Mitigating Health Disparities in EHR via Deconfounder

Zheng Liu
zliu212@uic.edu
University of Illinois at Chicago
Chicago, USA

Xiaohan Li
xli241@uic.edu
University of Illinois at Chicago
Chicago, USA

Philip Yu
psyu@uic.edu
University of Illinois at Chicago
Chicago, USA

ABSTRACT
Health disparities, or inequalities between different patient demographics, are becoming a crucial issue in medical decision-making, especially in Electronic Health Record (EHR) predictive modeling. In order to ensure the fairness of sensitive attributes, conventional studies mainly adopt calibration or re-weighting methods to balance the performance on among different demographic groups. However, we argue that these methods have some limitations. First, these methods usually mean making a trade-off between the model’s performance and fairness. Second, many methods attribute the existence of unfairness completely to the data collection process, which lacks substantial evidence. In this paper, we provide an empirical study to discover the possibility of using deconfounder to address the disparity issue in healthcare. Our study can be summarized in two parts. The first part is a pilot study demonstrating the exacerbation of disparity when unobserved confounders exist. The second part proposed a novel framework, Parity Medical Deconfounder (PriMeD), to deal with the disparity issue in healthcare datasets. Inspired by the deconfounder theory, PriMeD adopts a Conditional Variational Autoencoder (CVAE) to learn latent factors (substitute confounders) for observational data, and extensive experiments are provided to show its effectiveness.

CCS CONCEPTS
• Applied computing → Health informatics; • Social and professional topics → User characteristics; • Computing methodologies → Machine learning.

KEYWORDS
health disparity, fairness, deconfounder, deep generative model

1 INTRODUCTION
Machine learning models have demonstrated the promising potential on Electronic Health Records (EHRs), such as risk prediction [7, 30], auxiliary diagnosis [37], and automated prescription [48]. However, recent evidence shows these models also exacerbate bias and disparities in healthcare, which raises considerable concern and criticism [4–6, 14, 31, 34, 49]. Many studies have shown that a machine learning model may provide disparate results for people with different backgrounds. For example, for patients of different races, the accuracy and quality of a machine learning model’s prediction may vary significantly [32]. This disparity reduces the utility of machine learning models and is especially detrimental to disadvantaged and underrepresented populations [6, 34].

To deal with this fairness issue, conventional studies adopt certain criteria such as equalized odds [15] or counterfactual fairness [27] to mitigate the disparities [23, 33, 35]. However, we argue that there are limitations to these approaches. First, maintaining fairness usually means sacrificing some precision of the model, especially when the model penalizes/calibrates the majority of datapoints [11, 35]. If a model’s performance deteriorates after adopting certain fairness criteria, the utility of the model does not necessarily improve. Second, many methods attribute the existence of unfairness completely to the data collection process, which lacks substantial evidence. For example, many causality-based fairness methods assumes “no unobserved confounders” assumption [10, 24], which usually doesn’t hold on EHR data.

To address these issues, we first focus on the causes of health disparity and construct a fairness model on this basis. Consider a predictive task that using the observation (including both sensitive attributes and other clinical observations) to predict the outcomes of patients. In this paper we attribute the cause of health disparity to two factors: the imbalance of sensitive attributes and the existence of unobserved confounders (e.g., ethnicity, social/economic status, etc.) that affects both clinical observations and outcomes.

Then, we propose a novel framework, Parity Medical Deconfounder (PriMeD), to deal with the above two factors. Inspired by the deconfounder theory [43, 44] and the Inverse Propensity Weighting (IPW) methods [16], PriMeD can provide accurate and fair prediction by addressing the above two factors of health disparity. In detail, PriMeD is a two-stage supervised framework that infers unobserved confounders in the first stage and makes predictions in the second stage. In the first stage, PriMeD resorts to a weighted Conditional Variational Auto-encoder (CVAE) [41] to learn fair latent representation for observational data, and the weight of each datapoint is the probability of the sensitive attributes exist. Therefore, the imbalance of sensitive attributes is addressed in this step, and according to the deconfounder theory, we regard the fair representation as substitute confounders. In the second stage, PriMeD adopts a self-attentive deep neural network to predict medical outcomes based on the clinical observations, sensitive attributes and the fair representation of the first stage. The goal of this step is to
We consider an EHR with the latent structure shown in Figure 1a. In this causal diagram, an EHR consists of three parts: the sensitive attributes \(X\), the clinical observations \(S\), and the outcome/predictor \(Y\). We also assume the existence of unobserved confounders \(Z\) (e.g., living habits, social/economic status, genotypes, etc.) in the diagram. Please note that although \(Z\) and \(S\) may be associated with \(B\) in some scenarios, in this paper we don’t consider this effect.

In practice, we have an EHR dataset \(D = \{(x_i, s_i, y_i)\}_{i=1}^N\) where \(x, s, y\) corresponds to \(X, S, Y\) in Figure 1a. Our goal is to achieve higher accuracy in prediction, and theoretically any predictive model is eligible here. In this paper, we adopt an attention neural network to achieve this goal.

In summary, in this paper our contribution is listed as follows:

- To the best of our knowledge, this paper is the first one using deconfounder theory to address health disparities in EHR analysis.
- We propose two causes of health disparity and design a model based on this judgement. Parity Medical Deconfounder (PriMeD), is the novel model we proposed to address the health disparity in EHR modeling.
- Extensive experiments are provided to show the superiority of PriMeD in achieving equalized odds[15] and making accurate decisions.

2 METHOD

2.1 Problem Definition

We consider an EHR with the latent structure shown in Figure 1a. In this causal diagram, an EHR consists of three parts: the sensitive attributes \(S\) (e.g., gender, race, etc.), the clinical observations \(X\) and the outcome/predictor \(Y\). We also assume the existence of unobserved confounders \(Z\) (e.g., living habits, social/economic status, genotypes, etc.) in the diagram. Please note that although \(Z\) may be associated with \(B\) in some scenarios, in this paper we don’t consider this effect.

2.2 PriMeD

Inspired by the deconfounder theory, PriMeD consists of two stages. The first stage is to address the disparity by learning fair representations, and the second stage is to make accurate predictions.

Stage 1: weighted Conditional Variational Auto-encoder (CVAE). In the first stage, PriMeD addresses the confounding effect of \(Z\) and \(S\) by learning fair latent representations. Here we adopt a weighted Conditional Variational Auto-encoder (CVAE) [41] to achieve this goal. As a variant of Variational Autoencoder (VAE) [25], CVAE regards the observations as consequences of a latent factor and a condition. In EHR modeling, the clinical observations \(X\) are the effect of latent factor \(Z\) and sensitive attributes \(S\), which is identical to the CVAE setting.

CVAE introduces a variational posterior \(q_\phi(z|x, s)\) as a recognition network, to deal with the intractability of true posterior \(p_\theta(z|x, s)\) in maximum likelihood inference. It exhibits an autoencoder structure, with the recognition network \(q_\phi(z|x, s)\) as encoder and the generation network \(p_\theta(x|z, s)\) as decoder. When optimizing CVAE, the objective is designed to maximize the variational lower bound (ELBO) of the log-likelihood, which is written as

\[
\log p_\theta(x|s) \geq -KL(q_\phi(z|x, s)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x, s)}[\log p_\theta(x|z, s)].
\]

To approximate the second term, we draw \(L\) samples \(z^{(l)}_i, l = 1, \cdots, L\) from the recognition distribution \(q_\phi(z|x, s)\), and rewrite the empirical objective for datapoint \((x, s)\) as

\[
\mathcal{L}(x, s; \theta, \phi) = -KL(q_\phi(z|x, s)||p_\theta(z)) + \frac{1}{L} \sum_{l=1}^L \log p_\theta(x|z^{(l)}_i, s).
\]

where the first term minimizes the difference between the posterior and the prior of \(z\), and the second term minimizes the difference between the input and output of CVAE. In this paper, we set \(p_\theta(z)\) as standard Gaussian distribution. By adopting such an architecture, the confounding effect of \(X\) and \(S\) can be addressed.

Moreover, to deal with the imbalanced distribution of \(S\), in the training process we assign a weight \(\omega_s\) for each datapoint \((x, s)\) in \(D\). The intuition of using the weight is straightforward: since the distribution of sensitive attributes \(s\) in the dataset is usually imbalanced, we would like to assign larger weights for rarer \(s\) to mitigate this imbalance.

To obtain \(\omega_s\), suppose there are \(J\) sensitive attributes \(s = [s_1, \cdots, s_J]\) for each patient in the dataset. The training weight \(\omega_s\) for patient with sensitive attributes \(s\) is defined as:

\[
\omega_s = \frac{1}{\prod_{j=1}^J freq(s_j)}
\]

where \(freq(s_j)\) is the frequency of certain attribute that occurs in the dataset. Hence, the more frequent one attribute occurs in the sensitive attribute vector \(s\), the smaller the corresponding \(\omega_s\) is. Since we usually won’t focus on too many sensitive attributes, such an approximation of propensity score is enough to mitigate the imbalance.

Stage 2: Attentive Prediction

To mitigate the imbalance, we would like to assign larger weights for rarer \(s\). In this paper, we adopt an attention architecture to achieve this goal.

Inspired by the deconfounder theory, PriMeD consists of two stages. The first stage is to address the disparity by learning fair representations, and the second stage is to make accurate predictions.
Table 1: Information used in our experiments.

| Datasets | MIMIC | Hip & Knee |
|----------|-------|------------|
| Sensitive Attributes | Insurance, Ethnicity | Gender | Race, Age |
| Clinical Features | Demographics | Height, Weight | Clinical events, Lab events | Position # |
| | Procedures, Diagnoses | | | 20-37, 51-63 |
| Outcomes | Mortality | Re-operation | Re-admission | DOptoDis |

imbalance of data distribution. The objective for the whole dataset can be written as

$$\mathcal{L}_{CV_{AE}}(\mathcal{D}; \theta, \phi) = \sum_{i=1}^{N} \omega_x \mathcal{L}(x_i, s_i; \theta, \phi)$$  \hspace{1cm} (4)

where $(x_i, s_i)$ is $x$ and $s$ of the $i$-th datapoint in dataset $\mathcal{D}$.

**Stage 2: Prediction with Attention.** Once the CVAE is trained successfully, the learned latent variable $z$ can be viewed as substitute confounders and be used to derive accurate and unbiased predictions. Theoretically, any model taking $x$, $s$, and $z$ as inputs is eligible in this stage. Here we use a deep learning model with attention layers to achieve higher prediction performance. Suppose the length of sensitive attributes $z$ is $K$, and the clinical features $x = [x_1, x_2, \ldots, x_M]$ is with $M$ digits. We use $z$ to learn attention weights $w_x = [w_1, w_2, \ldots, w_M] \in \mathbb{R}^{K \times M}$ in order to determine which digit of $x$ is more important in prediction. Formally, PriMeD learns $w_x$ with a simple attention module as follows:

$$w_x = \text{softmax}(W_x^{att} z^\top)$$ \hspace{1cm} (5)

where $W_x^{att}$ is the $M \times K$ parameter matrix and $\text{softmax}()$ is the softmax function. Similarly, we also have $w_s = \text{softmax}(W_s^{att} z^\top)$ to learn attention weights for sensitive attributes $s$.

Finally, the processed features are input into a feed-forward neural network to derive the final decision:

$$\hat{y} = \text{MLP}(x \odot w_x^\top ; s \odot w_s^\top ; z),$$ \hspace{1cm} (6)

where $[\cdot ; \cdot]$ denotes the concatenation operation, $\odot$ denotes the element-wise multiplication operation, and $\text{MLP}(\cdot)$ is a two-layer MLP. In this stage, by incorporating the latent factor $z$ in the prediction, the model can calibrate the bias from observational data and derive a higher-quality prediction.

### 3 EXPERIMENTS

In this section, we conduct experiments to evaluate the performance of our proposed PriMeD.

#### 3.1 Datasets

We use three real-world EHR datasets to measure PriMeD’s performance.

- **MIMIC-III**. MIMIC-III contains comprehensive clinical data of patients admitted to the Beth Israel Deaconess Medical Center. MIMIC-III contains EHRs associated with 46,520 patients, including over 20 tables such as medical events, diagnoses, prescriptions, etc.

- **Hip & Knee**. Hip & Knee two subsets of the National Surgical Quality Improvement Program (NSQIP) project \(^2\) \cite{8}. NSQIP provides clinical observations and observations for patients taking surgical operations to track surgical complications after operations. They are associated with patients receiving Hip Arthroplasty (CPT 27130) and Knee Arthroplasty (CPT 27447) surgical operations. There are 96,441 and 156,292 pieces of data in them respectively, and each data point contains demographics, pre-operative features, and the outcome of the surgery for a patient.

**data preprocessing.** According to the documentation of MIMIC-III \(^3\) and NSQIP \(^5\), we determine the columns/tables used in our experiments (shown in Table 1). In the MIMIC-III dataset, we regard the patient’s insurance status, ethnicity and gender as sensitive attributes, and use clinical events to predict mortality. In the Hip & Knee dataset, we take gender, race and age as sensitive attributes and use BMI (inferred from height and weight) and multiple clinical observations to predict the readmission, reoperation, and whether there is a prolonged hospitalization after the operation (DOptoDis > 5).

#### 3.2 Evaluation Metrics & Implementation Details

Since all tasks are binary classifications, we use Area Under the Receiver Operating Characteristic curve (AUROC) to measure its accuracy in classification. We implement the model using Pytorch 1.10 and adopt the $10^{-4}$ as the learning rate and $5 \times 10^{-4}$ as the weight decay. The ratio of training, validating, and testing set is 7:2:1.

#### 3.3 Comparison Experiments

In this subsection, we compare our model with several baseline models listed below.

- **DNN**. We adopt a deep neural network to make accurate predictions on observational data. This method is served as a baseline to show the original disparity without interference.

- **Re-weighting** \cite{23}. Kamiran et al. propose a pre-processing method based on the dataset re-weighting to remove bias from the dataset. Re-weighting can reduce the discrimination while maintaining the overall positive class probability for the training set. We use this method to assign weights for the data, while using the same model as **DNN** to make the prediction.

- **CE Odds** \cite{35}. Pleiss et al. propose a post-processing method based on calibration constraints to minimize error disparity while maintaining calibrated probability estimates.

---

\(^1\)https://physionet.org/content/mimiciii/1.4/

\(^2\)https://www.facs.org/quality-programs/data-and-registries/acq-nsqip/

\(^3\)https://physionet.org/content/mimiciii/1.4/

\(^4\)https://www.facs.org/quality-programs/data-and-registries/acs-nsqip/participant-use-data-file/
Table 2: Accuracy of PriMeD and other baselines in prediction.

| Dataset | MIMIC-III | Knee | Hip |
|---------|-----------|------|-----|
| Task    | Mortality | Reoperation | Readmission | Prolonged Hospitalization | Reoperation | Readmission | Prolonged Hospitalization |
| DNN     | 0.6984    | 0.6752 | 0.6342 | 0.7061 | 0.7044 | 0.6448 | 0.7403 |
| Re-weighting | 0.6553    | 0.6457 | 0.5973 | 0.6590 | 0.6793 | 0.6125 | 0.6983 |
| CE odds | 0.6810    | 0.6579 | 0.6192 | 0.6773 | 0.6889 | 0.6313 | 0.7045 |
| WFC     | 0.6741    | 0.6481 | 0.6008 | 0.6717 | 0.6865 | 0.6217 | 0.7102 |
| RFC     | 0.6889    | 0.6630 | 0.6289 | 0.6928 | 0.6956 | 0.6375 | 0.7363 |
| FuCS    | 0.6944    | 0.6774 | 0.6375 | 0.7024 | 0.7029 | 0.6426 | 0.7345 |
| PriMeD  | 0.7013    | 0.6821 | 0.6413 | 0.7146 | 0.7059 | 0.6453 | 0.7457 |

- WFC [19]. Wasserstein Fair Classification is a post-processing method that enforces independence between the classifier outputs and sensitive information by minimizing Wasserstein-1 distances.
- RFC [1]. Agarwal et al. propose an in-processing method, focusing on reducing fair classification to a sequence of cost-sensitive classification problems to achieve the lowest error subject to the desired constraints.
- FuCS [38]. Rezaei et al. propose a pre-processing method to guarantee fairness under covariate shift in which the covariates change while the conditional label distribution remains the same.

### 3.4 Performance of Prediction

Table 2 shows the performance of all baselines together with PriMeD on all three tasks. From the table, we can observe that PriMeD outperforms all other baselines significantly in AUROC. It is because the architecture of PriMeD is specially designed for EHRs, while other baselines are universal classifiers. PriMeD also outperforms the vanilla classifier, DNN, on all tasks, showing its superiority in modeling EHR. Apart from PriMeD, FuCS achieves the second-best performance, which may due to its ability to deal with non-iid data. When the confounder is unobserved, EHR data may exhibit multiple distributions and is not iid. FuCS also outperforms DNN on some tasks, indicating its ability to learn knowledge from non-iid data. Then, RFC and CE odds achieve lower performance at the price of guaranteeing equalized odds in prediction. WFC and Re-weighting method have the worst performances. This fact indicates their means of controlling fairness clearly prevent them from learning from observational data.

If we compare different datasets/tasks, we will find the difficulty of tasks is different as well. As for the reoperation task, the differences between the performances of methods are relatively small, indicating the confounding effect on this predictor is relatively weaker. The difference between methods on the prolonged hospitalization task is quite large, indicating this predictor is strongly influenced by the confounding effect.

### 3.5 Mitigate Health Disparity

In this subsection, we demonstrate the disparity of prediction between different subgroups. Table 3 demonstrates the extent of disparities in prediction with respect to different datasets, predictors and baseline models. In this table, the Insurance column shows the difference in AUROC between patients with public and private insurances, the Ethnicity and Race columns show the difference between White and non-White patients, and the Age column shows the difference between patients with ages above and below 65.

From Table 3, we can observe that PriMeD can effectively reduce the difference between subgroups in prediction. It achieves the fairest performance across all datasets and all metrics. Apart from PriMeD, we can observe that FuCS and RFC achieve the second-best performance, which means they can balance the performance on different demographic subgroups. From the table, we also observe that the disparity in age is more stubborn than the disparity of race, which means age is a stronger confounder that can cause bias.

### 3.6 Ablation Study

In order to show the effectiveness of each stage of PriMeD, Table 4 demonstrates the performance of each module of PriMeD. In this module, PriMeD-Stage 1 means the performance of prediction while only using the stage 1 model (the CVAE). In PriMeD-Stage 1 we use the latent vector $z$ the CVAE learned to make classifications. Similarly, PriMeD-Stage 2 means simply using a deep neural network to make the prediction. In PriMeD-Stage 2, we use the input vector $b$ and $x$ to substitute the position of $z$ in the stage 2 model. In Table 4, we use AUROC to describe the performance in prediction and use the difference between White and non-White patients (denoted by Diff_Race) to measure the health disparity. We can observe that the CVAE can learn relatively fair predictions while the accuracy of classification is low. The attention neural network can achieve quite an accurate prediction which is similar to the DNN baseline, while the disparity issue is unaddressed.

### 4 RELATED WORKS

The related works of this paper consist of three parts: the observation of healthcare disparity, the review of fairness in machine learning, and a brief introduction of the deconfounder theory.

#### 4.1 Healthcare Disparity

With the increase in the number of machine learning applications in the healthcare domain, more concerns have been raised regarding the potential ethical issue of these models [4, 6, 34]. Recently, many studies have shown that these models worsen existing health disparities. Health disparity means the inequity for different patient subgroups when providing healthcare services [2, 12]. Although
Traditionally, fairness avoids any prejudice or favoritism towards any individual or group in the decision-making process. Based on this principle, many fairness metrics are proposed and can be categorized mainly into three classes. One standard fairness metric is group fairness, such as demographic parity, which requires the probability of a positive prediction to be the same for each group [46]. Another fairness metric is individual fairness, such as counterfactual fairness. It requires the probability of a positive prediction to be the same for both factual and counterfactual datapoints [27]. In this paper, our goal is to minimize the difference in utility between patient subgroups, which is similar to [17, 33, 36] and is different from either group or individual fairness.

Based on the above definitions of fairness, there are three types of algorithms to address the fairness of machine learning models: the pre-processing, in-processing, and post-processing methods. Pre-processing methods either change the label of training data or assign weights for data before the training process [22, 23]. Post-processing methods conduct calibration of prediction results after training. Some of them calibrate the prediction based on a holdout set, and others are based on certain constraints to ensure fairness [9, 35]. In-processing techniques develop model architectures to remove discrimination during the model training process [9, 29, 39]. Our model adopts the in-processing technique, because it is more flexible in dealing with complex latent data structures.

### 4.3 Causal Inference with Unobserved Confounders

In this part, we mainly discuss the deconfounder theory. Deconfounder [43, 44] is a theory to estimate unbiased treatment effects for observational data with the setting of multiple causes. Due to unobserved confounders of many tasks, traditional methods can hardly learn unbiased knowledge from the observational data. However, suppose we can observe multiple causes of the outcome. In this case, the dependencies between causes can be used to infer latent variables. The latent variables can be used as substitutes for the hidden confounders. Therefore, Wang & Blei propose a two-stage architecture to deal with the unobserved confounders. Recently, deconfounder has been applied to many areas such as recommender systems [45] and medical treatment estimation [47]. This paper demonstrates that deconfounder can also be used in the fairness domain to reduce inequity in prediction.

### 5 CONCLUSION & FUTURE DIRECTIONS

This paper analyzes the disparity between different demographic groups and proposes a solution to address it. Our goal is to minimize the disparity of different demographics while maintaining a high utility for machine learning healthcare applications. To achieve this goal, we propose our PriMeD to derive unbiased predictions. By incorporating CVAE as a module to infer latent factors for patients, PriMeD can naturally correct the disparity in the first stage of the model, while preserving high prediction accuracy. Experiments conducted on three real-world datasets have shown the superiority of PriMeD over other baselines, and the visualization of the learned latent factor further demonstrates its ability to learn fair representation without the influence of sensitive attributes.

### 6 ACKNOWLEDGEMENT

This work is supported in part by NSF under grants III-1763325, III-1909323, III-2106758, and SaTC-1930941.

**REFERENCES**

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In International Conference on Machine Learning. PMLR, 60–69.
