[22] M. Zehlike, K. Yang, and J. Stoyanovich, "Fairness in ranking: A survey," *arXiv preprint arXiv:2103.14000*, 2021.

[23] L. R. Soenksen, Y. Ma, C. Zeng, L. Boussioux, K. Villalobos Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, and D. Bertsimas, "Integrated multimodal artificial intelligence framework for healthcare applications," *NPJ digital medicine*, vol. 5, no. 1, p. 149, 2022.

[24] R. Caruana, *Multitask learning*. Springer, 1998.

[25] J. Fishman, A. I. D. C. of Practice, *et al.*, "Introduction: infection in solid organ transplant recipients," *American Journal of Transplantation*, vol. 9, pp. S3–S6, 2009.

[26] C. M. Delude, "Deep phenotyping: the details of disease," *Nature*, vol. 527, no. 7576, pp. S14–S15, 2015.

[27] S. Zhang, H. Li, R. Tang, S. Ding, L. Rasmy, D. Zhi, N. Zou, and X. Hu, "Pheme: A deep ensemble framework for improving phenotype prediction from multi-modal data," *arXiv preprint arXiv:2303.10794*, 2023.

[28] J. A. Williams, S. I. Hurst, J. Bauman, B. C. Jones, R. Hyland, J. P. Gibbs, R. S. Obach, and S. E. Ball, "Reaction phenotyping in drug discovery: moving forward with confidence?," *Current drug metabolism*, vol. 4, no. 6, pp. 527–534, 2003.

[29] R. R. Edwards, R. H. Dworkin, D. C. Turk, M. S. Angst, R. Dionne, R. Freeman, P. Hansson, S. Haroutounian, L. Arendt-Nielsen, N. Attal, *et al.*, "Patient phenotyping in clinical trials of chronic pain treatments: Immpact recommendations," *Pain Reports*, vol. 6, no. 1, 2021.

[30] L. Poissant, J. Pereira, R. Tamblyn, and Y. Kawasumi, "The impact of electronic health records on time efficiency of physicians and nurses: a systematic review," *Journal of the American Medical Informatics Association*, vol. 12, no. 5, pp. 505–516, 2005.

[31] J. M. Banda, M. Seneviratne, T. Hernandez-Boussard, and N. H. Shah, "Advances in electronic phenotyping: from rule-based definitions to machine learning models," *Annual review of biomedical data science*, vol. 1, pp. 53–68, 2018.

[32] H. Alzoubi, R. Alzubi, N. Ramzan, D. West, T. Al-Hadhrami, and M. Alazab, "A review of automatic phenotyping approaches using electronic health records," *Electronics*, vol. 8, no. 11, p. 1235, 2019.

[33] I. Chien, N. Deliu, R. Turner, A. Weller, S. Villar, and N. Kilbertus, "Multi-disciplinary fairness considerations in machine learning for clinical trials," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 906–924, 2022.

[34] J. C. Kirby, P. Speltz, L. V. Rasmussen, M. Basford, O. Gottesman, P. L. Peissig, J. A. Pacheco, G. Tromp, J. Pathak, D. S. Carrell, *et al.*, "Phekb: a catalog and workflow for creating electronic phenotype algorithms for transportability," *Journal of the American Medical Informatics Association*, vol. 23, no. 6, pp. 1046–1052, 2016.

[35] C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, and A. M. Lai, "A review of approaches to identifying patient phenotype cohorts using electronic health records," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 221–230, 2014.

[36] A. N. Kho, M. G. Hayes, L. Rasmussen-Torvik, J. A. Pacheco, W. K. Thompson, L. L. Armstrong, J. C. Denny, P. L. Peissig, A. W. Miller, W.-Q. Wei, *et al.*, "Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide

association study," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 212–218, 2012.

[37] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *the Journal of machine Learning research*, vol. 9, pp. 1871–1874, 2008.

[38] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[39] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[40] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[41] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *2014 science and information conference*, pp. 372–378, IEEE, 2014.

[42] R. J. Carroll, A. E. Eyler, and J. C. Denny, "Naïve electronic health record phenotype identification for rheumatoid arthritis," in *AMIA annual symposium proceedings*, vol. 2011, p. 189, American Medical Informatics Association, 2011.

[43] S. Yang, P. Varghese, E. Stephenson, K. Tu, and J. Gronsbell, "Machine learning approaches for electronic health records phenotyping: a methodical review," *Journal of the American Medical Informatics Association*, vol. 30, no. 2, pp. 367–381, 2023.

[44] Q. Li, K. Zhao, C. D. Bustamante, X. Ma, and W. H. Wong, "Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis," *Genetics in Medicine*, vol. 21, no. 9, pp. 2126–2134, 2019.

[45] T. Norman, N. Weinberger, and K. Y. Levy, "Robust linear regression for general feature distribution," in *International Conference on Artificial Intelligence and Statistics*, pp. 2405–2435, PMLR, 2023.

[46] Y. Park and J. C. Ho, "Tackling overfitting in boosting for noisy healthcare data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 7, pp. 2995–3006, 2019.

[47] R. G. Mantovani, T. Horváth, R. Cerri, S. B. Junior, J. Vanschoren, and A. C. P. d. L. F. de Carvalho, "An empirical study on hyperparameter tuning of decision trees," *arXiv preprint arXiv:1812.02207*, 2018.

[48] M. Ross, W. Wei, and L. Ohno-Machado, "big data and the electronic health record," *Yearbook of medical informatics*, vol. 23, no. 01, pp. 97–104, 2014.

[49] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[50] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *Journal of the American Medical Informatics Association*, vol. 24, no. 2, pp. 361–370, 2017.

[51] S. Gao, M. Alawad, M. T. Young, J. Gounley, N. Schaefferkoetter, H. J. Yoon, X.-C. Wu, E. B. Durbin, J. Doherty, A. Stroup, *et al.*, "Limitations of transformers on clinical text classification," *IEEE journal of biomedical and health informatics*, vol. 25, no. 9, pp. 3596–3607, 2021.

[52] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[53] E. Zihni, V. I. Madai, M. Livne, I. Galinovic, A. A. Khalil, J. B. Fiebach, and D. Frey, "Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome," *Plos one*, vol. 15, no. 4, p. e0231166, 2020.

[54] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Information Fusion*, vol. 77, pp. 29–52, 2022.

[55] J. Lee, C. Liu, J. Kim, Z. Chen, Y. Sun, J. R. Rogers, W. K. Chung, and C. Weng, "Deep learning for rare disease: A scoping review," *Journal of Biomedical Informatics*, p. 104227, 2022.

[56] J. C. Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin, and J. Sun, "Limestone: High-throughput candidate phenotype generation via tensor factorization," *Journal of biomedical informatics*, vol. 52, pp. 199–211, 2014.

[57] A. Afshar, I. Perros, H. Park, C. Defilippi, X. Yan, W. Stewart, J. Ho, and J. Sun, "Taste: temporal and static tensor factorization for phenotyping electronic health records," in *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 193–203, 2020.

[58] L. L. Coventry, J. Finn, and A. P. Bremner, "Sex differences in symptom presentation in acute myocardial infarction: a systematic review and meta-analysis," *Heart & Lung*, vol. 40, no. 6, pp. 477–491, 2011.

[59] J. L. Mehta, Z. Bursac, P. Mehta, D. Bansal, L. Fink, J. Marsh, R. Sukhija, and R. Sachdeva, "Racial disparities in prescriptions for cardioprotective drugs and cardiac outcomes in veterans affairs hospitals," *The American journal of cardiology*, vol. 105, no. 7, pp. 1019–1023, 2010.

[60] T. Y. Sun, S. A. Bhave, J. Altosaar, and N. Elhadad, "Assessing phenotype definitions for algorithmic fairness," in *AMIA Annual Symposium Proceedings*, vol. 2022, p. 1032, American Medical Informatics Association, 2022.

[61] S. Ding, R. Tang, D. Zha, N. Zou, K. Zhang, X. Jiang, and X. Hu, "Fairly predicting graft failure in liver transplant for organ assigning," in *AMIA Annual Symposium Proceedings*, vol. 2022, p. 415, American Medical Informatics Association, 2022.

[62] C. Li, S. Ding, N. Zou, X. Hu, X. Jiang, and K. Zhang, "Multi-task learning with dynamic re-weighting to achieve fairness in healthcare predictive modeling," *Journal of Biomedical Informatics*, vol. 143, p. 104399, 2023.

[63] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[64] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," *Advances in neural information processing systems*, vol. 30, 2017.

[65] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5310–5319, 2019.

[66] Y. Elazar and Y. Goldberg, "Adversarial removal of demographic attributes from text data," *arXiv preprint arXiv:1808.06640*, 2018.

[67] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," *arXiv preprint arXiv:1703.03717*, 2017.

[68] F. Liu and B. Avci, "Incorporating priors with feature attribution on text classification," *arXiv preprint arXiv:1906.08286*, 2019.

[69] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," *International Conference on Machine Learning*, 2018.

[70] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," *arXiv preprint arXiv:1707.09457*, 2017.

[71] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, pp. 3315–3323, 2016.

[72] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[73] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[74] C. W. Seymour, J. N. Kennedy, S. Wang, C.-C. H. Chang, C. F. Elliott, Z. Xu, S. Berry, G. Clermont, G. Cooper, H. Gomez, *et al.*, "Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis," *Jama*, vol. 321, no. 20, pp. 2003–2017, 2019.

[75] L. Gattinoni, D. Chiumello, P. Caironi, M. Busana, F. Romitti, L. Brazzi, and L. Camporota, "Covid-19 pneumonia: different respiratory treatments for different phenotypes?," 2020.

[76] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Transactions on Circuits and Systems*, vol. 8, no. 7, pp. 579–588, 2009.

[77] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, p. 96, 2019.

[78] R. A. Harshman *et al.*, "Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis," 1970.

[79] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, pp. 60–69, PMLR, 10–15 Jul 2018.

[80] E. Midding, S. M. Halbach, C. Kowalski, R. Weber, R. Würstlein, and N. Ernstmann, "Men with a womans disease: Stigmatization of male breast cancer patientsa mixed methods anal-

ysis," *American journal of men's health*, vol. 12, no. 6, pp. 2194–2207, 2018.

[81] G. M. Abouna, "Organ shortage crisis: problems and possible solutions," in *Transplantation proceedings*, vol. 40, pp. 34–38, Elsevier, 2008.

[82] R. Saidi and S. H. Kenari, "Challenges of organ shortage for transplantation: solutions and opportunities," *International journal of organ transplantation medicine*, vol. 5, no. 3, p. 87, 2014.

[83] R. Wiesner, E. Edwards, R. Freeman, A. Harper, R. Kim, P. Kamath, W. Kremers, J. Lake, T. Howard, R. M. Merion, *et al.*, "Model for end-stage liver disease (meld) and allocation of donor livers," *Gastroenterology*, vol. 124, no. 1, pp. 91–96, 2003.

[84] S. W. Biggins, W. R. Kim, N. A. Terrault, S. Saab, V. Balan, T. Schiano, J. Benson, T. Therneau, W. Kremers, R. Wiesner, *et al.*, "Evidence-based incorporation of serum sodium concentration into meld," *Gastroenterology*, vol. 130, no. 6, pp. 1652–1660, 2006.

[85] S. V. McDiarmid, R. M. Merion, D. M. Dykstra, and A. M. Harper, "Selection of pediatric candidates under the peld system.," *Liver Transplantation: Official Publication of the American Association for the Study of Liver Diseases and the International Liver Transplantation Society*, vol. 10, no. 10 Suppl 2, pp. S23–30, 2004.

[86] G. R. Silberhumer, H. Hetz, S. Rasoul-Rockenschaub, M. Peck-Radosavljevic, T. Soliman, R. Steininger, F. Muehlbacher, and G. A. Berlakovich, "Is meld score sufficient to predict not only death on waiting list, but also post-transplant survival?," *Transplant international*, vol. 19, no. 4, pp. 275–281, 2006.

[87] R. M. Merion, R. A. Wolfe, D. M. Dykstra, A. B. Leichtman, B. Gillespie, and P. J. Held, "Longitudinal assessment of mortality risk among candidates for liver transplantation," *Liver transplantation*, vol. 9, no. 1, pp. 12–18, 2003.

[88] R. P. Myers, A. A. M. Shaheen, P. Faris, A. I. Aspinall, and K. W. Burak, "Revision of meld to include serum albumin improves prediction of mortality on the liver transplant waiting list," *PloS one*, vol. 8, no. 1, p. e51926, 2013.

[89] D. Delen, A. Oztekin, and Z. J. Kong, "A machine learning-based approach to prognostic analysis of thoracic transplantations," *Artificial Intelligence in Medicine*, vol. 49, no. 1, pp. 33–42, 2010.

[90] J. Yoon, A. Alaa, M. Cadeiras, and M. Van Der Schaar, "Personalized donor-recipient matching for organ transplantation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[91] J. Berrevoets, J. Jordon, I. Bica, M. van der Schaar, *et al.*, "Organite: Optimal transplant donor organ offering using an individual treatment effect," *Advances in neural information processing systems*, vol. 33, pp. 20037–20050, 2020.

[92] J. Byrd, S. Balakrishnan, X. Jiang, and Z. C. Lipton, "Predicting mortality in liver transplant candidates," in *Explainable AI in Healthcare and Medicine*, pp. 321–333, Springer, 2021.

[93] L. Lau, Y. Kankanige, B. Rubinstein, R. Jones, C. Christophi, V. Muralidharan, and J. Bailey, "Machine-learning algorithms predict graft failure after liver transplantation," *Transplantation*, vol. 101, no. 4, p. e125, 2017.

[94] J. Berrevoets, A. Alaa, Z. Qian, J. Jordon, A. E. Gimson, and M. Van Der Schaar, "Learning queueing policies for organ transplantation allocation using interpretable counterfactual survival analysis," in *International Conference on Machine Learning*, pp. 792–802, PMLR, 2021.

[95] D. Bertsimas, T. Papalexopoulos, N. Trichakis, Y. Wang, R. Hirose, and P. A. Vagefi, "Balancing efficiency and fairness in liver transplant access: tradeoff curves for the assessment of organ distribution policies," *Transplantation*, vol. 104, no. 5, pp. 981–987, 2020.

[96] B. Parent and A. L. Caplan, "Fair is fair: We must re-allocate livers for transplant," *BMC medical ethics*, vol. 18, no. 1, pp. 1–7, 2017.

[97] S. R. Kaufman, "Fairness and the tyranny of potential in kidney transplantation," *Current Anthropology*, vol. 54, no. S7, pp. S56–S66, 2013.

[98] A. M. Bishara, D. S. Lituiev, D. Adelmann, R. P. Kothari, D. J. Malinoski, J. D. Nudel, M. B. Sally, R. Hirose, D. D. Hadley, and C. U. Niemann, "Machine learning prediction of liver allograft utilization from deceased organ donors using the national donor management goals registry," *Transplantation direct*, vol. 7, no. 10, 2021.

[99] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pp. 604–613, 1998.

[100] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," *Biomedical Signal Processing and Control*, vol. 52, pp. 456–462, 2019.

[101] R. Sapir-Pichhadze and B. Kaplan, "Seeing the forest for the trees: random forest models for predicting survival in kidney transplant recipients," *Transplantation*, vol. 104, no. 5, pp. 905–906, 2020.

[102] M. Wan, D. Zha, N. Liu, and N. Zou, "Modeling techniques for machine learning fairness: A survey," *arXiv preprint arXiv:2111.03015*, 2021.

[103] S. Caton and C. Haas, "Fairness in machine learning: A survey," *arXiv preprint arXiv:2010.04053*, 2020.

[104] G. Ke, Z. Xu, J. Zhang, J. Bian, and T.-Y. Liu, "Deepgbm: A deep learning framework distilled by gbdt for online prediction tasks," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 384–394, 2019.

[105] M. D. Good, C. James, B. J. Good, and A. E. Becker, "The culture of medicine and racial, ethnic, and class disparities in healthcare," *The Blackwell companion to social inequalities*, pp. 396–423, 2005.

[106] L. E. Egede, "Race, ethnicity, culture, and disparities in health care," *Journal of general internal medicine*, vol. 21, no. 6, p. 667, 2006.

[107] K. Hamberg, "Gender bias in medicine," *Womens health*, vol. 4, no. 3, pp. 237–243, 2008.

[108] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in AI," Tech. Rep. MSR-TR-2020-32, Microsoft, May 2020.

[109] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A convex framework for fair regression," *arXiv preprint arXiv:1706.02409*, 2017.

[110] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, *et al.*, "Wide & deep learning for recommender systems," in *Proceedings of the 1st workshop on deep learning for recommender systems*, pp. 7–10, 2016.

[111] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, "Deepfm: a factorization-machine based neural network for ctr prediction," *arXiv preprint arXiv:1703.04247*, 2017.

[112] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[113] OPTN, "Ethical principles in the allocation of human organs," n.d. Accessed: 22 Nov 2022.

[114] D. Bertsimas, V. F. Farias, and N. Trichakis, "Fairness, efficiency, and flexibility in organ allocation for kidney transplantation," *Operations Research*, vol. 61, no. 1, pp. 73–87, 2013.

[115] "Questions and answers about liver allocation - optn." Website. [cited 22 Nov 2022].

[116] A. Kwong, W. Kim, J. Lake, J. Smith, D. Schladt, M. Skeans, S. Noreen, J. Foutz, E. Miller, J. Snyder, *et al.*, "Optn/srtr 2018 annual data report: liver," *American Journal of Transplantation*, vol. 20, pp. 193–299, 2020.

[117] C.-L. Liu, R.-S. Soong, W.-C. Lee, G.-W. Jiang, and Y.-C. Lin, "Predicting short-term survival after liver transplantation using machine learning," *Scientific reports*, vol. 10, no. 1, p. 5654, 2020.

[118] J. Rhu, J. M. Kim, K. Kim, H. Yoo, G.-S. Choi, and J.-W. Joh, "Prediction model for early graft failure after liver transplantation using aspartate aminotransferase, total bilirubin and coagulation factor," *Scientific Reports*, vol. 11, no. 1, p. 12909, 2021.

[119] O. Nitski, A. Azhie, F. A. Qazi-Arisar, X. Wang, S. Ma, L. Lilly, K. D. Watt, J. Levitsky, S. K. Asrani, D. S. Lee, *et al.*, "Long-term mortality risk stratification of liver transplant recipients: real-time application of deep learning algorithms on longitudinal data," *The Lancet Digital Health*, vol. 3, no. 5, pp. e295–e305, 2021.

[120] C. Xu, A. Alaa, I. Bica, B. Ershoff, M. Cannesson, and M. van der Schaar, "Learning matching representations for individualized organ transplantation allocation," in *International Conference on Artificial Intelligence and Statistics*, pp. 2134–2142, PMLR, 2021.

[121] J. Berrevoets, A. Alaa, Z. Qian, J. Jordon, A. E. Gimson, and M. Van Der Schaar, "Learning queueing policies for organ transplantation allocation using interpretable counterfactual survival analysis," in *International Conference on Machine Learning*, pp. 792–802, PMLR, 2021.

[122] A. Ferrarese, G. Sartori, G. Orrù, A. C. Frigo, F. Pelizzaro, P. Burra, and M. Senzolo, "Machine learning in liver transplantation: a tool for some unsolved questions?," *Transplant International*, vol. 34, no. 3, pp. 398–411, 2021.

[123] A. Singh and T. Joachims, "Fairness of exposure in rankings," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2219–2228, 2018.

[124] A. Singh and T. Joachims, "Policy learning for fairness in ranking," *Advances in neural information processing systems*, vol. 32, 2019.

[125] G. Acharya, R. M. Kaushik, R. Gupta, and R. Kaushik, "Child-turcotte-pugh score, meld score and meld-na score as predictors of short-term mortality among patients with end-stage liver disease in northern india," *Inflammatory intestinal diseases*, vol. 5, no. 1, pp. 1–10, 2020.

[126] R. Wiesner, E. Edwards, R. Freeman, A. Harper, R. Kim, P. Kamath, W. Kremers, J. Lake, T. Howard, R. M. Merion, *et al.*, "Model for end-stage liver disease (meld) and allocation of donor livers," *Gastroenterology*, vol. 124, no. 1, pp. 91–96, 2003.

[127] R. M. Ghobrial, J. Gornbein, R. Steadman, N. Danino, J. F. Markmann, C. Holt, D. Anselmo, F. Amersi, P. Chen, D. G. Farmer, *et al.*, "Pretransplant model to predict posttransplant survival in liver transplant patients," *Annals of surgery*, vol. 236, no. 3, p. 315, 2002.

[128] H. Valizadegan, R. Jin, R. Zhang, and J. Mao, "Learning to rank by optimizing ndcg measure," *Advances in neural information processing systems*, vol. 22, 2009.

[129] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the international workshop on software fairness*, pp. 1–7, 2018.

[130] W. Fleisher, "What's fair about individual fairness?," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 480–490, 2021.

[131] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," *arXiv preprint arXiv:1804.07612*, 2018.

[132] R. L. Plackett, "The analysis of permutations," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 24, no. 2, pp. 193–202, 1975.

[133] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems*, vol. 12, 1999.

[134] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, pp. 229–256, 1992.

[135] A. K. Mathur, C. J. Sonnenday, and R. M. Merion, "Race and ethnicity in access to and outcomes of liver transplantation: a critical literature review," *American Journal of Transplantation*, vol. 9, no. 12, pp. 2662–2668, 2009.

[136] M. Abecassis, S. T. Bartlett, A. J. Collins, C. L. Davis, F. L. Delmonico, J. J. Friedewald, R. Hays, A. Howard, E. Jones, A. B. Leichtman, *et al.*, "Kidney transplantation as primary therapy for end-stage renal disease: a national kidney foundation/kidney disease outcomes quality initiative (nkf/kdoqi) conference," *Clinical Journal of the American Society of Nephrology*, vol. 3, no. 2, pp. 471–480, 2008.

[137] R. Kumar, U. Anand, and R. N. Priyadarshi, "Liver transplantation in acute liver failure: Dilemmas and challenges," *World Journal of Transplantation*, vol. 11, no. 6, p. 187, 2021.

[138] M. Ravaioli, G. Ercolani, F. Neri, M. Cescon, G. Stacchini, M. Del Gaudio, A. Cucchetti, and A. D. Pinna, "Liver transplantation for hepatic tumors: a systematic review," *World Journal of Gastroenterology: WJG*, vol. 20, no. 18, p. 5345, 2014.

[139] C. Tovikkai, S. C. Charman, R. K. Praseedom, A. E. Gimson, and J. van der Meulen, "Time-varying impact of comorbidities on mortality after liver transplantation: a national cohort study using linked clinical and administrative data," *BMJ open*, vol. 5, no. 5, p. e006971, 2015.

[140] K. D. Watt, R. A. Pedersen, W. K. Kremers, J. K. Heimbach, and M. R. Charlton, "Evolution of causes and risk factors for mortality post-liver transplant: results of the niddk long-term follow-up study," *American journal of transplantation*, vol. 10, no. 6, pp. 1420–1427, 2010.

[141] R. Moreno and M. Berenguer, "Post-liver transplantation medical complications," *Annals of hepatology*, vol. 5, no. 2, pp. 77–85, 2006.

[142] M. L. Volk, N. Goodrich, J. C. Lai, C. Sonnenday, and K. Shedden, "Decision support for organ offers in liver transplantation," *Liver transplantation*, vol. 21, no. 6, pp. 784–791, 2015.

[143] M. Bhat, S. A. Al-Busafi, M. Deschênes, and P. Ghali, "Care of the liver transplant patient," *Canadian Journal of Gastroenterology and Hepatology*, vol. 28, no. 4, pp. 213–219, 2014.

[144] N. Gong, C. Jia, H. Huang, J. Liu, X. Huang, and Q. Wan, "Predictors of mortality during initial liver transplant hospitalization and investigation of causes of death," *Annals of Transplantation*, vol. 25, pp. e926020–1, 2020.

[145] N. Gotlieb, A. Azhie, D. Sharma, A. Spann, N.-J. Suo, J. Tran, A. Orchanian-Cheff, B. Wang, A. Goldenberg, M. Chassé, *et al.*, "The promise of machine learning applications in solid organ transplantation," *NPJ digital medicine*, vol. 5, no. 1, p. 89, 2022.

[146] S. Ding, R. Tang, D. Zha, N. Zou, K. Zhang, X. Jiang, and X. Hu, "Fairly predicting graft failure in liver transplant for organ assigning," in *AMIA Annual Symposium Proceedings*, vol. 2022, p. 415, American Medical Informatics Association, 2022.

[147] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–36, 2019.

[148] J. Li, Y. Liu, R. Yin, H. Zhang, L. Ding, and W. Wang, "Multi-class learning: From theory to algorithm," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[149] M. Sayed, D. Riano, and J. Villar, "Predicting duration of mechanical ventilation in acute respiratory distress syndrome using supervised machine learning," *Journal of Clinical Medicine*, vol. 10, no. 17, p. 3824, 2021.

[150] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.

[151] A. B. Massie, L. Kuricka, and D. L. Segev, "Big data in organ transplantation: registries and administrative claims," *American Journal of Transplantation*, vol. 14, no. 8, pp. 1723–1730, 2014.

[152] Z. Zhang and C. Jung, "Gbdt-mo: gradient-boosted decision trees for multiple outputs," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 7, pp. 3156–3167, 2020.

[153] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.

[154] M. D. P. Hernandez, P. Martin, and J. Simkins, "Infectious complications after liver transplantation," *Gastroenterology & hepatology*, vol. 11, no. 11, p. 741, 2015.

[155] W. Chang, Y. Liu, Y. Xiao, X. Yuan, X. Xu, S. Zhang, and S. Zhou, "A machine-learning-based prediction method for hypertension outcomes based on medical data," *Diagnostics*, vol. 9, no. 4, p. 178, 2019.

[156] H. Seto, A. Oyama, S. Kitora, H. Toki, R. Yamamoto, J. Kotoku, A. Haga, M. Shinzawa, M. Yamakawa, S. Fukui, *et al.*, "Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data," *Scientific Reports*, vol. 12, no. 1, p. 15889, 2022.

[157] 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*, vol. 22, no. 5, pp. 1589–1604, 2017.

[158] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[159] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, *et al.*, "A survey of uncertainty in deep neural networks," *arXiv preprint arXiv:2107.03342*, 2021.

[160] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[161] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring fairness in machine learning to advance health equity," *Annals of internal medicine*, vol. 169, no. 12, pp. 866–872, 2018.

[162] C.-Y. Chang, J. Yuan, S. Ding, Q. Tan, K. Zhang, X. Jiang, X. Hu, and N. Zou, "Towards fair patient-trial matching via patient-criterion level fairness constraint," *arXiv preprint arXiv:2303.13790*, 2023.

[163] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, "Mining electronic health records (ehrs) a survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, pp. 1–40, 2018.

[164] M. Sung, S. Hahn, C. H. Han, J. M. Lee, J. Lee, J. Yoo, J. Heo, Y. S. Kim, and K. S. Chung, "Event prediction model considering time and input error using electronic medical records in the intensive care unit: Retrospective study," *JMIR medical informatics*, vol. 9, no. 11, p. e26426, 2021.

[165] A. Pakbin, P. Rafi, N. Hurley, W. Schulz, M. H. Krumholz, and J. B. Mortazavi, "Prediction of icu readmissions using data at patient discharge," in *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pp. 4932–4935, IEEE, 2018.

[166] K. Yin, W. K. Cheung, B. C. Fung, and J. Poon, "Learning inter-modal correspondence and phenotypes from multi-modal electronic health records," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 9, pp. 4328–4341, 2020.

[167] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," *Journal of Biomedical Informatics*, vol. 100, p. 100057, 2019.

[168] 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, 2020.

[169] M. Tayefi, P. Ngo, T. Chomutare, H. Dalianis, E. Salvi, A. Budrionis, and F. Godtliebsen, "Challenges and opportunities beyond structured data in analysis of electronic health

records," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 13, no. 6, p. e1549, 2021.

[170] O. Hoekstra, W. Hurst, and J. Tummers, "Healthcare related event prediction from textual data with machine learning: A systematic literature review," *Healthcare Analytics*, vol. 2, p. 100107, 2022.

[171] H. Duan, Z. Sun, W. Dong, K. He, and Z. Huang, "On clinical event prediction in patient treatment trajectory using longitudinal electronic health records," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 2053–2063, 2019.

[172] D. Zhang, C. Yin, J. Zeng, X. Yuan, and P. Zhang, "Combining structured and unstructured data for predictive models: a deep learning approach," *BMC medical informatics and decision making*, vol. 20, no. 1, pp. 1–11, 2020.

[173] N. Tomašev, N. Harris, S. Baur, A. Mottram, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, V. Magliulo, *et al.*, "Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records," *Nature Protocols*, vol. 16, no. 6, pp. 2765–2787, 2021.

[174] Z. D. Cohen, J. Delgadillo, and R. J. DeRubeis, "Personalized treatment approaches.," 2021.

[175] S. C. Mathews, M. J. McShea, C. L. Hanley, A. Ravitz, A. B. Labrique, and A. B. Cohen, "Digital health: a path to validation," *NPJ digital medicine*, vol. 2, no. 1, p. 38, 2019.

[176] C. D. Kemp and J. V. Conte, "The pathophysiology of heart failure," *Cardiovascular Pathology*, vol. 21, no. 5, pp. 365–371, 2012.

[177] D. S. Lee, S. E. Straus, M. E. Farkouh, P. C. Austin, M. Taljaard, A. Chong, C. Fahim, S. Poon, P. Cram, S. Smith, *et al.*, "Trial of an intervention to improve acute heart failure outcomes," *New England Journal of Medicine*, vol. 388, no. 1, pp. 22–32, 2023.

[178] G. Savarese and L. H. Lund, "Global public health burden of heart failure," *Cardiac failure review*, vol. 3, no. 1, p. 7, 2017.

[179] F. R. Vogenberg, "Predictive and prognostic models: implications for healthcare decision-making in a modern recession," *American health & drug benefits*, vol. 2, no. 6, p. 218, 2009.

[180] S. Oparil, M. C. Acelajado, G. L. Bakris, D. R. Berlowitz, R. Cífková, A. F. Dominiczak, G. Grassi, J. Jordan, N. R. Poulter, A. Rodgers, *et al.*, "Hypertension," *Nature reviews Disease primers*, vol. 4, no. 1, pp. 1–21, 2018.

[181] S. Kumar, M. H. Selim, and L. R. Caplan, "Medical complications after stroke," *The Lancet Neurology*, vol. 9, no. 1, pp. 105–118, 2010.

[182] I. Chabot, J. Moisan, J.-P. Grégoire, and A. Milot, "Pharmacist intervention program for control of hypertension," *Annals of Pharmacotherapy*, vol. 37, no. 9, pp. 1186–1193, 2003.

[183] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.

[184] C. Lu, T. Han, and Y. Ning, "Context-aware health event prediction via transition functions on dynamic disease graphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 4567–4574, 2022.

[185] T. M. Seinen, E. A. Fridgeirsson, S. Ioannou, D. Jeannetot, L. H. John, J. A. Kors, A. F. Markus, V. Pera, A. Rekkas, R. D. Williams, *et al.*, "Use of unstructured text in prognostic clinical prediction models: a systematic review," *Journal of the American Medical Informatics Association*, vol. 29, no. 7, pp. 1292–1302, 2022.

[186] C. Lu, C. K. Reddy, P. Chakraborty, S. Kleinberg, and Y. Ning, "Collaborative graph learning with auxiliary text for temporal event prediction in healthcare," *arXiv preprint arXiv:2105.07542*, 2021.

[187] C. Mugisha and I. Paik, "Pneumonia outcome prediction using structured and unstructured data from ehr," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2640–2646, IEEE, 2020.

[188] J. Qiu, L. Li, J. Sun, J. Peng, P. Shi, R. Zhang, Y. Dong, K. Lam, F. P.-W. Lo, B. Xiao, *et al.*, "Large ai models in health informatics: Applications, challenges, and the future," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[189] Y. Kim, X. Xu, D. McDuff, C. Breazeal, and H. W. Park, "Health-llm: Large language models for health prediction via wearable sensor data," *arXiv preprint arXiv:2401.06866*, 2024.

[190] Y. Zhao, C.-Y. Lin, K. Zhu, Z. Ye, L. Chen, S. Zheng, L. Ceze, A. Krishnamurthy, T. Chen, and B. Kasikci, "Atom: Low-bit quantization for efficient and accurate llm serving," *arXiv preprint arXiv:2310.19102*, 2023.

[191] S. Kim, S. Yun, H. Lee, M. Gubri, S. Yoon, and S. J. Oh, "Propile: Probing privacy leakage in large language models," *arXiv preprint arXiv:2307.01881*, 2023.

[192] J. Zhang, R. Krishna, A. H. Awadallah, and C. Wang, "Ecoassistant: Using llm assistant more affordably and accurately," *arXiv preprint arXiv:2310.03046*, 2023.

[193] A. Zhang, L. Xing, J. Zou, and J. C. Wu, "Shifting machine learning for healthcare from development to deployment and from models to data," *Nature Biomedical Engineering*, vol. 6, no. 12, pp. 1330–1345, 2022.

[194] Y. Li, H. Wen, W. Wang, X. Li, Y. Yuan, G. Liu, J. Liu, W. Xu, X. Wang, Y. Sun, *et al.*, "Personal llm agents: Insights and survey about the capability, efficiency and security," *arXiv preprint arXiv:2401.05459*, 2024.

[195] A. Belyaeva, J. Cosentino, F. Hormozdiari, K. Eswaran, S. Shetty, G. Corrado, A. Carroll, C. Y. McLean, and N. A. Furlotte, "Multimodal llms for health grounded in individual-specific data," in *Workshop on Machine Learning for Multimodal Healthcare Data*, pp. 86–102, Springer, 2023.

[196] Y. Han, C. Zhang, X. Chen, X. Yang, Z. Wang, G. Yu, B. Fu, and H. Zhang, "Chartllama: A multimodal llm for chart understanding and generation," *arXiv preprint arXiv:2311.16483*, 2023.

[197] H. Nguyen and J. Patrick, "Text mining in clinical domain: Dealing with noise," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 549–558, 2016.

[198] M. Moradi, K. Blagec, and M. Samwald, "Deep learning models are not robust against noise in clinical text," *arXiv preprint arXiv:2108.12242*, 2021.

[199] Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, J. Gao, and A. Zhang, "Deep patient similarity learning for personalized healthcare," *IEEE transactions on nanobioscience*, vol. 17, no. 3, pp. 219–227, 2018.

[200] L. Ma, J. Gao, Y. Wang, C. Zhang, J. Wang, W. Ruan, W. Tang, X. Gao, and X. Ma, "Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 825–832, 2020.

[201] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[202] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.

[203] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," *Advances in neural information processing systems*, vol. 29, 2016.

[204] 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," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1903–1911, 2017.

[205] M. Adamczak, M. Zeier, R. Dikow, and E. Ritz, "Kidney and hypertension," *Kidney International*, vol. 61, pp. S62–S67, 2002.

[206] K. Altemose and J. M. Dionne, “Neonatal hypertension: concerns within and beyond the neonatal intensive care unit,” *Clinical and Experimental Pediatrics*, vol. 65, no. 8, p. 367, 2022.