FairEHR-CLP: Towards Fairness-Aware Clinical Predictions with Contrastive Learning in Multimodal Electronic Health Records

Yuqing Wang YWANG216@STANFORD.EDU
Stanford University

Malvika Pillai MPILLAI@STANFORD.EDU
Stanford University

Yun Zhao YUNZHAO20@META.COM
Meta Platforms, Inc.

Catherine Curtin CCURTIN@STANFORD.EDU
Stanford University

Tina Hernandez-Boussard BOUSSARD@STANFORD.EDU
Stanford University

Abstract

In the high-stakes realm of healthcare, ensuring fairness in predictive models is crucial. Electronic Health Records (EHRs) have become integral to medical decision-making, yet existing methods for enhancing model fairness restrict themselves to unimodal data and fail to address the multifaceted social biases intertwined with demographic factors in EHRs. To mitigate these biases, we present *FairEHR-CLP*: a general framework for **Fair**ness-aware Clinical **P**redictions with **C**ontrastive **L**earning in **EHR**s. FairEHR-CLP operates through a two-stage process, utilizing patient demographics, longitudinal data, and clinical notes. First, synthetic counterparts are generated for each patient, allowing for diverse demographic identities while preserving essential health information. Second, fairness-aware predictions employ contrastive learning to align patient representations across sensitive attributes, jointly optimized with an MLP classifier with a softmax layer for clinical classification tasks. Acknowledging the unique challenges in EHRs, such as varying group sizes and class imbalance, we introduce a novel fairness metric to effectively measure error rate disparities across subgroups. Extensive experiments on three diverse EHR datasets on three tasks demonstrate the effectiveness of FairEHR-CLP in terms of fairness and utility compared with competitive baselines. FairEHR-CLP represents an advancement towards ensuring both accuracy and equity in predictive healthcare models. Our code is available at https://github.com/EternityYW/FairEHR-CLP.

1. Introduction

The growing availability of Electronic Health Records (EHRs) holds significant potential for enhancing healthcare delivery and patient outcomes (Zhao et al., 2021a; Wang et al., 2022b). However, their use in predictive modeling raises substantial challenges, particularly in ensuring algorithmic fairness and addressing inherent data biases (Chen et al., 2023; Giovanola and Tiribelli, 2023). EHR data often mirror social and systemic biases, which if unaddressed, can perpetuate inequalities in healthcare outcomes. For example, studies have shown racial disparities in healthcare, such as Black patients being 40% less likely to receive pain medication than White patients for similar conditions (Lee et al., 2019). Such biases, when ingrained in training data, can lead models to perpetuate or even exacerbate these inequalities, resulting in disparities in patient care based on race, gender, or socioeconomic status. In a field where decisions can have life-altering consequences, it is crucial to ensure that predictive tools do not inadvertently disadvantage marginalized patient groups (Vela et al., 2022). Therefore, developing fair and effective predictive models is essential.

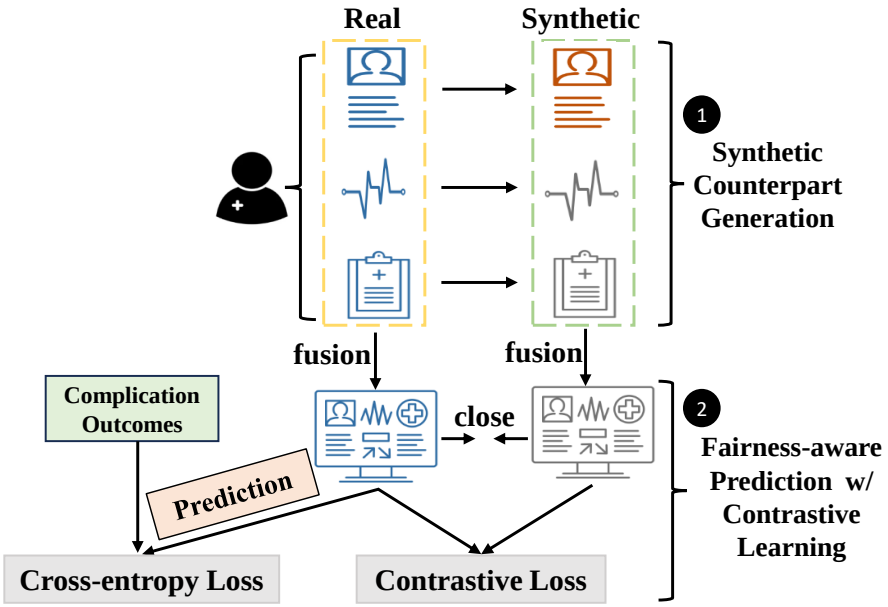


Figure 1: Overview of our FairEHR-CLP framework.

Existing methods to enhance fairness in EHR predictive models fall into three categories, each with respective limitations. Pre-processing techniques that alter training data distributions, such as sampling (Iosifidis and Ntoutsi, 2018) and perturbation (Wang et al., 2022c) can lead to overfitting or data distortion. Post-processing methods, involving modifications after training (Du et al., 2021) or prediction relabeling (Lohia et al., 2019), are slow and resource-intensive. In-processing strategies like loss function regularization (Kim et al., 2018) and adversarial training (Yang et al., 2023), overlook the interplay and complex nature of social biases (Wang et al., 2021; Boyd et al., 2023; Rajendran et al., 2024). The intricacy involved in these techniques highlights a fundamental question: *How can we develop a fair prediction model that effectively addresses the varied social biases from demographic factors in EHRs?*

To address this question, developing a fair prediction model that utilizes the value of demographic data as predictors while minimizing associated social biases is essential. Consider a scenario where a model assesses patients with similar health issues but from varied demographics, such as two individuals with cardiovascular symptoms, differing in gender and ethnicity. By applying contrastive analysis to these cases, the model can identify clinical patterns that span across demographic lines, focusing on health similarities. This strategy strengthens the model's ability to make unbiased, clinically relevant recommendations, prioritizing health factors over demographic differences. The above process aligns with the principles of contrastive learning (CL), a prominent representation learning method that differentiates similar and dissimilar instances within an embedding space (Chen et al., 2020; Chuang et al., 2020; Zhang et al., 2022; Sun et al., 2023; Ge et al., 2023). We aim to harness CL in balancing the use of demographics for informative predictions and the imperative for bias mitigation.

To this end, we introduce a general framework for Fairness-aware Clinical Predictions with Contrastive Learning in EHRs, which we call *FairEHR-CLP*. The framework involves two distinct stages: first, synthetic counterpart generation creates synthetic instances for each patient, representing varied demographics while preserving vital health data. The second stage involves fairness-aware predictions using CL, which aims to minimize the representation distance between real patients and their synthetic counterparts who share similar health conditions but differ demographically, in tandem with a multi-layer perceptron (MLP) classifier equipped with a softmax layer for downstream classification tasks. Figure 1 presents an overview of our FairEHR-CLP framework.

In our experiments, we incorporate patient demographics, longitudinal data, and clinical notes into the FairEHR-CLP framework for clinical predictions. We focus on five sensitive attributes linked to social biases: gender, race, ethnicity, age, and socioeconomic status (represented by insurance type). We demonstrate the effectiveness of our method across three diverse EHR datasets: STARR (Sun et al., 2021), MIMIC-III (Johnson et al., 2016), and MIMIC-IV (Johnson et al., 2023), focusing on surgical patient outcomes, which are often subject to social bias (Raso et al., 2023). We consider three binary classification tasks, identifying delirium, opioid use disorder (OUD), and 30-day readmission, all of which have a direct impact on postoperative care. Our extensive experiments show that FairEHR-CLP not only outperforms existing debiasing methods in terms of fairness but also achieves competitive predictive performance when compared to standard classification baselines.

To summarize, our contributions are three-fold:

(1) We develop FairEHR-CLP, a general fairness-aware clinical prediction framework that employs contrastive learning in multimodal EHRs, aiming at mitigating social biases arising from demographic factors.

(2) We propose a new fairness metric, the Error Distribution Disparity Index (EDDI), by quantifying the deviation in error rates for each subgroup from the overall error rate, particularly relevant in clinical settings with diverse group sizes and class imbalance.

(3) Extensive experiments on three large-scale EHR datasets across three classification tasks illustrate the effectiveness of our proposed method in terms of fairness and utility compared with multiple baselines.

Generalizable Insights about Machine Learning in the Context of Healthcare

Healthcare data often contain biases due to incomplete records, inaccuracies, inconsistencies, and an overrepresentation of individuals with structural privileges. These biases can be perpetuated by machine learning models, leading to disparities in predictions across different demographic groups. Additionally, healthcare data encompass a range of modalities like clinical notes, lab measurements, and demographic details. Leveraging these diverse data types can enhance model fairness by providing a more complete representation of patients. Our study underscores the critical importance of fairness in healthcare predictive models by proposing a framework, FairEHR-CLP, for fairness-aware clinical predictions that employs contrastive learning (CL) in multimodal EHRs. This method effectively reduces social biases related to demographics in the data through fairness evaluation while minimizing performance loss. The CL-based framework can be applied across clinical domains and is scalable to other types of biases in EHR data, offering a robust solution for fair machine learning applications in healthcare.

2. Related Work

In this section, we explore existing methods to mitigate bias and enhance fairness in EHRs, review CL applications in EHRs, and discuss fairness evaluation approaches.

Bias and Fairness. EHRs, rich in patient data, often exhibit systemic biases, stemming from demographic, socioeconomic, and access disparities (Zhao et al., 2021b; Chin et al., 2023; Wang et al., 2023; Rajendran et al., 2024). Such biases in EHRs risk being reinforced or exacerbated by algorithms trained on these datasets, potentially harming underrepresented groups. To combat this, recent research has focused on reducing algorithmic bias. Representative approaches include adversarial training (Yang et al., 2023), which involves parallel training of a task-specific classifier and a bias-exploiting adversary model, and using stacked denoising autoencoders with weighted reconstruction loss to enhance representation of underrepresented classes (Sivarajkumar et al., 2023). However, these approaches fail to account for the complex interactions between social biases that are embedded in demographic features and the multimodal nature of EHR data (Wang et al., 2022a). In contrast, our proposed method leverages multimodal EHRs and addresses a spectrum of social biases through a unified framework.

Contrastive Learning. Contrastive learning (CL), originally developed for vision tasks, employs the principle of contrasting samples to identify attributes common to and differentiating between data classes (Khosla et al., 2020; Chen et al., 2020; Jaiswal et al., 2020). In essence, CL generates varied views of original data through random augmentation, treating views from the same source as positive pairs. The model then learns effective representations by minimizing the distance between these positive pair representations. Recently, CL has been adapted for patient representation in EHRs, applied in critical event prediction for COVID-19 (Wanyan et al., 2021), clinical risk prediction (Zang and Wang, 2021), and survival analysis (Nayebi Kerdabadi et al., 2023). However, existing CL applications in EHRs neglect potential fairness issues. To address this oversight, our method introduces a fairness-oriented contrastive loss for training models that learn fair representations, incorporating tailored contrasting sample designs specific to EHRs.

Fairness Evaluation. Traditional fairness metrics such as equalized odds, equal opportunity (Hardt et al., 2016), demographic parity (Jiang et al., 2022), and disparate impact assess fairness are based on aggregate outcomes across diverse demographic groups (Feldman et al., 2015). However, these metrics may not fully capture the heterogeneity and distinct distribution patterns in EHR data, particularly when considering variability in subgroup sizes. To address this gap, we propose the Error Distribution Disparity Index (EDDI), a metric specifically designed for EHRs. EDDI measures fairness by evaluating the disparities in error rates across subgroups relative to the overall error rate, which is crucial in clinical settings characterized by imbalanced outcome labels and varying patient group sizes.

3. Methods

In this section, we begin by presenting an overview of the problem formulation and the workflow of our FairEHR-CLP. Then, the process of generating synthetic counterparts for each patient during the training phase is detailed. Finally, we discuss fairness-aware predictions with CL, in conjunction with the outcome prediction with the MLP classifier.

3.1. Problem Formulation and Method Overview

We define a dataset as $\mathscr{D} = \{(x_k, y_k, s_k)\}_{k=1}^{n}$, where $x_k \in \mathcal{X}$ corresponds to the input features extracted from patient demographics, longitudinal health records, and clinical notes; $y_k \in \{0, 1\} \subseteq \mathcal{Y}$ denotes the binary target label; and $s_k \in \mathcal{S}$ signifies the sensitive attribute indicative of potential social bias. These attributes encompass gender (male, female), race (White, Black, Asian, etc.), ethnicity (including categories such as Latino/Hispanic), age (categorized into ranges like 50-60, 60-70, etc.), and socioeconomic status (SES), represented by the type of insurance (private, government, etc.). The inclusion of insurance type as a proxy for SES allows for the examination of disparities that may arise due to economic barriers to healthcare access (Green et al., 2021). Our objective is to develop an effective and fair prediction model $f : \mathcal{X} \to \mathcal{Y}$ that aims to accurately predict outcomes without discriminating against the subgroups defined by sensitive attributes $\mathcal{S}$ from demographics.

Our approach unfolds in two primary stages: 1) **Synthetic Counterpart Generation**, where we generate synthetic demographic counterparts to represent a spectrum of demographic identities. For creating corresponding synthetic longitudinal data, we employ EHR-based Generative Adversarial Networks (GANs) (Li et al., 2023). Simultaneously, Llama2-70b (Touvron et al., 2023) is used to synthesize clinical notes, thereby enriching our dataset to mirror demographic diversity while maintaining clinical accuracy. An example of a patient profile is illustrated in Appendix A. 2) **Fairness-Aware Predictions with CL**, in which we align the representations of real and synthetic data to address biases. It incorporates an MLP classifier with a softmax layer for downstream classification tasks, leveraging aligned real data representations for final prediction.

3.2. Synthetic Counterpart Generation

Our initial step involves generating synthetic counterparts for sensitive attributes (i.e., gender, age, race, ethnicity, and insurance), along with longitudinal data (including vital signs and lab measurements), and clinical notes for patients during the training phase. This process

creates pairs of patients with similar health conditions but distinct demographic factors. For example, if the original patient is a 55-year-old diabetic White male, the corresponding synthetic counterpart might be a 60-year-old diabetic Black female. This step enhances the representation of diverse demographics while preserving the consistency of health-related information. These synthetic samples are used alongside real data for contrastive training but not in final predictions in full FairEHR-CLP experiments.

Sensitive Attributes. We consider five sensitive attributes, which range from binary to multi-class subgroups for each attribute. Four of these attributes are categorical, except for age, which is a continuous variable. For the categorical variables, we randomly assign a new category to each patient to create their corresponding synthetic counterpart (e.g., male to female). For age, we segment our patient cohort into 10-year age bins (e.g., 50-60, 60-70, etc.). Then, we assign a random age within a different bin for each patient's synthetic age.

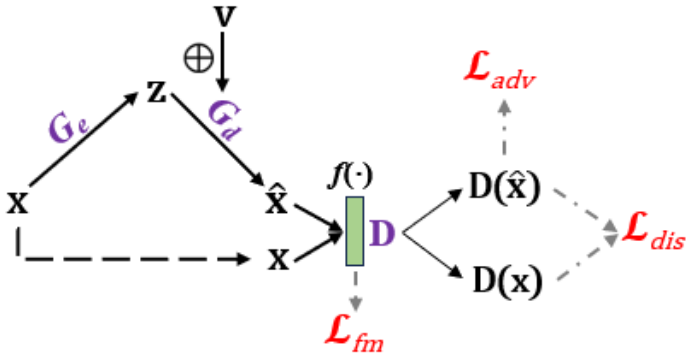


Figure 2: Architecture of EHR-GAN. The green box indicates the output from an intermediate layer of the discriminator D.

Longitudinal Data. We generate synthetic longitudinal data, which is data collected from the same individuals over a period of time, using the EHR-M-GAN model (Li et al., 2023), focusing exclusively on continuous data streams, which we designate as EHR-GAN, as shown in Figure 2. The architecture of EHR-GAN comprises a generator G, which includes an encoder G_e, a decoder G_d, and a discriminator D. The encoder G_e transforms the input x into a latent space representation z. Subsequently, the decoder G_d utilizes z, along with random noise v, to generate synthetic data $\hat{x}$. The discriminator D is responsible for distinguishing between real and synthetic data. The training process involves optimizing three joint losses: 1) The discriminative loss l_{dis}, provided by the discriminator D, ensures that the generated longitudinal data appear realistic. It is defined as:

$$l_{dis} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log D(x_i) + (1-y_i)\log(1 - D(G_d(z_i)))],$$

where y_i denotes the label indicating whether the data is real or synthetic. 2) The adversarial loss l_{adv} encourages the decoder G_d to produce data that the discriminator will classify as

real:

$$l_{adv} = -\mathbb{E}_{z\sim p_z(z)}\left[\log D(G_d(z))\right],$$

where $p_z(z)$ represents the prior distribution over the latent space representation z. 3) The feature matching loss l_{fm} ensures that the decoder G_d creates data with statistical properties that are similar to real data:

$$l_{fm} = \sqrt{\mathbb{E}_{x\sim p_x(x), z\sim p_z(z)}\left[\left(f(D(x)) - f(D(G_d(z)))\right)^2\right]},$$

thereby minimizing the discrepancy between the discriminative features of the real and synthetic data. Here, $f(\cdot)$ denotes the output of an intermediate layer of the discriminator D, and $p_x(x)$ is the distribution of the real data.

The total loss is $\beta_0 l_{dis} + \beta_1 l_{adv} + \beta_2 l_{fm}$, where β_0, β_1, and β_2 are the weighting coefficients that balance the importance of each loss component.

Clinical Notes. We utilize Llama2-70b-chat (Touvron et al., 2023) to generate synthetic clinical notes. The model receives specific instructions to ensure the preservation of essential elements in clinical documentation: "*Please paraphrase the provided clinical notes, ensuring no critical medical components such as medical history, diagnoses, and treatments are omitted while maintaining the integrity of authentic documentation*". After generation, a random subset of these synthetic notes undergoes manual review by clinical experts. This process is crucial to confirm the fidelity and accuracy of the content, ensuring it aligns with authentic clinical records in accordance with the given prompt. Details regarding the review guidelines are in Appendix B.

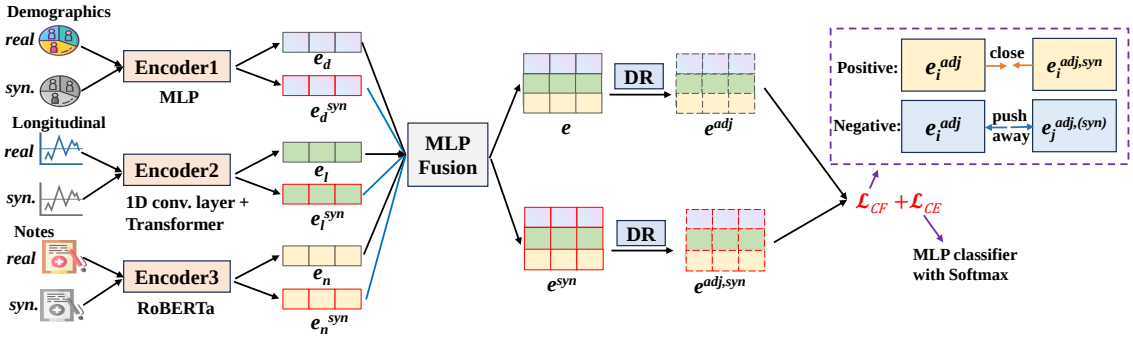


Figure 3: Workflow of our second stage: fairness-aware predictions with contrastive learning. The term $e_j^{adj,(syn)}$ represents both e_j^{adj} and $e_j^{adj,syn}$, with $j \neq i$. Also, conv. layer denotes the convolutional layer.

3.3. Fairness-Aware Predictions with Contrastive Learning

In our augmented dataset, which includes both real patient data and synthetic counterparts spanning demographics, longitudinal records, and clinical notes during the training phase, we implement fairness-aware predictions with contrastive learning. For each patient, positive

samples (x^+) are defined as their respective synthetic counterparts. These counterparts differ in sensitive demographic attributes but are matched to share similar health conditions, as determined by corresponding synthetic longitudinal and note data. In contrast, negative samples (x^-) are all other patients present in the same minibatch during training. To encode features from both real and synthetic data during training, demographic characteristics are processed using an MLP encoder, while longitudinal data are handled with a convolutional layer followed by a standard Transformer encoder to capture temporal dynamics. Clinical note embeddings are derived using RoBERTa-large (Liu et al., 2019). The encoded demographic, longitudinal, and note data are denoted as e_d, e_l, and e_n, respectively. Following this, an MLP-based fusion combines these modality-specific representations into a unified representation that captures inter-modal dependencies and interactions: $F_{\text{fusion}}(\cdot) = \text{MLP}(e_{\text{d}} \oplus e_{\text{l}} \oplus e_{\text{n}}; \theta_{\text{fusion}})$, where θ_{fusion} represents the set of trainable parameters within the fusion layer. Integrated representations for real and synthetic data are labeled as e and e^{syn}, respectively. To dynamically address potential biases across different data types, we introduce a Dynamic Relevance (DR) layer, defined as $F_{\text{DR}}(e) = \sigma(w) \odot e$, using e as an example, where w represents adjustable weights and σ is the sigmoid function. This gating mechanism modulates the influence of each feature in the final representation. Post-DR, the adjusted embeddings are referred to as e^{adj} and $e^{adj,syn}$ for real and synthetic data, respectively. The joint learning objective combines a fairness-oriented contrastive loss (l_{CF}) for bias mitigation and cross entropy loss (l_{CE}) to enhance classification performance. Formally,

$$l_{CF} = \sum_{k=1}^{N} -\log \frac{\exp(\text{sim}(e_k^{adj}, e_{k^+}^{adj,syn})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(e_k^{adj}, e_{j^-}^{adj,syn})/\tau)} + \gamma \left(\frac{1}{N} \sum_{k=1}^{N} \left\| e_k^{adj,syn} - \mu_{\text{syn}}^{adj} \right\|_2^2 \right),$$

where N denotes the number of real embeddings (and corresponding synthetic counterparts) in a minibatch, $\text{sim}(u, v)$ calculates cosine similarity, τ is a temperature parameter, γ is a regularization parameter, and μ_{syn}^{adj} is the mean of $e^{adj,syn}$ across a minibatch. The first term is inspired by NT-Xent loss (Chen et al., 2020), while the second term encourages the synthetic embeddings to cluster tightly around their mean, mitigating overfitting to outliers in synthetic data. Additionally,

$$l_{CE}(e, y) = -\sum_{k=1}^{N} y_k \log(C(e_k^{adj})),$$

where y_k corresponds to the true label for each of the N real embeddings and $C(e_k^{adj})$ signifies the softmax probability of the predicted class. The total loss is $\sum_k (\alpha l_{CF} + (1-\alpha) l_{CE})$, with α balancing fairness and performance. The detailed workflow of our second stage (fairness-aware predictions with CL) is depicted in Figure 3. For clarity, all notations used throughout this section can be found in Appendix C. Implementation details are in Appendix D.

4. Experimental Setup

In this section, we outline the experimental setup, including the datasets used, the baseline models for comparison, and the evaluation metrics employed.

4.1. Datasets

We evaluate our proposed framework using three EHR datasets: STAnford medicine Research data Repository (STARR) from Stanford Medicine, MIMIC-III, and MIMIC-IV. Our focus is on surgical patients aged 50 years or older, a cohort often subject to social bias in medical treatments and outcomes due to age-related factors like impaired cognition. For the MIMIC-III and MIMIC-IV datasets, we specifically employ the MIMIC-III Clinical Database CareVue subset (Johnson et al., 2022) to ensure there is no overlap of patient data. The study targets three critical tasks: classifying delirium, OUD, and 30-day readmission. These tasks are chosen for their direct impact on enhancing postoperative care, improving patient safety, and reducing healthcare costs. Demographic indicators are excluded from clinical notes to focus solely on health conditions. We extract patient data from a 24-hour postoperative period and employ MICE imputation (Van Buuren and Groothuis-Oudshoorn, 2011) to address missing values for all datasets. Each task is approached as a binary classification problem. The class distribution for each task is summarized in Table 1 with more details in Appendix E.

Table 1: Class distribution in three prediction tasks over all datasets.

Dataset	Delirium *class 0 / 1*	OUD *class 0 / 1*	30-day Readmission *class 0 / 1*
STARR	39,516 / 7,417	42,156 / 4,777	34,919 / 12,014
MIMIC-III	4,030 / 272	3,998 / 304	3,974 / 328
MIMIC-IV	7,956 / 7,962	14,169 / 1,749	9,136 / 6,782

4.2. Baselines

To assess our method in terms of performance and fairness, we compare it with a variety of established methods. Our evaluation begins with the Demographic-free Classification (DfC) approach, based on the premise that models, if unaware of demographic features often central to socially sensitive biases, should demonstrate minimal differences in performance. Additionally, we explore two notable debiasing strategies tailored for EHR: Adversarial Debiasing (AdvDebias)(Zhang et al., 2018; Yang et al., 2023), a technique that simultaneously trains a classifier and an adversary model to neutralize bias, and Fair Patient Model (FPM)(Sivarajkumar et al., 2023), which employs a Stacked Denoising Autoencoder and a weighted reconstruction loss for equitable patient representations. Furthermore, we include comparisons with embedding methods RoBERTa-large (Liu et al., 2019) and ClinicalBERT (Alsentzer et al., 2019), widely used in general and healthcare-specific applications, respectively. The embeddings generated by these models are utilized as inputs for an MLP classifier equipped with a softmax layer for prediction.

4.3. Evaluation Metrics

For classification performance evaluation, we employ F1 and AUROC as metrics. Regarding fairness metrics, we adopt a variant of the Equalized Odds (EO) metric (Hardt et al., 2016), a widely recognized notion of group fairness (Dwork et al., 2012). Traditionally, EO suggests

that a model achieves fairness when the True Positive Rates (TPR) and False Positive Rates (FPR) are consistent across all subgroups defined by the sensitive attribute. However, this conventional interpretation of EO may not fully account for practical challenges such as data variability or differences in group sizes in clinical settings. Therefore, we employ the Average Disparity in EO to measure the average deviation from the ideal EO condition:

$$\mathrm{EO}_{\mathrm{TPR}} = \frac{1}{\binom{|S|}{2}} \sum_{s_i} \sum_{s_j > s_i} \left| \mathrm{TPR}_{s_i} - \mathrm{TPR}_{s_j} \right|,$$

$$\mathrm{EO}_{\mathrm{FPR}} = \frac{1}{\binom{|S|}{2}} \sum_{s_i} \sum_{s_j > s_i} \left| \mathrm{FPR}_{s_i} - \mathrm{FPR}_{s_j} \right|,$$

where

$$\mathrm{TPR}_s = \frac{\mathrm{TP}_s}{\mathrm{TP}_s + \mathrm{FN}_s}$$

and

$$\mathrm{FPR}_s = \frac{\mathrm{FP}_s}{\mathrm{FP}_s + \mathrm{TN}_s}.$$

Here, for each subgroup $s \in S$, where S is the set of subgroups determined by a sensitive attribute (e.g., race), TP_s, FN_s, FP_s, and TN_s represent the counts of true positives, false negatives, false positives, and true negatives for each subgroup s, respectively. We adopt the pairwise comparison approach, averaging the differences in TPR and FPR across all pairs of subgroups (e.g., White, Black, etc.) within a sensitive attribute (e.g., race). We then compute the arithmetic mean of $\mathrm{EO}_{\mathrm{TPR}}$ and $\mathrm{EO}_{\mathrm{FPR}}$ to establish a singular EO metric.

A critical limitation of the traditional EO metric is its tendency to oversimplify fairness across subgroups that are diverse and unevenly represented, failing to adequately capture subgroup-specific error rate disparities. To overcome this, we introduce the Error Distribution Disparity Index (EDDI), a new fairness metric designed to address the complexities of clinical settings, especially those with significant data variability and diverse group sizes. It is formulated as:

$$\mathrm{EDDI} = \frac{1}{|S|} \sum_{s \in S} \frac{\mathrm{ER}_s - \mathrm{OER}}{\max(\mathrm{OER}, 1 - \mathrm{OER})},$$

where

$$\mathrm{ER}_s = \frac{1}{N_s} \sum_{i \in s} \mathbb{I}(y_i \neq \hat{y}_i)$$

represents the error rate for each subgroup s and

$$\mathrm{OER} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i \neq \hat{y}_i)$$

denotes the overall error rate across the dataset. Here, y_i and $\hat{y}_i$ denote the true and predicted labels, respectively. N_s and N indicate the number of instances within each subgroup and the total number of instances in the dataset, respectively. EDDI quantifies the error rate deviation for each subgroup from the overall error rate. We contend that a model is fair if it maintains consistent error rates across all demographic subgroups. In general, reduced values of EO and EDDI signify enhanced fairness in the model.

Table 2: Performance and fairness evaluation across three Datasets: STARR, MIMIC-III, and MIMIC-IV. We report average results and standard deviations over five runs. EO and EDDI results are averaged over five sensitive attributes. For each dataset and each task, results highlighted in bold indicate the highest performance, while those underlined denote the optimal fairness outcomes. Our method demonstrates superior classification performance and fairness in the majority of settings.

Model	Delirium				OUD				30-Day Readmission			
	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)
Dataset 1: STARR												
DfC	$79.6_{\pm 1.4}$	$81.8_{\pm 1.2}$	$5.2_{\pm 0.8}$	$2.6_{\pm 0.5}$	$85.7_{\pm 1.8}$	$89.2_{\pm 1.5}$	$3.4_{\pm 0.6}$	$2.8_{\pm 0.7}$	$80.9_{\pm 1.6}$	$83.4_{\pm 1.5}$	$0.2_{\pm 0.6}$	$3.7_{\pm 0.5}$
AdvDebias	$81.5_{\pm 1.7}$	$83.8_{\pm 1.4}$	$6.6_{\pm 0.9}$	$4.2_{\pm 0.5}$	$83.6_{\pm 2.0}$	$87.3_{\pm 1.6}$	$3.8_{\pm 0.8}$	$2.9_{\pm 0.6}$	$81.2_{\pm 1.8}$	$84.2_{\pm 1.4}$	$0.8_{\pm 0.4}$	$4.8_{\pm 0.6}$
FPM	$80.2_{\pm 1.7}$	$82.6_{\pm 1.4}$	$7.0_{\pm 0.9}$	$4.4_{\pm 0.6}$	$84.3_{\pm 2.1}$	$88.1_{\pm 1.8}$	$3.8_{\pm 0.9}$	$3.0_{\pm 0.8}$	$80.6_{\pm 1.2}$	$83.1_{\pm 1.0}$	$0.9_{\pm 0.3}$	$4.7_{\pm 0.6}$
RoBERTa	$83.6_{\pm 1.5}$	$86.2_{\pm 1.3}$	$8.7_{\pm 1.0}$	$5.2_{\pm 0.8}$	$\mathbf{87.5_{\pm 1.8}}$	$\mathbf{91.3_{\pm 1.5}}$	$4.8_{\pm 0.7}$	$4.0_{\pm 0.8}$	$82.3_{\pm 1.7}$	$85.9_{\pm 1.6}$	$1.4_{\pm 0.3}$	$5.9_{\pm 0.9}$
ClinicalBERT	$82.8_{\pm 1.6}$	$84.1_{\pm 1.4}$	$8.0_{\pm 1.1}$	$4.6_{\pm 0.7}$	$85.2_{\pm 1.4}$	$88.9_{\pm 1.2}$	$4.2_{\pm 0.8}$	$3.5_{\pm 0.7}$	$81.6_{\pm 1.9}$	$84.7_{\pm 1.3}$	$1.1_{\pm 0.4}$	$5.6_{\pm 0.8}$
FairEHR-CLP (Ours)	$\mathbf{84.1_{\pm 1.3}}$	$\mathbf{87.3_{\pm 1.0}}$	$5.7_{\pm 0.7}$	$3.4_{\pm 0.5}$	$86.3_{\pm 1.6}$	$90.6_{\pm 1.4}$	$3.5_{\pm 0.6}$	$2.8_{\pm 0.5}$	$\mathbf{83.2_{\pm 1.3}}$	$\mathbf{87.8_{\pm 1.5}}$	$0.4_{\pm 0.2}$	$4.4_{\pm 0.6}$
Dataset 2: MIMIC-III												
DfC	$82.9_{\pm 1.4}$	$85.8_{\pm 1.3}$	$5.8_{\pm 0.7}$	$3.6_{\pm 0.6}$	$86.8_{\pm 1.5}$	$88.3_{\pm 1.6}$	$3.3_{\pm 0.5}$	$1.8_{\pm 0.5}$	$83.7_{\pm 1.2}$	$86.6_{\pm 1.0}$	$2.1_{\pm 0.3}$	$1.3_{\pm 0.4}$
AdvDebias	$74.5_{\pm 1.6}$	$77.6_{\pm 1.7}$	$7.0_{\pm 0.9}$	$4.9_{\pm 0.8}$	$85.2_{\pm 1.5}$	$87.1_{\pm 1.3}$	$3.9_{\pm 0.7}$	$2.4_{\pm 0.5}$	$85.4_{\pm 1.3}$	$88.1_{\pm 1.0}$	$5.1_{\pm 0.3}$	$3.8_{\pm 0.2}$
FPM	$75.8_{\pm 1.6}$	$79.2_{\pm 1.8}$	$6.6_{\pm 0.8}$	$4.3_{\pm 0.6}$	$83.7_{\pm 1.3}$	$86.4_{\pm 1.5}$	$4.5_{\pm 0.3}$	$2.6_{\pm 0.4}$	$84.3_{\pm 1.4}$	$87.2_{\pm 1.2}$	$5.4_{\pm 0.5}$	$4.0_{\pm 0.4}$
RoBERTa	$83.7_{\pm 1.4}$	$86.9_{\pm 1.6}$	$7.2_{\pm 0.7}$	$4.0_{\pm 0.6}$	$87.2_{\pm 1.3}$	$89.7_{\pm 1.5}$	$4.6_{\pm 0.6}$	$2.9_{\pm 0.6}$	$86.1_{\pm 1.4}$	$89.5_{\pm 1.2}$	$5.6_{\pm 0.2}$	$4.5_{\pm 0.3}$
ClinicalBERT	$85.1_{\pm 1.5}$	$87.6_{\pm 1.7}$	$6.7_{\pm 0.8}$	$4.2_{\pm 0.7}$	$86.9_{\pm 1.4}$	$88.5_{\pm 1.3}$	$4.2_{\pm 0.5}$	$3.5_{\pm 0.7}$	$85.3_{\pm 1.4}$	$87.9_{\pm 1.6}$	$5.3_{\pm 0.6}$	$4.8_{\pm 0.8}$
FairEHR-CLP (Ours)	$\mathbf{85.5_{\pm 1.2}}$	$\mathbf{89.7_{\pm 1.1}}$	$6.2_{\pm 0.3}$	$3.8_{\pm 0.5}$	$\mathbf{89.4_{\pm 1.4}}$	$\mathbf{91.9_{\pm 1.5}}$	$3.7_{\pm 0.5}$	$2.0_{\pm 0.4}$	$\mathbf{88.2_{\pm 1.3}}$	$\mathbf{91.4_{\pm 1.1}}$	$3.3_{\pm 0.4}$	$2.1_{\pm 0.6}$
Dataset 3: MIMIC-IV												
DfC	$76.1_{\pm 1.6}$	$79.4_{\pm 1.3}$	$4.9_{\pm 0.6}$	$3.5_{\pm 0.4}$	$75.2_{\pm 1.9}$	$79.5_{\pm 1.8}$	$1.3_{\pm 0.6}$	$2.1_{\pm 0.5}$	$76.9_{\pm 1.5}$	$79.3_{\pm 1.4}$	$2.2_{\pm 0.6}$	$4.5_{\pm 0.7}$
AdvDebias	$73.6_{\pm 1.8}$	$76.6_{\pm 1.6}$	$5.3_{\pm 0.7}$	$4.0_{\pm 0.8}$	$74.7_{\pm 1.5}$	$78.6_{\pm 1.3}$	$5.8_{\pm 0.5}$	$3.0_{\pm 0.6}$	$77.8_{\pm 1.3}$	$80.6_{\pm 1.2}$	$3.1_{\pm 0.3}$	$5.9_{\pm 0.3}$
FPM	$70.4_{\pm 2.0}$	$73.1_{\pm 1.8}$	$5.6_{\pm 0.8}$	$4.2_{\pm 0.9}$	$72.9_{\pm 1.5}$	$76.0_{\pm 1.3}$	$5.0_{\pm 0.8}$	$2.6_{\pm 0.7}$	$79.2_{\pm 1.4}$	$82.7_{\pm 1.5}$	$3.0_{\pm 0.5}$	$5.6_{\pm 0.7}$
RoBERTa	$77.9_{\pm 1.4}$	$81.1_{\pm 1.6}$	$5.7_{\pm 0.5}$	$4.3_{\pm 0.7}$	$\mathbf{86.3_{\pm 1.9}}$	$\mathbf{89.6_{\pm 1.7}}$	$4.2_{\pm 0.8}$	$2.3_{\pm 0.9}$	$81.3_{\pm 1.4}$	$85.7_{\pm 1.5}$	$3.6_{\pm 0.6}$	$5.6_{\pm 0.5}$
ClinicalBERT	$78.2_{\pm 1.7}$	$81.7_{\pm 1.5}$	$6.0_{\pm 0.6}$	$4.6_{\pm 0.8}$	$84.2_{\pm 2.1}$	$87.6_{\pm 1.8}$	$4.9_{\pm 0.9}$	$3.1_{\pm 0.9}$	$80.4_{\pm 1.2}$	$83.7_{\pm 1.1}$	$3.9_{\pm 0.5}$	$5.7_{\pm 0.6}$
FairEHR-CLP (Ours)	$\mathbf{78.8_{\pm 1.2}}$	$\mathbf{82.4_{\pm 1.0}}$	$6.1_{\pm 0.4}$	$3.5_{\pm 0.3}$	$84.8_{\pm 1.6}$	$88.9_{\pm 1.5}$	$1.5_{\pm 0.3}$	$3.0_{\pm 0.6}$	$\mathbf{81.6_{\pm 1.8}}$	$\mathbf{86.4_{\pm 1.6}}$	$2.8_{\pm 0.7}$	$5.2_{\pm 0.9}$

5. Results

In this section, we present a comprehensive comparison of our method with baselines across all datasets in Section 5.1, explore the effects of data modalities, model components, and hyperparameters in Section 5.2, provide visualizations of learned representations in Section 5.3, and analyze the model's impact on each sensitive attribute in Section 5.4.

5.1. Main Results

We report the classification and fairness results from the test set in the second stage of our approach (see Figure 1) across three tasks and three datasets (9 settings in total) in Table 2. We use F1 and AUROC as performance metrics, as well as EO and EDDI as fairness metrics, with EO and EDDI results averaged over five sensitive attributes. There are several key takeaways. Firstly, FairEHR-CLP consistently outperforms DfC in F1 and AUROC by 4.8% and 5.8% on average, respectively, highlighting the benefit of demographic features in enhancing predictive accuracy, despite potential bias risks. In terms of fairness, FairEHR-CLP achieves EO and EDDI levels comparable to DfC, affirming the effectiveness of our bias mitigation approach. Moreover, when compared with specialized debiasing methods like AdvDebias and FPM, FairEHR-CLP excels in both predictive accuracy and fairness in most settings. This superior performance can be attributed to its comprehensive integration of multimodal EHR data and concurrent bias mitigation across multiple sensitive attributes, in contrast to the single-attribute focus of AdvDebias and FPM. Lastly, against classification methods using embeddings such as RoBERTa and ClinicalBERT, FairEHR-CLP shows

superior performance in 7 out of 9 tasks, along with consistently lower EO and EDDI scores across all settings, demonstrating its robustness in balancing bias management with minimal performance loss.

5.2. Ablation Study

We conduct ablation studies on the STARR dataset to evaluate: (1) the effectiveness of various data modalities; (2) the impact of the main components of FairEHR-CLP; and (3) the influence of the key hyperparameter α, which balances fairness and performance. For additional results on other datasets, please refer to Appendix F.

Data Modalities. We study the effectiveness of different data modalities (demographics $\mathcal{D}$, longitudinal $\mathcal{L}$, and notes $\mathcal{N}$) within the full FairEHR-CLP framework. Considering our objective of mitigating social bias, often rooted in $\mathcal{D}$, we keep it constant in our ablation experiments. We then explore all combinations involving $\mathcal{D}$ and present the results on the STARR dataset in Table 3. We observe that the $\mathcal{D}+\mathcal{L}$ combination marginally outperforms the $\mathcal{D}+\mathcal{N}$ combination. Utilizing the full dataset ($\mathcal{D}+\mathcal{L}+\mathcal{N}$) results in a 2.2% improvement in F1 and a 2.5% increase in AUROC compared to the second-best results ($\mathcal{D}+\mathcal{L}$). From a fairness perspective, the complete data combination consistently demonstrates a reduction in bias, indicating a more nuanced understanding and representation of patient profiles, leading to more equitable outcome predictions.

Table 3: Effects of different data modalities as inputs for FairEHR-CLP on the STARR dataset. Here, $\mathcal{D}$, $\mathcal{L}$, and $\mathcal{N}$ represent demographics, longitudinal data, and clinical notes, respectively.

Data Modalities	Delirium				OUD				30-Day Readmission			
	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)
$\mathcal{D}$	$78.5_{\pm 1.7}$	$80.2_{\pm 1.6}$	$7.8_{\pm 0.9}$	$5.5_{\pm 0.6}$	$81.2_{\pm 1.5}$	$85.0_{\pm 1.3}$	$4.7_{\pm 0.8}$	$4.1_{\pm 0.7}$	$79.3_{\pm 1.4}$	$82.1_{\pm 1.2}$	$6.0_{\pm 0.5}$	$6.2_{\pm 0.6}$
$\mathcal{D}+\mathcal{L}$	$82.3_{\pm 1.4}$	$85.5_{\pm 1.2}$	$6.2_{\pm 0.8}$	$4.2_{\pm 0.5}$	$84.1_{\pm 1.6}$	$88.3_{\pm 1.4}$	$3.9_{\pm 0.7}$	$3.0_{\pm 0.6}$	$81.8_{\pm 1.3}$	$85.4_{\pm 1.1}$	$1.1_{\pm 0.3}$	$4.9_{\pm 0.5}$
$\mathcal{D}+\mathcal{N}$	$81.7_{\pm 1.5}$	$84.8_{\pm 1.3}$	$6.7_{\pm 0.7}$	$4.8_{\pm 0.6}$	$83.7_{\pm 1.7}$	$87.6_{\pm 1.5}$	$4.1_{\pm 0.6}$	$3.5_{\pm 0.7}$	$81.5_{\pm 1.2}$	$85.2_{\pm 1.0}$	$1.6_{\pm 0.4}$	$5.0_{\pm 0.7}$
$\mathcal{D}+\mathcal{L}+\mathcal{N}$	**$84.1_{\pm 1.3}$**	**$87.3_{\pm 1.0}$**	$5.7_{\pm 0.7}$	$3.4_{\pm 0.5}$	**$86.3_{\pm 1.6}$**	**$90.6_{\pm 1.4}$**	$3.5_{\pm 0.6}$	$2.8_{\pm 0.5}$	**$83.2_{\pm 1.3}$**	**$87.8_{\pm 1.5}$**	$0.4_{\pm 0.2}$	$4.4_{\pm 0.6}$

Model Components. We investigate the key model components in the FairEHR-CLP, namely the CL approach and the DR layer. We maintain synthetic counterparts for data augmentation during the training phase when CL is not applied. Results from the STARR dataset, as shown in Table 4, reveal that removing both CL and DR results in the most significant performance degradation, averaging a 2.6% drop in F1 and 4.1% in AUROC across three tasks. This setup also yields the most biased predictions. The absence of either CL or DR (full w/o CL or full w/o DR) leads to only a slight decline in performance but shows a tendency towards more biased outcomes compared to those from the full model. This can be attributed to the complementary roles of CL and DR in balancing accurate predictions with fairness.

Effect of α. We investigate the effect of α on the trade-off between fairness and utility across a range from 0.0 to 1.0. Figure 5 demonstrates that, generally, a lower α prioritizes utility, resulting in higher F1 scores at the expense of fairness, as reflected by increased EO and EDDI values. Conversely, a higher α enhances fairness, evidenced by lower EO and EDDI, but leads to decreased F1 scores. The figure indicates that the optimal $\alpha = 0.6$,

Table 4: Effects of different model components for FairEHR-CLP (full) on the STARR dataset ($\mathcal{D} + \mathcal{L} + \mathcal{N}$). Here, 'w/o CL' and 'w/o DR' represent the full model without contrastive learning and without the Dynamic Relevance layer, respectively.

Model Components	Delirium				OUD				30-Day Readmission			
	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)
Full w/o CL + DR	$80.3_{\pm 1.4}$	$82.1_{\pm 1.1}$	$7.2_{\pm 0.8}$	$5.3_{\pm 0.5}$	$85.9_{\pm 1.7}$	$89.7_{\pm 1.3}$	$5.1_{\pm 0.7}$	$4.3_{\pm 0.6}$	$81.2_{\pm 1.2}$	$83.5_{\pm 1.3}$	$1.9_{\pm 0.3}$	$5.6_{\pm 0.5}$
Full w/o CL	$81.1_{\pm 1.3}$	$83.0_{\pm 1.0}$	$6.7_{\pm 0.7}$	$4.5_{\pm 0.4}$	$86.0_{\pm 1.5}$	$89.9_{\pm 1.2}$	$4.7_{\pm 0.6}$	$3.8_{\pm 0.5}$	$81.7_{\pm 1.1}$	$84.6_{\pm 1.2}$	$1.6_{\pm 0.2}$	$5.1_{\pm 0.4}$
Full w/o DR	$82.5_{\pm 1.2}$	$85.4_{\pm 0.9}$	$6.4_{\pm 0.6}$	$4.1_{\pm 0.3}$	$86.2_{\pm 1.3}$	$90.3_{\pm 1.1}$	$3.9_{\pm 0.5}$	$3.2_{\pm 0.4}$	$82.1_{\pm 1.0}$	$86.2_{\pm 1.1}$	$1.2_{\pm 0.2}$	$4.8_{\pm 0.3}$
Full	$\mathbf{84.1}_{\pm \mathbf{1.3}}$	$\mathbf{87.3}_{\pm \mathbf{1.0}}$	$5.7_{\pm 0.7}$	$3.4_{\pm 0.5}$	$\mathbf{86.3}_{\pm \mathbf{1.6}}$	$\mathbf{90.6}_{\pm \mathbf{1.4}}$	$3.5_{\pm 0.6}$	$2.8_{\pm 0.5}$	$\mathbf{83.2}_{\pm \mathbf{1.3}}$	$\mathbf{87.8}_{\pm \mathbf{1.5}}$	$0.4_{\pm 0.2}$	$4.4_{\pm 0.6}$

positioned at the top-left corner of both plots, signifies an equitable compromise between fairness and utility.

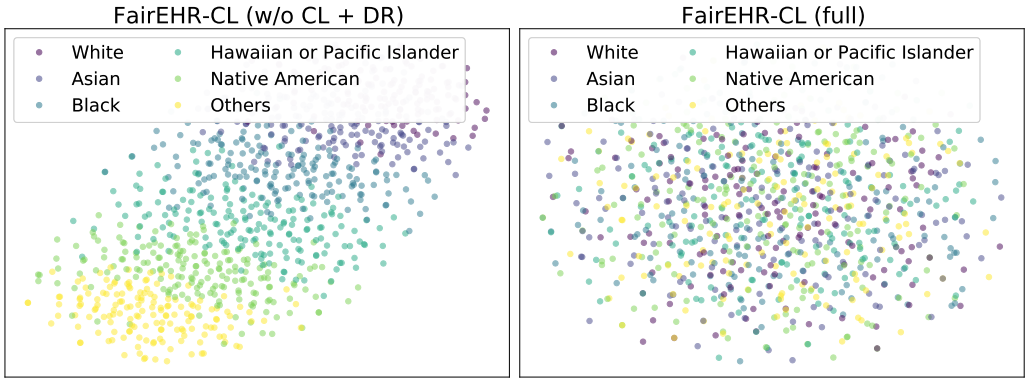


Figure 4: t-SNE visualization of learned representations from FairEHR-CLP with and without bias mitigation components CL and DR on the STARR dataset w.r.t. the sensitive group race.

5.3. Visualization

To assess the quality of the learned representations and the effectiveness of our method, we employ t-SNE (Van der Maaten and Hinton, 2008) to visualize projections of 1000 patient records from the STARR test set, focusing on the sensitive attribute race, as shown in Figure 4. The left panel depicts a vanilla model lacking the CL and DR components, which are integral to bias mitigation in FairEHR-CLP. We observe that the vanilla model learns information about race, as the representations given by vanilla exhibit distinct clusters along racial lines. It suggests that the model may be disproportionately weighting race when forming representations. In contrast, our full FairEHR-CLP model on the right shows a more homogeneous distribution across racial groups, suggesting a reduced impact of race on the representations, thereby diminishing reliance on biased attributes and advancing towards more equitable predictions.

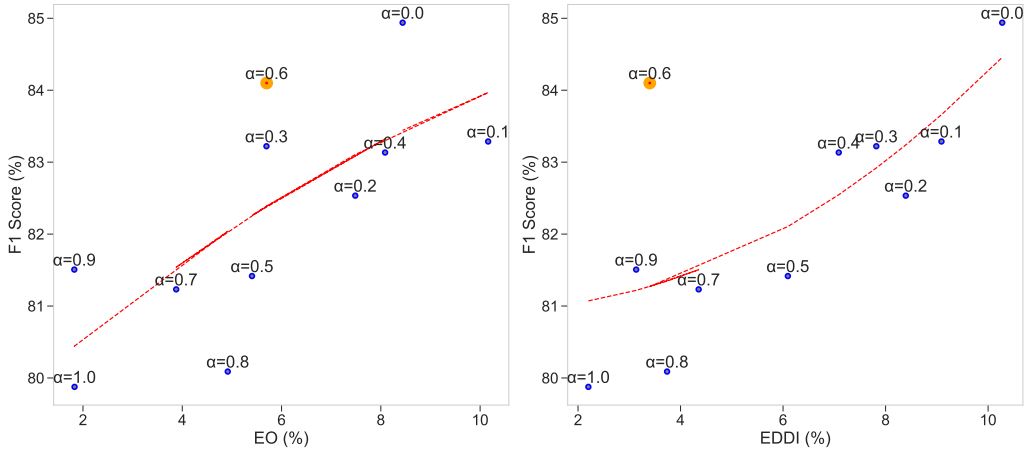


Figure 5: Effect of α on fairness-utility trade-off in the STARR dataset (delirium task). Left: EO vs. F1; Right: EDDI vs. F1.

5.4. Sensitive Attributes Analysis

We investigate the impact of our method on each sensitive attribute from a fairness perspective. Table 5 presents the EO and EDDI values for each sensitive attribute across three datasets. Our approach consistently demonstrates the least bias in gender, with EO as low as 1.7% and EDDI at 1.8%, followed by a slightly increasing bias in SES. The most biased sensitive attribute is race, exhibiting up to 5.9% in EO and 4.7% in EDDI. Similarly, age bias is also pronouncedly high. The variability in bias levels across different sensitive attributes and datasets underscores the impact of dataset-specific characteristics on model fairness.

Table 5: Fairness evaluation of FairEHR-CLP across individual sensitive attributes in three datasets, averaged over three tasks. Bold values represent the least bias, while underlined values indicate the most bias among sensitive attributes.

Attributes	**STARR**		**MIMIC-III**		**MIMIC-IV**	
	EO (↓)	EDDI (↓)	EO (↓)	EDDI (↓)	EO (↓)	EDDI (↓)
Gender	$\mathbf{1.7_{\pm 0.5}}$	$\mathbf{2.4_{\pm 0.5}}$	$\mathbf{3.3_{\pm 0.3}}$	$\mathbf{1.8_{\pm 0.3}}$	$\mathbf{2.8_{\pm 0.4}}$	$\mathbf{3.3_{\pm 0.4}}$
Race	$\underline{5.2_{\pm 0.8}}$	$\underline{4.7_{\pm 0.7}}$	$\underline{5.9_{\pm 0.6}}$	$\underline{3.2_{\pm 0.4}}$	$3.8_{\pm 0.6}$	$4.2_{\pm 0.7}$
Ethnicity	$3.0_{\pm 0.5}$	$3.6_{\pm 0.3}$	$4.4_{\pm 0.4}$	$2.6_{\pm 0.6}$	$3.5_{\pm 0.3}$	$3.9_{\pm 0.6}$
Age	$3.5_{\pm 0.3}$	$3.7_{\pm 0.4}$	$4.6_{\pm 0.4}$	$3.0_{\pm 0.7}$	$\underline{4.1_{\pm 0.8}}$	$\underline{4.4_{\pm 0.8}}$
SES	$2.6_{\pm 0.4}$	$3.1_{\pm 0.6}$	$3.8_{\pm 0.3}$	$2.4_{\pm 0.5}$	$3.3_{\pm 0.4}$	$3.7_{\pm 0.5}$

6. Discussion

In this paper, we have presented a novel approach to address the challenges of fairness in clinical predictions using EHRs. Our findings suggest that the FairEHR-CLP framework, which integrates patient demographics, longitudinal data, and clinical notes through a unique two-stage process: synthetic counterpart generation and fairness-aware predictions with CL, significantly reduces disparities in error rates across different demographic subgroups. This improvement is critical in the context of healthcare, where equitable treatment and diagnosis are paramount. The integration of contrastive learning in fairness-aware predictions, combined with our novel fairness metric, represents a substantial advancement in the pursuit of equitable healthcare outcomes.

Limitations and Future Work. A concern in our study is the quality of synthetic data generated. Inaccuracies in capturing the complexity of real patient data could limit the model's effectiveness in mitigating biases. Future research should explore diverse synthetic data generation techniques, especially for longitudinal data and notes, to identify those that most accurately mirror the statistical characteristics of real data. Additionally, our approach encounters challenges with ambiguous categories in sensitive attributes, such as 'Unknown' or 'Other'. Refining categorization strategies is crucial to address biases more precisely. We will also extend our experiments to various clinical contexts, thereby enhancing the robustness and adaptability of our approach.

7. Broader Impacts

This paper introduces a general framework aimed at enhancing fairness in clinical predictions using multimodal EHRs by addressing social biases from demographic factors. Our approach highlights the potential for more equitable healthcare outcomes through ethically conscious AI, underscoring the importance of responsible usage. FairEHR-CLP offers a promising avenue to close the disparities gap in health outcomes by ensuring more accurate and unbiased healthcare predictive models, paving the way for a more inclusive future in medical decision-making.

Acknowledgments

This project was supported by grant number R01HS024096 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. Publicly available clinical bert embeddings. *NAACL HLT 2019*, page 72, 2019.

Andrew D Boyd, Rosa Gonzalez-Guarda, Katharine Lawrence, Crystal L Patil, Miriam O Ezenwa, Emily C O'Brien, Hyung Paek, Jordan M Braciszewski, Oluwaseun Adeyemi, Allison M Cuthel, et al. Potential bias and lack of generalizability in electronic health

record data: reflections on health equity from the national institutes of health pragmatic trials collaboratory. *Journal of the American Medical Informatics Association*, page ocad115, 2023.

Richard J Chen, Judy J Wang, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6):719–742, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Marshall H Chin, Nasim Afsar-Manesh, Arlene S Bierman, Christine Chang, Caleb J Colón-Rodríguez, Prashila Dullabh, Deborah Guadalupe Duran, Malika Fair, Tina Hernandez-Boussard, Maia Hightower, et al. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. *JAMA Network Open*, 6 (12):e2345050–e2345050, 2023.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33: 8765–8775, 2020.

Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. *Advances in Neural Information Processing Systems*, 34:12091–12103, 2021.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6840–6849, 2023.

Benedetta Giovanola and Simona Tiribelli. Beyond bias and discrimination: redefining the ai ethics principle of fairness in healthcare machine-learning algorithms. *AI & society*, 38 (2):549–563, 2023.

Colin Green, Bruce Hollingsworth, and Miaoqing Yang. The impact of social health insurance on rural populations. *The European Journal of Health Economics*, 22:473–483, 2021.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Vasileios Iosifidis and Eirini Ntoutsi. Dealing with bias via data augmentation in supervised learning scenarios. *Jo Bates Paul D. Clough Robert Jäschke*, 24(11), 2018.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

Zhimeng Jiang, Xiaotian Han, Chao Fan, Fan Yang, Ali Mostafavi, and Xia Hu. Generalized demographic parity for group fairness. In *International Conference on Learning Representations*, 2022.

Alistair Johnson, Tom Pollard, and Roger Mark. Mimic-iii clinical database carevue subset (version 1.4), 2022.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Michael Kim, Omer Reingold, and Guy Rothblum. Fairness through computationally-bounded awareness. *Advances in neural information processing systems*, 31, 2018.

Paulyne Lee, Maxine Le Saux, Rebecca Siegel, Monika Goyal, Chen Chen, Yan Ma, and Andrew C Meltzer. Racial and ethnic disparities in the management of acute pain in us emergency departments: meta-analysis and systematic review. *The American journal of emergency medicine*, 37(9):1770–1777, 2019.

Jin Li, Benjamin J Cairns, Jingsong Li, and Tingting Zhu. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *NPJ Digital Medicine*, 6(1):98, 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 2847–2851. IEEE, 2019.

Mohsen Nayebi Kerdabadi, Arya Hadizadeh Moghaddam, Bin Liu, Mei Liu, and Zijun Yao. Contrastive learning of temporal distinctiveness for survival analysis in electronic health

records. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1897–1906, 2023.

Suraj Rajendran, Weishen Pan, Mert R Sabuncu, Yong Chen, Jiayu Zhou, and Fei Wang. Learning across diverse biomedical data modalities and cohorts: Challenges and opportunities for innovation. *Patterns*, 2024.

Jon Raso, Pramod Kamalapathy, Andrew S Cuthbert, Alyssa Althoff, Pradip Ramamurti, and Brian C Werner. Social determinants of health disparities are associated with increased costs, revisions, and infection in patients undergoing arthroscopic rotator cuff repair. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, 39(3):673–679, 2023.

Sonish Sivarajkumar, Yufei Huang, and Yanshan Wang. Fair patient model: Mitigating bias in the patient representation learned from the electronic health records. *arXiv preprint arXiv:2306.03179*, 2023.

Ran Sun, Selen Bozkurt, Marcy Winget, Mark R Cullen, Tina Seto, and Tina Hernandez-Boussard. Characterizing patient flow after an academic hospital merger and acquisition. *The American Journal of Managed Care*, 27(10):e343–e348, 2021.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. Contrastive learning reduces hallucination in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13618–13626, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Monica B Vela, Amarachi I Erondu, Nichole A Smith, Monica E Peek, James N Woodruff, and Marshall H Chin. Eliminating explicit and implicit biases in health care: evidence and research needs. *Annual review of public health*, 43:477–501, 2022.

Yuqing Wang, Yun Zhao, Rachael Callcut, and Linda Petzold. Empirical analysis of machine learning configurations for prediction of multiple organ failure in trauma patients. *arXiv preprint arXiv:2103.10929*, 2021.

Yuqing Wang, Yun Zhao, Rachael Callcut, and Linda Petzold. Integrating physiological time series and clinical notes with transformer for early prediction of sepsis. *arXiv preprint arXiv:2203.14469*, 2022a.

Yuqing Wang, Yun Zhao, and Linda Petzold. Predicting the need for blood transfusion in intensive care units with reinforcement learning. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10, 2022b.

Yuqing Wang, Yun Zhao, and Linda Petzold. Are large language models ready for healthcare? a comparative study on clinical language understanding. *arXiv preprint arXiv:2304.05368*, 2023.

Zhibo Wang, Xiaowei Dong, Henry Xue, Zhifei Zhang, Weifeng Chiu, Tao Wei, and Kui Ren. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10379–10388, 2022c.

Tingyi Wanyan, Hossein Honarvar, Suraj K Jaladanki, Chengxi Zang, Nidhi Naik, Sulaiman Somani, Jessica K De Freitas, Ishan Paranjpe, Akhil Vaid, Jing Zhang, et al. Contrastive learning improves critical event prediction in covid-19 patients. *Patterns*, 2(12), 2021.

Jo Ellen Wilson, Matthew F Mart, Colm Cunningham, Yahya Shehabi, Timothy D Girard, Alasdair MJ MacLullich, Arjen JC Slooter, and E Wesley Ely. Delirium. *Nature Reviews Disease Primers*, 6(1):90, 2020.

Jenny Yang, Andrew AS Soltan, David W Eyre, Yang Yang, and David A Clifton. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digital Medicine*, 6(1):55, 2023.

Chengxi Zang and Fei Wang. Scehr: Supervised contrastive learning for clinical risk prediction using electronic health records. In *Proceedings. IEEE International Conference on Data Mining*, volume 2021, page 857. NIH Public Access, 2021.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Fengda Zhang, Kun Kuang, Long Chen, Yuxuan Liu, Chao Wu, and Jun Xiao. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *The Eleventh International Conference on Learning Representations*, 2022.

Yun Zhao, Qinghang Hong, Xinlu Zhang, Yu Deng, Yuqing Wang, and Linda Petzold. Bertsurv: Bert-based survival models for predicting outcomes of trauma patients. *arXiv preprint arXiv:2103.10928*, 2021a.

Yun Zhao, Yuqing Wang, Junfeng Liu, Haotian Xia, Zhenni Xu, Qinghang Hong, Zhiyang Zhou, and Linda Petzold. Empirical quantitative analysis of covid-19 forecasting models. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 517–526. IEEE, 2021b.

Appendix A. EHR Data Examples

We provide a sample of EHR data from MIMIC-IV for one patient, including both real and synthetic data, encompassing static demographic features, longitudinal data, and clinical notes.

Demographics. Figure 6 provides an example of real and synthetic demographic features for a patient.

Feature	Value
Gender	Male
Age	66
Race	White
Ethnicity	Non-Hispanic/Non-Latino
Insurance Type	Medicare

(a) Real demographic features.

Feature	Value
Gender	Female
Age	81
Race	NHPI
Ethnicity	Other
Insurance Type	Medicaid

(b) Synthetic demographic features.

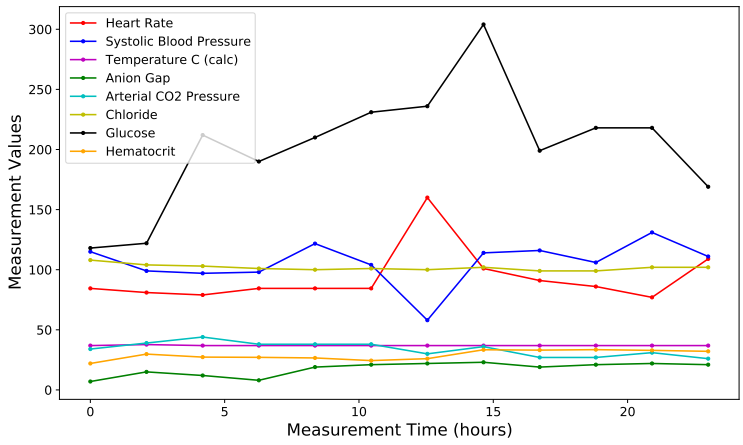
Figure 6: Demographic examples (real and synthetic) from an EHR data sample. NHPI denotes Native Hawaiians and Pacific Islanders.

Longitudinal Data. Figure 7 presents an example of real and synthetic longitudinal data for a patient.

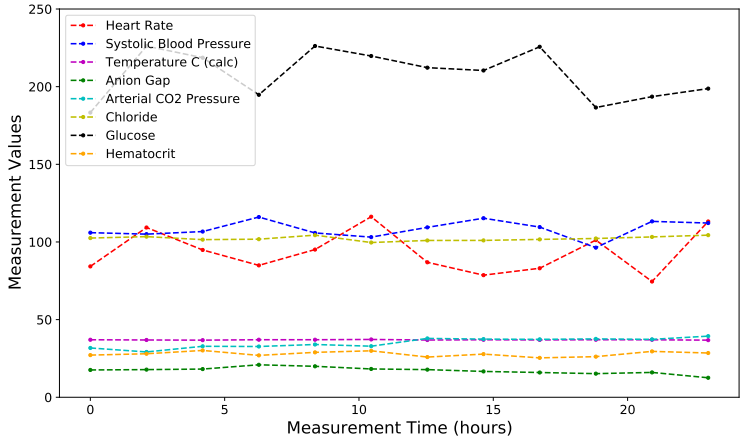
Notes. We provide the following examples of real and synthetic clinical notes for the patient described earlier. The texts in bold indicate the patient's primary medical or health conditions.

Real: The patient exhibited a **progressive exacerbation of dyspnea and edema** over four days, ultimately found in a tripod position with a resting arterial oxygen saturation of 90%. Initially managed as a COPD exacerbation and later excluding non-ST-elevation myocardial infarction (NSTE-ACS), the patient was stabilized in the ICU with **BiPAP support**. Subsequent cardiac catheterization identified **multivessel coronary artery disease**, including **in-stent stenosis** in the left anterior descending artery. Despite these complications, the patient remained hemodynamically stable in **normal sinus rhythm** and was subsequently shifted for **revascularization evaluation**.

Synthetic: The patient arrived with an **escalating severity in breathing difficulty and swelling** over four days, observed in a respiratory distress posture with an oxygen saturation level at 90%. Initially treated for a chronic obstructive pulmonary disease flare-up, myocardial infarction without ST-elevation was later ruled out. The patient was maintained in a stable condition under **BiPAP respiratory support** in the intensive care unit. Cardiac catheterization conducted recently revealed a **complex coronary artery disease**, notably including a **narrowed segment within a stent** in the left anterior descending artery. Notwithstanding these heart-related complexities, the patient's hemodynamic status was stable with a **normal heart rhythm**, leading to a transfer for further assessment and planning for **revascularization therapy**.



(a) Real longitudinal features.



(b) Synthetic longitudinal features.

Figure 7: Longitudinal feature examples (real and synthetic) from an EHR data sample.

Appendix B. Synthetic Notes Review Guidelines

For quality assurance, we randomly select 100 synthetic patient notes from each of the three datasets. The manual review process adheres to the following principles:

(1) **Exclusion of Demographic Factors**: Demographic identifiers such as gender, race, age, ethnicity, and socioeconomic status (SES) associated with insurance type are excluded to ensure the notes primarily focus on health conditions, aligning with our objective to mitigate social bias stemming from demographic factors in clinical predictions.

(2) **Inclusion of Major Treatments and Diagnoses**: We verify the presence and accuracy of essential health information, including diagnoses, treatments, and medical history, to ensure the synthetic notes retain critical medical content for predictive modeling relevance.

(3) **Consistency with Real Records:** The synthetic notes are compared against authentic clinical records to ascertain their fidelity in mirroring the structure, terminology, and clinical reasoning typical of real medical documentation.

Appendix C. Notations

All the notations corresponding to the FairEHR-CLP framework are summarized in Table 6 and Table 7.

Table 6: Notation definitions in FairEHR-CLP (Part 1).

Reference	Notation	Description
Section 3.1 Problem Formulation	$\mathscr{D}$	Dataset with patient data, labels, and sensitive attributes
	$x_k \in \mathcal{X}$	Input features from demographics, longitudinal data, and clinical notes
	$y_k \in \{0,1\} \subseteq \mathcal{Y}$	Binary target label for patient outcomes
	$s_k \in \mathcal{S}$	Sensitive attributes from demographic features
	$\mathcal{S}$	Set of sensitive attributes including gender, race, ethnicity, age, and SES
	$f : \mathcal{X} \to \mathcal{Y}$	Prediction model from features to outcomes
Section 3.2 Longitudinal Data EHR-GAN	G	Generator
	G_e	Encoder component of generator
	G_d	Decoder component of generator
	D	Discriminator in EHR-GAN
	x	Input data to encoder
	z	Latent space representation from encoder
	v	Random noise input to decoder
	$\hat{x}$	Synthetic data generated by decoder
	l_{dis}	Discriminative loss by discriminator
	l_{adv}	Adversarial loss for generator
	l_{fm}	Feature matching loss for generator
	$\beta_0, \beta_1, \beta_2$	Weighting coefficients for loss components
	$p_z(z)$	Prior distribution over latent space
	$p_x(x)$	Distribution of real data
	$f(\cdot)$	Output of intermediate layer in discriminator
	y_i	Label indicating real or synthetic data

Continued on next page

Appendix D. Implementation Details

All of the experiments are conducted on four NVIDIA A100 GPUs. We apply a random train/test split in an 80%/20% ratio for each dataset. In training our EHR-GAN, we primarily adhere to the experimental settings of the baseline EHR-M-GAN as described in Li et al. (2023), omitting the discrete-valued time-series data and focusing solely on continuous longitudinal data. Maximum Mean Discrepancy (MMD) is employed to assess the similarity between real and synthetic data, aiding in the adjustment of hyperparameters in EHR-GAN for quality control. For detailed implementation specifics, please refer to Li et al. (2023). Based on the results in the original paper and our experiments, we set the MMD threshold at 0.68 to ensure a reasonable quality of synthetic longitudinal data. After the first stage of FairEHR-CLP, which involves synthetic counterpart generation, and considering that we have both synthetic and real data for each patient in the training set (demographics,

Table 7: Notation definitions in FairEHR-CLP (Part 2).

Reference	Notation	Description
Section 3.3 Fairness-aware Prediction with Contrastive Learning	x^+	Positive samples: synthetic counterparts
	x^-	Negative samples: other patient data in minibatch
	e_d	Encoded demographic data
	e_l	Encoded longitudinal data
	e_n	Encoded clinical notes
	F_{fusion}	MLP-based fusion function
	θ_{fusion}	Trainable parameters in fusion layer
	e, e^{syn}	Integrated representations for real and synthetic data
	F_{DR}	Dynamic Relevance (DR) layer function
	w	Adjustable weights in DR layer
	σ	Sigmoid function
	e^{adj}, $e^{adj,syn}$	Adjusted embeddings post-DR layer
	l_{CF}	Fairness-oriented contrastive loss
	l_{CE}	Cross entropy loss
	N	Number of embeddings in minibatch
	τ	Temperature parameter
	γ	Regularization parameter
	μ_{syn}^{adj}	Mean of adjusted synthetic embeddings
	y_k	True label for each real embedding
	$C(e_k^{adj})$	softmax probability of predicted class
	α	Parameter balancing fairness and performance

longitudinal, and notes), we employ fairness-aware predictions with CL. The Adam optimizer is utilized with its default parameters for optimization. The hyperparameter search space for all datasets is detailed in Table 8. Hyperparameter optimization is conducted via random search.

Table 8: Hyperparameter search space of FairEHR-CLP on three datasets.

Hyperparameters	Search Space
Batch size	[16, 32, 64, 128, 256]
Learning rate	[1e−5, 5e−5, 1e−6, 5e−6]
# of epochs	[20, 30, 50]
τ	[0.1, 0.3, 0.5, 0.7]
λ	[0.3, 0.4, 0.5, 0.6, 0.7]

Appendix E. Datasets

We summarize the clinical predictors, including vital signs and laboratory measurements, used in the MIMIC-III/IV and STARR datasets, in Table 9 and Table 10, respectively. These predictors are used for all three prediction tasks: classifying delirium, OUD, and 30-day readmission. Due to the absence of explicit codes for identifying surgical patients

in the MIMIC-III/IV datasets, we extract patient data from the Surgical Intensive Care Unit (SICU). Delirium refers to a condition characterized by confusion and a reduced ability to maintain attention and clear awareness, with its incidence increasing with age (Wilson et al., 2020). Bias could arise from healthcare professionals' age-related stereotypes, leading to underdiagnosis in older patients or overdiagnosis in those with pre-existing cognitive impairments, which could affect treatment decisions and ultimately patient recovery. OUD is a medical condition characterized by the problematic use of opioid medications, commonly prescribed for pain relief, and can lead to a high risk of dependence and misuse. OUD can be influenced by biases related to prescribing practices, such as biases based on patients' race or socioeconomic status, which might affect the likelihood of being prescribed opioids, the dosage, or the duration of use, potentially leading to disparities in the risk of developing OUD. Lastly, 30-day readmission is defined as the rehospitalization of a patient within 30 days following their discharge from a hospital, serving as an important indicator of the quality of care and patient outcomes. For example, elderly patients might receive less comprehensive discharge planning or follow-up care due to assumptions about their support systems or ability to manage their own care, leading to higher readmission rates.

Table 9: Summary of clinical predictors in longitudinal data for MIMIC-III/IV datasets.

Category	**Predictors**
Vital Signs	Heart Rate, Systolic Blood Pressure, Diastolic Blood Pressure, Mean Blood Pressure, Respiratory Rate, Body Temperature, Oxygen Saturation
Blood Gases	Arterial Base Excess, Arterial Carbon Dioxide Pressure, Arterial Oxygen Pressure, Arterial pH
Renal Function	Blood Urea Nitrogen, Creatinine
Metabolic Panel	Ionized Calcium, Serum Chloride, Serum Glucose, Fingerstick Glucose, Anion Gap, Serum Bicarbonate, Magnesium, Phosphorus, Serum Potassium, Serum Sodium
Hematology	Serum Hematocrit, Hemoglobin, Platelet Count, White Blood Cell Count

Table 10: Summary of clinical predictors in longitudinal data for the STARR dataset.

Category	**Predictors**
Vital Signs	Heart Rate, Pulse, Respiratory Rate, Oxygen Saturation, Body Temperature, Systolic Blood Pressure, Diastolic Blood Pressure
Blood Gases	CO2, Anion Gap
Renal Function	Blood Urea Nitrogen, Creatinine
Metabolic Panel	Calcium, Chloride, Glucose, Potassium, Sodium
Hematology	Hematocrit, Hemoglobin, Mean Corpuscular Volume, Mean Corpuscular Hemoglobin, White Blood Cell Count, Platelet Count, Red Blood Cell Count, Red Cell Distribution Width, Mean Corpuscular Hemoglobin Concentration
Liver Function	ALT (SGPT), Albumin

Appendix F. Ablation Study

Data Modalities. Table 11 and Table 12 demonstrate the impact of different data modalities on the performance of our FairEHR-CLP method for the MIMIC-III and MIMIC-IV datasets, respectively. Similar to the trends observed in the STARR dataset, combining demographic ($\mathcal{D}$) and longitudinal ($\mathcal{L}$) data surpasses the mix of $\mathcal{D}$ with clinical notes. In MIMIC-III, using the complete dataset ($\mathcal{D}+\mathcal{L}+\mathcal{N}$) results in a 5.8% increase in F1 and a 4.9% improvement in AUROC compared to the second-best combination ($\mathcal{D}+\mathcal{L}$). Likewise, for MIMIC-IV, employing the full dataset ($\mathcal{D}+\mathcal{L}+\mathcal{N}$) leads to a 2.0% enhancement in F1 and a 2.4% increase in AUROC over the second-best results ($\mathcal{D}+\mathcal{L}$). In terms of fairness metrics, the full dataset consistently yields lower EO and EDDI values compared to the use of partial data. This highlights the effectiveness of comprehensive patient representation in achieving more equitable predictions.

Table 11: Effects of different data modalities as inputs for FairEHR-CLP on the MIMIC-III dataset.

Data Modalities	Delirium				OUD				30-Day Readmission			
	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)
$\mathcal{D}$	$77.3_{\pm 1.8}$	$79.6_{\pm 1.7}$	$8.1_{\pm 1.0}$	$5.8_{\pm 0.7}$	$80.5_{\pm 1.6}$	$84.3_{\pm 1.4}$	$5.0_{\pm 0.9}$	$4.4_{\pm 0.8}$	$80.6_{\pm 1.5}$	$83.4_{\pm 1.3}$	$6.1_{\pm 0.6}$	$6.3_{\pm 0.7}$
$\mathcal{D}+\mathcal{L}$	$81.0_{\pm 1.5}$	$85.3_{\pm 1.3}$	$6.5_{\pm 0.8}$	$4.5_{\pm 0.6}$	$83.8_{\pm 1.7}$	$87.9_{\pm 1.5}$	$4.2_{\pm 0.7}$	$3.3_{\pm 0.6}$	$83.9_{\pm 1.4}$	$87.1_{\pm 1.2}$	$5.3_{\pm 0.5}$	$5.0_{\pm 0.6}$
$\mathcal{D}+\mathcal{N}$	$79.8_{\pm 1.6}$	$83.3_{\pm 1.4}$	$7.0_{\pm 0.7}$	$5.1_{\pm 0.6}$	$82.4_{\pm 1.8}$	$86.5_{\pm 1.6}$	$4.6_{\pm 0.6}$	$3.8_{\pm 0.7}$	$80.7_{\pm 1.3}$	$84.5_{\pm 1.1}$	$5.7_{\pm 0.4}$	$5.1_{\pm 0.8}$
$\mathcal{D}+\mathcal{L}+\mathcal{N}$	$\mathbf{85.5_{\pm 1.2}}$	$\mathbf{89.7_{\pm 1.1}}$	$6.2_{\pm 0.3}$	$3.8_{\pm 0.5}$	$\mathbf{89.4_{\pm 1.4}}$	$\mathbf{91.9_{\pm 1.5}}$	$3.7_{\pm 0.5}$	$2.0_{\pm 0.4}$	$\mathbf{88.2_{\pm 1.3}}$	$\mathbf{91.4_{\pm 1.1}}$	$3.3_{\pm 0.4}$	$2.1_{\pm 0.6}$

Table 12: Effects of different data modalities as inputs for FairEHR-CLP on the MIMIC-IV dataset.

Data Modalities	Delirium				OUD				30-Day Readmission			
	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)
$\mathcal{D}$	$75.6_{\pm 1.8}$	$77.8_{\pm 1.7}$	$8.5_{\pm 1.0}$	$6.2_{\pm 0.7}$	$81.9_{\pm 1.6}$	$84.9_{\pm 1.4}$	$5.4_{\pm 0.8}$	$4.7_{\pm 0.6}$	$76.3_{\pm 1.7}$	$79.5_{\pm 1.5}$	$6.7_{\pm 0.7}$	$6.9_{\pm 0.8}$
$\mathcal{D}+\mathcal{L}$	$78.2_{\pm 1.5}$	$81.6_{\pm 1.3}$	$7.0_{\pm 0.8}$	$5.1_{\pm 0.6}$	$83.7_{\pm 1.7}$	$87.8_{\pm 1.6}$	$4.7_{\pm 0.7}$	$3.6_{\pm 0.5}$	$78.5_{\pm 1.6}$	$82.4_{\pm 1.4}$	$6.0_{\pm 0.6}$	$5.7_{\pm 0.7}$
$\mathcal{D}+\mathcal{N}$	$76.9_{\pm 1.6}$	$80.3_{\pm 1.5}$	$7.8_{\pm 0.9}$	$5.6_{\pm 0.7}$	$82.5_{\pm 1.8}$	$86.7_{\pm 1.5}$	$5.1_{\pm 0.6}$	$4.2_{\pm 0.6}$	$77.1_{\pm 1.5}$	$80.7_{\pm 1.3}$	$6.4_{\pm 0.8}$	$6.1_{\pm 0.9}$
$\mathcal{D}+\mathcal{L}+\mathcal{N}$	$\mathbf{78.8_{\pm 1.2}}$	$\mathbf{82.4_{\pm 1.0}}$	$6.1_{\pm 0.4}$	$3.5_{\pm 0.3}$	$\mathbf{84.8_{\pm 1.6}}$	$\mathbf{88.9_{\pm 1.5}}$	$1.5_{\pm 0.3}$	$3.0_{\pm 0.6}$	$\mathbf{81.6_{\pm 1.8}}$	$\mathbf{86.4_{\pm 1.6}}$	$2.8_{\pm 0.7}$	$5.2_{\pm 0.9}$

Model Components. Table 13 and Table 14 demonstrate the impact of different model components on FairEHR when employing the full dataset for the MIMIC-III and MIMIC-IV datasets, respectively. During the training phase, synthetic counterparts are maintained for data augmentation when CL is not applied. For both datasets, the removal of both CL and DR leads to the most significant performance decline. Specifically, for MIMIC-III, the configuration without CL and DR (Full w/o CL + DR) results in a performance decrease of 2.7% in F1 and 3.3% in AUROC. For MIMIC-IV, the same configuration leads to a decrease of 4.2% in F1 and 4.4% in AUROC. In this case, it yields the most biased predictions with higher EO and EDDI values, while removing CL or DR moderately reduces performance but slightly increases fairness metrics.

Table 13: Effects of different model components for FairEHR-CLP (full) on the MIMIC-III dataset ($\mathcal{D} + \mathcal{L} + \mathcal{N}$).

Model Components	Delirium				OUD				30-Day Readmission			
	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)
Full w/o CL + DR	$83.2_{\pm 1.3}$	$86.4_{\pm 1.2}$	$7.0_{\pm 0.8}$	$5.1_{\pm 0.6}$	$87.1_{\pm 1.6}$	$89.4_{\pm 1.4}$	$4.4_{\pm 0.7}$	$3.9_{\pm 0.5}$	$85.8_{\pm 1.2}$	$88.6_{\pm 1.3}$	$4.8_{\pm 0.5}$	$3.9_{\pm 0.6}$
Full w/o CL	$83.4_{\pm 1.1}$	$86.9_{\pm 1.0}$	$6.6_{\pm 0.7}$	$4.7_{\pm 0.4}$	$88.3_{\pm 1.5}$	$90.6_{\pm 1.3}$	$4.1_{\pm 0.6}$	$3.4_{\pm 0.4}$	$86.9_{\pm 1.1}$	$89.8_{\pm 1.2}$	$4.3_{\pm 0.4}$	$3.6_{\pm 0.5}$
Full w/o DR	$84.2_{\pm 1.0}$	$87.5_{\pm 0.9}$	$6.3_{\pm 0.6}$	$4.3_{\pm 0.3}$	$88.9_{\pm 1.6}$	$91.1_{\pm 1.2}$	$3.9_{\pm 0.5}$	$3.1_{\pm 0.3}$	$87.5_{\pm 1.0}$	$90.2_{\pm 1.3}$	$3.4_{\pm 0.3}$	$2.9_{\pm 0.4}$
Full	$\mathbf{85.5_{\pm 1.2}}$	$\mathbf{89.7_{\pm 1.1}}$	$6.2_{\pm 0.3}$	$3.8_{\pm 0.5}$	$\mathbf{89.4_{\pm 1.4}}$	$\mathbf{91.9_{\pm 1.5}}$	$3.7_{\pm 0.5}$	$2.0_{\pm 0.4}$	$\mathbf{88.2_{\pm 1.3}}$	$\mathbf{91.4_{\pm 1.1}}$	$3.3_{\pm 0.4}$	$2.1_{\pm 0.6}$

Table 14: Effects of different model components for FairEHR-CLP (full) on the MIMIC-IV dataset ($\mathcal{D} + \mathcal{L} + \mathcal{N}$).

Model Components	Delirium				OUD				30-Day Readmission			
	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)	F1 (↑)	AUROC (↑)	EO (↓)	EDDI (↓)
Full w/o CL + DR	$76.4_{\pm 1.3}$	$79.8_{\pm 1.2}$	$7.5_{\pm 0.9}$	$6.3_{\pm 0.7}$	$80.5_{\pm 1.7}$	$84.9_{\pm 1.6}$	$5.2_{\pm 0.8}$	$4.5_{\pm 0.6}$	$78.3_{\pm 1.7}$	$82.2_{\pm 1.5}$	$3.5_{\pm 0.6}$	$6.6_{\pm 0.8}$
Full w/o CL	$76.8_{\pm 1.1}$	$80.9_{\pm 1.1}$	$6.9_{\pm 0.8}$	$5.7_{\pm 0.5}$	$83.3_{\pm 1.5}$	$87.1_{\pm 1.4}$	$4.8_{\pm 0.7}$	$3.9_{\pm 0.5}$	$80.6_{\pm 1.6}$	$84.7_{\pm 1.4}$	$3.1_{\pm 0.5}$	$6.2_{\pm 0.7}$
Full w/o DR	$77.6_{\pm 1.0}$	$81.7_{\pm 1.0}$	$6.4_{\pm 0.7}$	$5.1_{\pm 0.4}$	$84.1_{\pm 1.4}$	$87.6_{\pm 1.3}$	$4.3_{\pm 0.6}$	$3.6_{\pm 0.4}$	$81.1_{\pm 1.5}$	$85.8_{\pm 1.3}$	$2.9_{\pm 0.4}$	$5.5_{\pm 1.0}$
Full	$\mathbf{78.8_{\pm 1.2}}$	$\mathbf{82.4_{\pm 1.0}}$	$6.1_{\pm 0.4}$	$3.5_{\pm 0.3}$	$\mathbf{84.8_{\pm 1.6}}$	$\mathbf{88.9_{\pm 1.5}}$	$1.5_{\pm 0.3}$	$3.0_{\pm 0.6}$	$\mathbf{81.6_{\pm 1.8}}$	$\mathbf{86.4_{\pm 1.6}}$	$2.8_{\pm 0.7}$	$5.2_{\pm 0.9}$