Implementing Fairness in Real-World Healthcare Machine Learning through *Datasheet for Database*

by

Anand Murugan

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2024

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Healthcare Machine Learning (HML) models are revolutionizing the healthcare industry, promising improved patient outcomes and enhanced public health. However, it is essential to ensure fairness, i.e., models delivering equitable performance to all individuals, irrespective of their inherent or acquired characteristics. This requires a thorough examination of the data used and the specific applications of these models.

This study conducted a six-year systematic survey of models trained on the Medical Information Mart for Intensive Care (MIMIC) clinical research database (CRD) – one of the most popular and widely used HML databases to explore the link between data and fairness in HML.

The results were striking: for the popular MIMIC IV – ICU mortality task, a naive baseline outperformed the state-of-the-art (SOTA) model in prediction performance, demonstrating greater fairness across subgroups (while still somewhat unfair). These findings demonstrate the urgent need to integrate fairness into healthcare machine learning models and a greater need to include practitioners in HML modeling.

To achieve this, we propose a data-centric approach to fairness through our 'Datasheet for MIMIC IV v2.0 CRD', modeled after the recent works recommending datasheets for datasets. Given that MIMIC is large and complex, this datasheet will assist practitioners in identifying data anomalies and task-specific feature-target relationships during modeling, thereby fostering the development of equitable HML models.

Acknowledgments

The work presented in this thesis would not have been possible without the invaluable assistance of my advisors, Professor. Alexander Wong and Professor. Sirisha Rambhatla. Their unwavering support and mentorship over the past two years have been instrumental. They allowed me to explore diverse ideas, significantly contributing to my growth as a researcher.

This work was supported by the J.W. Graham Trust at the University of Waterloo. I am thankful for their generous support.

Dedication

This is sincerely dedicated to my family and friends, who are the rock-solid foundation of my life and the silent force that propels me forward.

To my Mom, Dad, and Sister: your unwavering support and love have been my beacon of hope. You were there to lend a helping hand whenever I stumbled. Despite the vast seas and countries separating us, your kindness and compassion found their way to me. Your faith in me has greatly influenced my journey at the University of Waterloo; every step I took was encouraged by the assurance of your steadfast support. This accomplishment is not just mine; it is evidence of our family's unwavering love and strength. Thank you!!

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the rapidly evolving landscape of Canadian healthcare, the system faces critical challenges: an aging population [1], the impacts of climate change [2], and the repercussions of the COVID-19 pandemic [3]. CIHI report forecasts a 68% increase in the population aged 65 and over in the next two decades, a shift to significantly strain healthcare services and escalate care demands for older adults [4]. Concurrently, climate change exacerbates healthcare burdens, with more frequent extreme weather events—heatwaves, flooding, and wildfires—increasing healthcare demand. The COVID-19 pandemic further revealed and deepened the healthcare system's vulnerabilities, emphasizing the gap in providing timely and equitable care [5]. Together, these issues underscore the pressing need to build up the healthcare workforce and develop tools to assist them in delivering quality healthcare for all.

Artificial Intelligence (AI)/Machine Learning (ML) models can help achieve this goal. It has emerged as a powerful tool in the healthcare industry, enabling the creation of accurate prediction models that can significantly improve public health outcomes [6, 7]. However, in the high-stakes context of healthcare, it is critical to ensure unbiased predictive outcomes [8]. Despite significant progress in healthcare machine learning (HML), unfairness persists due to biases in both data and algorithms. Data biases include minority bias, where certain demographic groups are underrepresented; missing data bias, where data are randomly missing, impacting model reliability; and label bias, where inaccuracies in labels can misguide learning algorithms [9]. Algorithmic biases occur when the models themselves systematically generate unfair outcomes for certain groups, often due to inherent flaws in their design choices, such as the use of certain optimization functions, regularizations, etc, or training processes. [10–12]. Addressing these biases in Canada's diverse healthcare landscape is crucial to ensure that HML benefits everyone, regardless of age, race, gender, or

socio-economic status.

FairML refers to the process of designing, developing, and deploying machine learning models that ensure equitable treatment for all individuals, regardless of their inherent characteristics [9]. It is an active area of research studying how models can perpetuate inequalities and develop methods to mitigate these issues. A common strategy in ML fairness involves implementing fairness constraints during model training [13–20]. These constraints, such as decision boundary covariance [21], are algorithmic modifications made during the training process to ensure fair outcomes across different groups, distinguished by sensitive attributes like ethnicity or gender [22–25]. However, a significant challenge in this model-centric approach is the data itself, which may encode real-world biases and inconsistencies, thus complicating the achievement of fairness in healthcare settings [26]. This insight can be traced back to 2018, when Chen et al. [15] proposed that the predictive fairness evaluation must consider model training in the context of the input data.

To investigate the progress towards fairness in HML, we analyze the state of fair HML research through the lens of Medical Information Mart for Intensive Care (MIMIC) [27, 28] database, one of the most popular clinical research databases (CRD) [18]; see Figure 1.1 for its popularity across the world. Our survey identified ICU mortality prediction as the most widely researched task in MIMIC III/IV. To analyze the ICU mortality task, we compared the performance and fairness properties (across ethnicity) of the current SOTA: STraTS [1] [29] with simple baseline models (such as XGBoost) on MIMIC IV. Surprisingly, we find that the SOTA model is outperformed by XGBoost in performance and group fairness metrics, with disparate impact (DI) metrics revealing substantial inequalities across ethnic subgroups [2]. This demonstrates that despite more than half a decade since [15], it has been challenging to incorporate fairness in HML modeling, highlighting a need to reconsider fairness in the context of (i) the task and (ii) the data driving the model, and ways to make it a central part of any HML analysis.

The aforementioned example is not to undermine the work by SOTA; it is to highlight the complexities involved in accomplishing healthcare fairness. For instance, MIMIC IV is a large database, which makes it especially challenging for practitioners to focus on modeling and fairness concurrently since they may only be interested in small parts of the database. With recent calls in the community to develop "Datasheets for Dataset", spearheaded by Gebru et al. [30], we develop a blueprint to enable the next generation of HML by introducing the 'Datasheet for CRD (MIMIC IV v2.0)' in Chapter 4. It is a

[1] STraTS reported SOTA performance on MIMIC III, and MIMIC IV is an augmented version of MIMIC III. No fairness metrics were reported by the paper.

[2] Moreover, note that MIMIC IV originates from Boston, and its use world-wide further underscores a need to understand its characteristics for downstream applications.

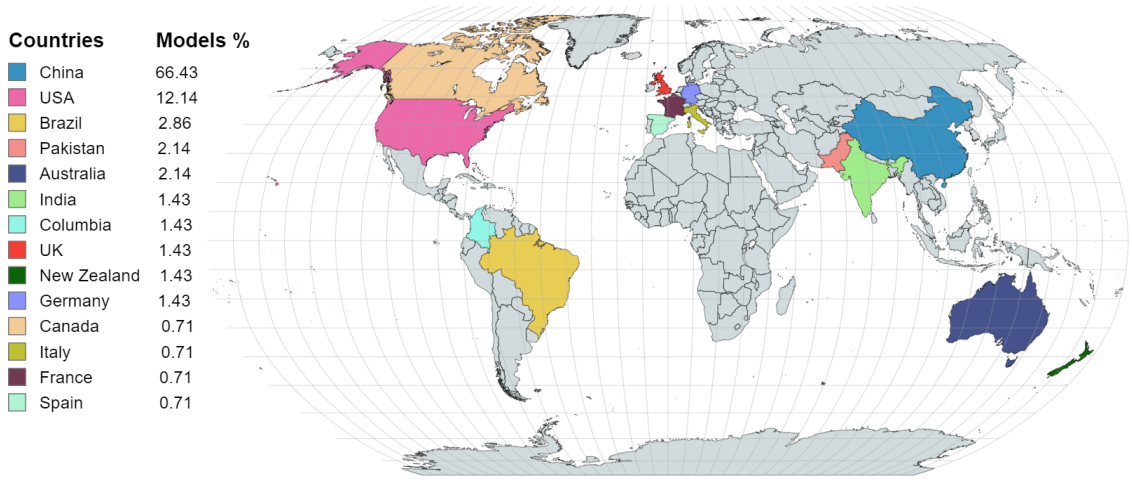


Figure 1.1: Distribution of healthcare prediction models using MIMIC data across the world from 2018 to 2024. The country list represents the widespread usage of MIMIC data for healthcare predictions across the globe, and the % value indicates the Healthcare ML prediction model researched by the respective country.

comprehensive resource tailored for complex clinical databases. Here, we call for data and tasks to take center stage in any HML modeling to accomplish fairness. To this end, our datasheet is designed to support HML practitioners in identifying data biases like data quality issues, representation bias, etc, across the database and analyzing task-specific associations between features and the target, enabling fair predictions for all individuals irrespective of their sensitive attributes.

To build a fair HML model that does not perpetuate societal biases, we advocate for a comprehensive examination of the data, the context of its use, and algorithmic biases together. This approach is especially significant in the light of recent works in fairness theory, which establish that it is not feasible to satisfy all fairness criteria at once [31]. There have also been calls to include end-users in determining key fairness goals [32]. Yet, the lack of fairness in the analysis (by SOTA) is alarming, and our work aims to make it easier for practitioners to incorporate these metrics.

The contributions highlighted in this thesis are detailed below:

1. In chapter 2, we discuss how important the data is for modeling, particularly the EHR data, considering HML predictive tools. In section 2.1, we detail the methods and techniques we used in the systematic survey of MIMIC III/IV trained HML models from 2018 to 2024. Survey results highlighted the predominant Healthcare ML risk

predictive tasks, with ICU Mortality prediction being the most widely researched (67%) globally.

2. Chapter 3 provides an overview of fairness in ML, specifically focusing on the Healthcare setting. It highlights the importance of fairness in high-stakes healthcare applications in addition to the model performance, while section 3.3 details how we can evaluate it.

3. Chapter 4 explains the role of Clinical Research database in HML modeling and provides an in-depth look at MIMIC IV v2.0. This chapter introduces the 'Datasheet for MIMIC IV v2.0'. It highlights the data inconsistencies across MIMIC IV and includes task-specific feature-target analysis to streamline fairness evaluations for equitable and trustworthy HML predictive models.

4. In chapter 5, we show how researchers can use the 'Datasheet for Database' for their modeling task by implementing an ICU Mortality prediction task utilizing the 'Datasheet for MIMIC IV v2.0'. Section 5.3 analyses the SOTA and several baseline model's performance, and section 5.3.2 highlights how fair the model's prediction is across the demographics. Analysis results underscore the immediate need to incorporate fairness during HML modeling.

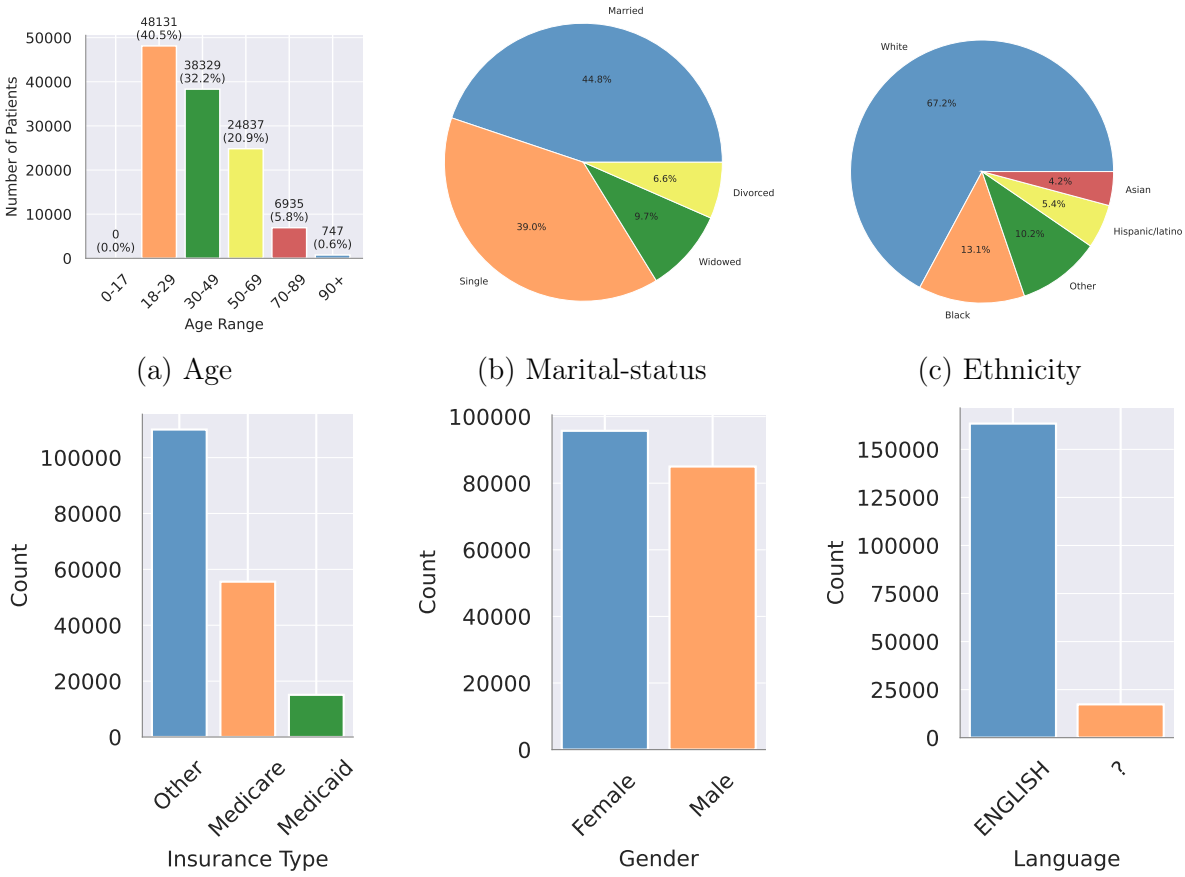
5. Chapter 6 highlights how the 'Datasheet for Database' can help build equitable and trustworthy HML models and discusses its potential in aiding synthetic data generation and Healthcare Gen AI.

Chapter 2

Healthcare ML data source - MIMIC Clinical Research Database

Data is the foundation of machine learning, and it is essential for models to learn, predict, and make informed decisions. Within the Healthcare Machine Learning (HML), the accuracy and efficacy of models are determined by the quality and relevance of Electronic Health Records (EHR). The Medical Information Mart for Intensive Care (MIMIC), a comprehensive repository for HML model development, is the leading Clinical Research Database (CRD).

MIMIC encompasses de-identified health data from patients admitted to Beth Israel Deaconess Medical Center's critical care units, with its latest iteration, MIMIC-IV, capturing records from 2008 to 2019. This database is meticulously organized into four modules: 'hosp' and 'icu' for hospital and intensive care data, 'ed' for emergency department insights, 'note' for clinician's notes, and 'cxr' for chest X-ray information [33]. Each module caters to specific research needs, making MIMIC an invaluable asset for advancing healthcare through machine learning. A comprehensive overview of the MIMIC IV CRD is provided in the Datasheet section, and Figure 2.1 shows the demographic attributes statistics of the entire MIMIC IV CRD where only age, marital status, ethnicity, insurance, gender, and language were recorded. There are 33 ethnic values recorded, which are grouped into White, Asian, Hispanic/Latino, Other, and Black following [20]. There are several representation issues, such as only males and females being recorded as part of the gender demographic and only English for language attributes while others are marked as '?'. HML practitioners need to be wary of these inconsistencies, and the Datasheet for Database captures these kinds of information in section 4.1.2 specifically in question C12.



(a) Age (b) Marital-status (c) Ethnicity

(d) Insurance (e) Gender (f) Language

Figure 2.1: MIMIC IV Sensitive attributes statistics.

2.1 Systematic Survey

We conducted a six-year systematic survey from 2018 to 2024 to examine MIMIC-trained HML models to identify common prediction tasks, monitor progress, and assess the field's current state.

2.1.1 Survey Method

PubMed and Google Scholar databases were extensively searched because of their medical focus, broad coverage, and potential to uncover emerging trends beyond specialized

Table 2.1: Inclusion and Exclusion Criteria of MIMIC trained Healthcare ML Studies.

Criterion	Included	Excluded
Study Design	Study that develops a prediction model	Review articles, database innovation studies, medical data mining studies, etc
MIMIC	III/IV	Other older versions
Outcome	Mortality, Readmission, LOS, Phenotype labeling/ICD code grouping	Other outcomes
Performance	AUC, sensitivity, specificity, accuracy, etc	No reported performance

databases. We adopted a broad search term approach following [34] to capture extensive research on MIMIC-trained HML models. PubMed is searched using *('Medical information mart for intensive care AND MIMIC AND MIMIC-IV') AND ('machine learning' OR 'artificial intelligence' OR 'deep learning' OR 'neural network' OR 'prediction model')* search terms yielding 819 works whereas, Google Scholar yielded 1000 records with the search term *'machine learning'|'prediction model'|'artificial intelligence'|'deep learning' AND 'MIMIC IV'|'the medical information mart for intensive care'|'MIMIC III'*. Only 220 studies were included after the initial abstract and title screening, as listed in Table 2.1.

Following PRISMA guideline [35], studies were screened to exclusively consider those utilizing MIMIC III/IV—the recent publicly available database versions—as the primary data source. A total of 140 papers from both databases, including recent research up until February 2024, were selected after analyzing the citation count, methodologies used, and clarity of the work. The selected scientific works span conferences like NeurIPS, IJCAI, ICML, ICLR, AAAI, ACM FAccT, and journals like Nature, JAMA, JAMIA, BMC, PLoS One, etc.

We meticulously adhered to the TRIPOD guidelines, as outlined by Moons et al. in [36]. This adherence was pivotal in guiding our data abstraction, evaluation, and synthesis methodology, specifically focusing on identifying and addressing data bias. The TRIPOD guidelines provided a comprehensive framework that ensured our approach was systematic, methodical, and transparent, facilitating the development of robust and reliable HML risk prediction models that deliver equitable performance to all individuals, irrespective of their inherent or acquired characteristics.

We abstracted information on:

1. The CRD recorded demographic variables like Age, Gender, Ethnicity, Insurance, Marital status and Language,
2. Consideration of demographic variables in the dataset and its selection/rejection criterion based on feature engineering and
3. Algorithm used for the model and its final prediction outcome.

2.1.2 Healthcare Risk Prediction Task

We grouped the selected research works based on the predicted task and observed that ICU mortality is the most extensively studied HML prediction task worldwide (70%). It is closely followed by ICU readmission and ICU length of stay (LOS). The following is a comprehensive list of the everyday HML prediction tasks.

1. Mortality Prediction [37–109] - Predict the likelihood of a patient dying.
   (a) In-hospital and ICU [76, 110] - Estimating the risk of a patient dying during their hospital stay and in ICU.
   (b) Short-term [105, 111] - Assessing the risk of death shortly after ICU admission, typically within 2-3 days.
   (c) *Long-term* [39, 112] - Evaluating the likelihood of death over a longer period, typically from 30 days up to a year after hospital discharge.
2. Length of Stay (LOS) [77, 95, 113–121]
   (a) Predicting the duration of hospital stays for admissions, focusing on stays longer than 3 and 7 days. A custom number of days is also a topic of research interest.
3. Readmission [122, 123]
   (a) Identifying patients at risk of returning to the hospital within 30, 60, 90, 120 days, or custom time frames after discharge.
4. Phenotype Labeling and ICD-9/10 Code Grouping
   (a) *Phenotype Labeling* [124, 125] - Classifying patients into specific groups based on clinical data for disease prediction and treatment customization.

(b) *ICD Code Grouping* [126, 127] - Categorizing diseases based on diagnosis codes to streamline classification.

5. Specific Health Conditions

(a) *Heart failure* [48, 83] - Predicting the occurrence, progression, and prognosis of heart failure in patients, using historical and real-time health data.

(b) *Chronic Kidney Disease (CKD)* [128, 129] - Assessing the risk and progression of CKD to inform treatment plans and manage patient health outcomes.

(c) *Chronic obstructive pulmonary disease (COPD)* [130] - Forecasting COPD exacerbations and identifying patients at higher risk for hospitalization or severe outcomes.

(d) *Coronary artery disease (CAD)* [76, 79] - Utilizing clinical data to predict the development or worsening of CAD for early intervention.

(e) *Sepsis* [128, 131] - Early detection and prediction of sepsis in hospitalized patients to improve treatment response and survival rates.

(f) *Cancer* [132] - Leveraging patient data to predict cancer risks, progression, and treatment responses.

(g) *Ventilation failure* [133] - Predicting the risk of ventilation failure in critically ill patients to guide intervention strategies.

Our survey found that about 67% of HML models trained on the MIMIC CRD came from China [1], with significant contributions from the USA, Brazil, Pakistan, Australia, and India, as shown in Figure 1.1. These findings highlight that ICU mortality prediction is a key area within HML models, making it the main focus of this thesis.

[1] Over 90% of the research is focused on predicting mortality

Chapter 3

Fairness in Healthcare ML

In high-stakes healthcare settings, fairness becomes paramount to ensure that predictive models do not inadvertently perpetuate health disparities or unequal resource distribution among different demographic groups. As such, ensuring fairness in healthcare machine learning is garnering heightened interest, particularly as it stands at the critical juncture of advanced analytics and patient care. Despite its importance, the convergence of fairness-integrated HML modeling remains an area in need of deeper exploration and understanding. Addressing fairness concerns is essential for ethical imperatives, improving health outcomes, and fostering trust in HML applications across diverse populations.

3.1 Problem Formulation

Without losing generality, we only consider the Intensive Care Unit (ICU) mortality prediction task, formulated as a binary classification task in this study. Let the binary label $y_i \in \{0, 1\}$ for the i-th patient; where $y_i = 1$ denotes mortality and $y_i = 0$ survival. We define the sparse irregular time series dataset $D = \{(s_i, X_i, y_i)\}_{i=1}^{N}$ of N observations.

For each patient i, $s_i \in \mathbb{R}^S$ is a sensitive attribute vector like age, gender, insurance, etc. But, based on our analysis results in this study, ethnicity is used. X_i is the multivariate time-series data represented by $X_i = \{(t_j, x_j, v_j)\}_{j=1}^{M}$ for M observations. It is a tuple containing time t_j, event feature $x_j \in X$ representing the physiological/clinical indicator, and its corresponding value $v_j \in \mathbb{R}$. So, in this study, the model $f(\,)$ is trained with the data D to predict $\hat{y}$, given by

$$\hat{y} = f(D) \tag{3.1}$$

STraTS is modeled to predict ICU Mortality using the data D; however, for the time-series LSTM model, we use the same data, but the dataset is formatted as $D^t = f(X_i^t; y_i)g_{i=1}^N$ with the sensitive attribute being a part of the features x_j. In the case of non-time-series/static models, the mean values of the time-series feature over the time t_j are calculated to construct $X_i^s = f(x_j^s; v_j^s)g_{j=1}^M$ of the dataset $D^s = f(X_i^s; y_i)g_{i=1}^N$.

Given these predictions, we then use the fairness metrics defined in section 3.3 to evaluate the models.

3.2 Healthcare ML Fairness Measurements

Before going further, defining some key terminology in fairness research is important. A 'sensitive attribute' is an individual characteristic deemed potentially discriminatory, such as ethnicity, gender, or age. 'Protected groups' are demographics that could face unfair treatment based on these attributes, whereas 'privileged groups' are typically exempt from such bias. Fairness metrics quantitatively evaluate an AI model's impartiality, and bias mitigation strategies are the methods employed to reduce discrimination within these models [134].

Fairness in HML ensures equitable model performance and decision-making across diverse patient groups, avoiding bias based on sensitive attributes like ethnicity, gender, or age [1]. The fairness problem of machine learning methods in healthcare can be grouped into two categories based on differences in the resources allocated [8].

1. Equal allocation - Resources should be distributed proportionally to patients in protected groups.
2. Equal performance - The model is guaranteed to be equally accurate for patients in protected and non-protected groups.

As mentioned above, for each patient i, s_i 2 R^S is a sensitive attribute vector. To identify which sensitive attribute is most closely associated with the target variable, we performed a comprehensive statistical analysis as documented in our datasheet. Chi-square analysis revealed a significant association between ethnicity and ICU mortality, identifying ethnicity s_i^{eth} as the sensitive attribute. This informed our selection of fairness group metrics [9] to

[1] Fairness metrics are evaluated on any available sensitive attributes. However, unfairness might exist because of unrecorded sensitive attributes as well

evaluate disparities across ethnic groups. In the subsequent sections, we omit patient index i for notational simplicity.

Fairness metrics can be evaluated on sensitive attributes like gender, insurance, etc. But, this study uses ethnicity based on our correlation analysis results. The ethnic attribute s^{eth} is categorized into five groups: Asian, Black, Hispanic/Latino, White, and Other. For fairness assessments, we compare pairs of these groups, and the function priv() assigns status for the sub-groups, a for 'privileged' and b for 'protected' based on the ethnic groups being compared. The ground truth is denoted by y, and the model $f()$ prediction is $\hat{y}$.

3.3 Fairness Metrics

Fairness metrics in machine learning serve as quantitative benchmarks to evaluate and ensure that algorithms perform equitably across all user groups, particularly when decisions impact individual's lives. In healthcare, these metrics are vital as they directly influence the quality of patient care and resource allocation. Various fairness metrics exist, each with different implications for model assessment, and based on our analysis results, group fairness metrics are used in this study.

Demographic Parity, which advocates for equal opportunity allocation, stipulates that the probability of a favorable outcome should be independent of the sensitive attribute [135], i.e. the probability of a positive prediction should be equal across different ethnic groups.

$$P(\hat{y} = 1 | \text{priv}(s^{eth}) = a)) = P(\hat{y} = 1 | \text{priv}(s^{eth}) = b)) \tag{3.2}$$

Equalized Odds, a metric that promotes equal performance, requires equal decision rates for privileged and unprivileged groups and is defined as [136], i.e. the model's decision rates for predicting an outcome should be the same across different demographic groups, given the actual outcome.

$$P(\hat{y} = y | \text{priv}(s^{eth}) = a; y) = P(\hat{y} = y | \text{priv}(s^{eth}) = b; y); \quad \forall y \in \{0; 1\} \tag{3.3}$$

Equal Opportunity advocates equal true positive rates across different ethnic groups [21], aiming for fairness in model sensitivity.

$$P(\hat{y} = 1 | \text{priv}(s^{eth}) = a; y = 1) = P(\hat{y} = 1 | \text{priv}(s^{eth}) = b; y = 1) \tag{3.4}$$

Disparate Impact, a group metric that assesses the ratio of favorable outcomes for unprivileged to privileged groups. It evaluates the ratio of positive predictions from one ethnic group to another, with a value of 1 indicating perfect fairness.

$$\frac{P(\text{priv}(s^{eth}) = a; \hat{y} = 1)}{P(\text{priv}(s^{eth}) = b; \hat{y} = 1)} = \frac{P(\text{priv}(s^{eth}) = a)}{P(\text{priv}(s^{eth}) = b)} \tag{3.5}$$

In addition to evaluating their predictive accuracy, we assess the fairness of HML models to measure their trustworthiness. This assessment is crucial because these models are trained on the data sourced from CRD that may contain societal biases. Without fair data assessment, there's a risk that these models could unintentionally perpetuate existing biases.

Chapter 4

Datasheet for Clinical Research Database

Understanding the inherent data and being aware of its inconsistencies during modeling are essential to achieve data-centric fairness. Practitioners may find it challenging to navigate the documentation while concentrating on modeling and fairness due to the breadth of the database [28], as they might only be interested in datasets relevant to their task. So, we present the 'Datasheet for CRD' for MIMIC IV v2.0 as shown in Figure 4.1.

This resource will aid researchers in identifying and addressing data inconsistencies, guide the selection of sensitive attributes essential for fairness assessments, and facilitate the creation of robust, just, and data-conscious fair HML models. More than a mere inventory, the datasheet provides comprehensive insights into the entire database structure, data collection methodologies, management practices, and potential biases.

The datasheet for the MIMIC IV v2.0 provides the following.

1. A thorough overview of the MIMIC IV database, developed by expanding the [30] template to accommodate the intricacies of a complete CRD.
2. A detailed insight into the unique structure of the CRD. We extended the [30] template to incorporate all of the MIMIC IV modules, such as 'Hosp' and 'ICU', MIMIC IV-ED (entire emergency department data), MIMIC IV Notes (physician's notes on patients), and MIMIC-CXR (chest X-ray).
3. Custom queries unique to the CRD. These queries are formulated to provide an in-depth understanding of the data collection, composition, arrangement, task-specific

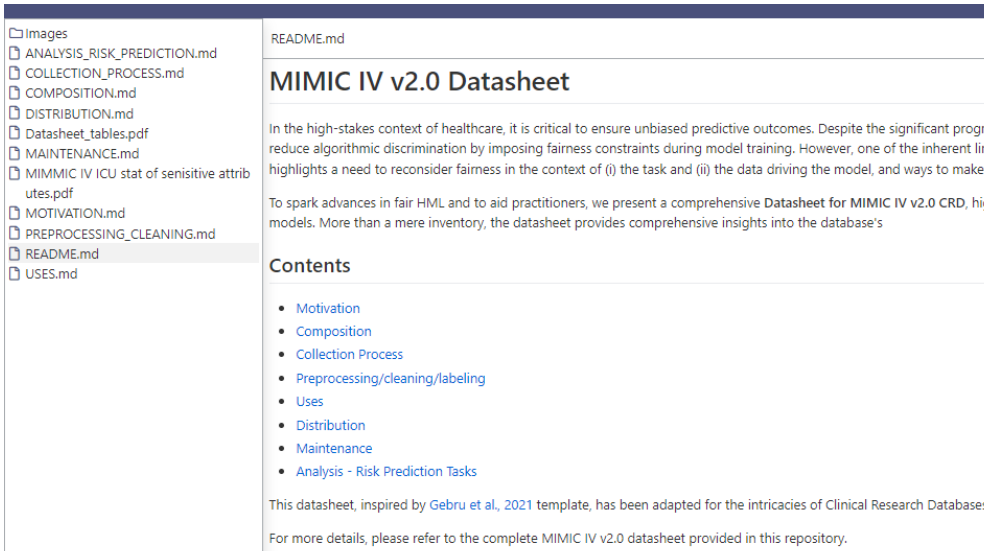


Figure 4.1: Snapshot of our comprehensive datasheet webpage, tailored for the MIMIC IV v2.0 database.

usage, and restrictions unique to the MIMIC IV database and prefixed with '+' in the datasheet. For example, *Can/How the dataset be/are constructed from the MIMIC database?'* is used in place of questions that are unique to the dataset, such as *What is the composition of the dataset?'*.

4. It also features task-specific association analysis between features and the target for a variety of HML prediction tasks such as ICU Mortality, Length of Stay, and Readmission, thus facilitating a deeper fairness evaluation of the models.

4.1 Datasheet for complex Clinical Research Database - MIMIC IV v2.0

4.1.1 Motivation

The questions in this section are primarily intended to clearly articulate the reasons for creating the database and to promote transparency about funding interests.

M1. For what purpose was the database created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The creation of the MIMIC-IV CRD aimed to improve patient care through knowledge discovery and algorithm development using a historically collected medical dataset. It was developed with an approach that allows permissive access, enabling extensive utilization of the MIMIC-IV database. Consequently, the database has been widely utilized in various healthcare applications, including assessing treatment effectiveness in specific patient groups and predicting critical outcomes such as mortality, readmission, and length of stay [28].

M2. Who created the database (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The MIMIC-IV database [28], developed by Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark from the Massachusetts Institute of Technology at the MIT Laboratory for Computational Physiology, is a collaborative effort involving various research groups.

M3. Who funded the creation of the database?

The work was supported by grants from the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health (NIH) under award numbers R01-EB001659 (2003-2013) and R01-EB017205 (2014-2018) [1].

M4. Any other comments?

MIMIC is a large and freely available database that contains deidentified health-related data from patients admitted to the critical care units of the Beth Israel Deaconess Medical Center. There are multiple versions of MIMIC that have been released:

1. MIMIC-IV encompasses data collected from 2008 to 2019, obtained from Metavision bedside monitors [28].
2. MIMIC-III comprises data collected from 2001 to 2012, obtained from both Metavision and CareVue bedside monitors [28].
3. MIMIC-II includes data collected from 2001 to 2008, obtained exclusively from CareVue bedside monitors. While MIMIC-II is no longer publicly available, its data can still be obtained from MIMIC-III by selectively including the data from the CareVue monitors [28].

MIMIC III and MIMIC IV have been extensively utilized throughout the surveyed timeline for healthcare machine learning (HML) prediction models. The datasheet provided is specifically for MIMIC IV v2.0, the latest available version of the database.

[1] https://mimic.mit.edu/

License - The licensing for the MIMIC files can be found in the *PhysioNet Credentialed Health Data License 1.5.0 (MIT-LCP)*[1].

4.1.2 Composition

The questions in this section will highlight the composition of the database intended to provide ML practitioners with the information needed to curate their custom task-specific MIMIC dataset.

C1. What is the composition of the database?

MIMIC IV database is grouped into four modules: MIMIC IV (hosp, and icu), MIMIC IV-ED (ed), MIMIC IV-Note (note), and MIMIC-CXR (cxr)[1].

1. MIMIC IV[28]
   (a) The *Hosp* module grants access to diverse data extracted from the hospital's electronic health record system, including patient and admission details, laboratory measurements, microbiology information, medication administration records, and billed diagnoses. These data are organized in the form of tables, including patient and admission-related tables (patients, admissions, transfers), laboratory measurement tables (labevents, d_labitems), microbiology culture table (microbiologyevents), provider order tables (poe, poe_detail), medication administration tables (emar, emar_detail), medication prescription tables (prescriptions, pharmacy), and hospital billing information tables (diagnoses_icd, d_icd_diagnoses, procedures_icd, d_icd_procedures, services).
   (b) The *ICU* module contains information collected from the clinical information system (BIDMC: MetaVision (iMDSoft)) used within the ICU. Documented data includes intravenous administrations, ventilator settings, and other charted items. Data documented in the icu module includes intravenous and fluid inputs (inputevents), ingredients of the aforementioned inputs (ingredientevents), patient outputs (outputevents), procedures (procedureevents), information documented as a date or time (datetimeevents), and other charted information (chartevents).
2. MIMIC IV-ED[28] - The *ED* module of MIMIC IV-ED focuses on emergency department patients and encompasses information regarding reasons for admission, triage assessments, vital signs, and medication reconciliation. The subject_id and

hadm_id identifiers within MIMIC-IV-ED allow linkage with other MIMIC-IV modules.

3. MIMIC IV-Note [28] - The *Note* module contains deidentified free-text clinical notes for hospitalized patients.

4. MIMIC IV-CXR [28] - The *CXR* module of MIMIC IV-CXR provides lookup tables that establish connections between patient identifiers and MIMIC-CXR study_id and dicom_id, facilitating the analysis of patient chest x-rays in conjunction with clinical data from other MIMIC-IV modules.

+ C2. How is the data arranged within each module, and for what purpose?

The data within each module is structured in tables, as MIMIC is a well-organized relational database. Each table within a module represents a specific type of data. Within each table, the data is organized into rows and columns. Each row corresponds to a particular patient or event, while each column represents a specific variable or attribute corresponding to that row. This organized structure allows researchers to extract customized datasets tailored to their research inquiries efficiently and facilitates the construction of machine learning models.

+ C3. Can the modules be linked to create a specific task dataset?

Yes. The tables within a module can be connected to others within the same module or across different modules using unique identifiers.

+ C4. Explain in detail the tables presented in each module.

MIMIC IV

Hosp Module

Table 4.1: Description of *Hosp* module Tables with detailed information about the features.

Table	Description	Features
omr	The Online Medical Record (OMR) table contains miscellaneous information from the EHR	subject_id, chartdate, seq_num, result_name, result_value

provider	The provider table lists deidentified provider identifiers used in the database	provider_id
admission	Detailed information about hospital stays	subject_id, hadm_id, admittime, dischtime, deathtime, admission_type, admit_provider_id, admission_location, discharge_location, insurance, language, marital_status, race, edregtime, edouttime, hospital_expire_flag
diagnoses_icd	Billed ICD-9/ICD-10 diagnoses for hospitalizations	subject_id, hadm_id, seq_num, icd_code, icd_version
drgcodes	Billed diagnosis-related group (DRG) codes for hospitalizations	subject_id, hadm_id, drg_type, drg_code, description, drg_severity, drg_mortality
emar	The Electronic Medicine Administration Record (eMAR); barcode scanning of medications at the time of administration	subject_id, hadm_id, emar_id, emar_seq, poe_id, pharmacy_id, enter_provider_id, charttime, medication, event_txt, scheduletime, storetime
emar_detail	Supplementary information for electronic administrations recorded in eMAR	subject_id, emar_id, emar_seq, parent_field_ordinal, administration_type, pharmacy_id, barcode_type, reason_for_no_barcode, complete_dose_not_given, dose_due, dose_due_unit, dose_given, dose_given_unit, will_remainder_of_dose_be_given, product_amount_given, product_unit, product_code, product_description, product_description_other, prior_infusi, n_rate, infusion_rate, infusion_rate_adjustment, infusion_rate_adjustment_amount, infusion_rate_unit, route, infusion_complete, completion_interval, new_iv_bag_hung, contin, ed_infusion_in_other_location, restart_interval, side, site, non_formulary_visual_verification

hpcsevents	Billed events occurring during the hospitalization. Includes CPT codes	subject_id, hadm_id, chartdate, hcpcs_cd, seq_num, short_description
labevents	Laboratory measurements sourced from patient-derived specimens	labevent_id, subject_id, hadm_id, specimen_id, itemid, order_provider_id, charttime, storetime, value, valuenum, valueuom, ref_range_lower, ref_range_upper, flag, priority, comments
microbiology events	Microbiology cultures	microevent_id,subject_id, hadm_id, micro_specimen_id, order_provider_id, chartdate, charttime, spec_itemid, spec_type_desc, test_seq, storedate, storetime, test_itemid, test_n, me, org_itemid, org_name, isolate_num, quantity, ab_itemid, ab_name, dilution_text, dilution_comparison, dilution_value, interpretation, comments
patients	Patients' gender, age, and date of death if information exists	subject_id, gender, anchor_age, anchor_year, anchor_year_group, dod
pharmacy	Formulary, dosing, and other information for prescribed medications	subject_id, hadm_id, pharmacy_id, poe_id, starttime, stoptime, medication, proc_type, status, entertime, verifiedtime, route, frequency, disp_sched, infusion_type, sliding_scale, lockout_interval, basal_rate, one_hr_max, doses_per_24_hrs, duration, duration_interval, expiration_value, expiration_unit, expirationdate, dispensation, fill_quantity
poe	Orders made by providers relating to patient care	poe_id, poe_seq, subject_id, hadm_id, ordertime, order_type, order_subtype, transaction_type, discontinue_of_poe_id, discontinued_by_poe_id, order_provider_id, order_status
poe_detail	Supplementary information for orders made by providers in the hospital	poe_id, poe_seq, subject_id, field_name, field_value
prescriptions	Prescribed medications	subject_id, hadm_id, pharmacy_id, poe_id, poe_seq, order_provider_id, starttime, stoptime, drug_type, drug, formulary_drug_cd, gsn, ndc, prod_strength, form_rx, dose_val_rx, dose_unit_rx, form_val_disp, form_unit_disp, doses_per_24_hrs, route

procedures_icd	Billed procedures for patients during their hospital stay	subject_id, hadm_id, seq_num, chartdate, icd_code, icd_version
services	The hospital service(s) that cared for the patient during their hospitalization	subject_id, hadm_id, transfertime, prev_service, curr_service
transfers	Detailed information about patients' unit transfers	subject_id, hadm_id, transfer_id, eventtype, careunit, intime, outtime
d_hcpcs	Dimension table for hpcsevents; provides a description of CPT codes	code, category, long_description, short_description
d_icd_diagnoses	Dimension table for diagnoses_icd; provides a description of ICD-9/ICD-10 billed diagnoses	icd_code, icd_version, long_title
d_icd_procedures	Dimension table for procedures_icd; provides a description of ICD-9/ICD-10 billed procedures	icd_code, icd_version, long_title
d_labitems	Dimension table for labevents provides a description of all lab items	itemid, label, fluid, category

ICU module

Table 4.2: Description of *ICU* module Tables with detailed information about the features.

Table	Description	Features
caregiver	The caregiver table lists deidentified provider identifiers used in the ICU module	caregiver_id
d_items	Dimension table describing itemid. Defines concepts recorded in the events table in the ICU module	itemid, label, abbreviation, linksto, category, unitname, param_type, lownormalvalue, highnormalvalue
chartevents	Charted items occurring during the ICU stay. Contains the majority of information documented in the ICU	subject_id, hadm_id, stay_id, caregiver_id, charttime, storetime, itemid, value, valuenum, valueuom, warning
datetimeevents	Documented information which is in a date format (e.g., date of last dialysis)	subject_id, hadm_id, stay_id, caregiver_id, charttime, storetime, itemid, value, valueuom, warning
icustays	Tracking information for ICU stays including admission and discharge times	subject_id, hadm_id, stay_id, first_careunit, last_careunit, intime, outtime, los

Ingredientevents	Ingredients of continuous or intermittent administrations including nutritional and water content	subject_id, hadm_id, stay_id, caregiver_id, starttime, endtime, storetime, itemid, amount, amountuom, rate, rateuom, orderid, linkorderid, statusdescription, originalamount, originalrate
inputevents	Information documented regarding continuous infusions or intermittent administrations	subject_id, hadm_id, stay_id, caregiver_id, starttime, endtime, storetime, itemid, amount, amountuom, rate, rateuom, orderid, linkorderid, ordercategoryname, secondaryordercategoryname, ordercomponenttypedescription, ordercategorydescription, patientweight, totalamount, totalamountuom, isopenbag, statusdescription, originalamount, originalrate
outputevents	Information regarding patient outputs including urine, drainage, and so on	subject_id, hadm_id, stay_id, caregiver_id, charttime, storetime, itemid, value, valueuom
procedureevent	Procedures documented during the ICU stay (e.g., ventilation), though not necessarily conducted within the ICU (e.g., x-ray imaging)	subject_id, hadm_id, stay_id, caregiver_id, starttime, endtime, storetime, itemid, value, valueuom, location, locationcategory, orderid, linkorderid, ordercategoryname, ordercategorydescription, patientweight, isopenbag, continueinnextdept, statusdescription, originalamount, originalrate

MIMIC IV-ED

Table 4.3: Description of *ED* module Tables with detailed information about the features.

Table	Description	Features
diagnosis	The diagnosis table provides billed diagnoses for patients. Diagnoses are determined after discharge from the emergency department	subject_id, stay_id, seq_num, icd_code, icd_version, icd_title
edstays	The edstays table is the primary tracking table for emergency department visits. It provides the time the patient entered the emergency department and the time they left the emergency department	subject_id, hadm_id, stay_id, intime, outtime, gender, race, arrival_transport, disposition

medrecon	On admission to the emergency departments, sta will ask the patient what current medications they are taking. This process is called medicine reconciliation, and the medrecon table stores the findings of the care providers	subject_id, stay_id, charttime, name, gsn, ndc, etc_rn, etccode, etcdescription
pyxis	The pyxis table provides information for medicine dispensations made via the Pyxis system	subject_id, stay_id, charttime, med_rn, name, gsn_rn, gsn
triage	The triage table contains information about the patient when they were first triaged in the emergency department	subject_id, stay_id, temperature, heartrate, resprate, o2sat, sbp, dbp, pain, acuity, chiefcomplaint
vitalsign	Patients admitted to the emergency department have routine vital signs taken every 1-4 hours. These vital signs are stored in the vitalsign table	subject_id, stay_id, charttime, temperature, heartrate, resprate, o2sat, sbp, dbp, rhythm, pain

MIMIC IV-CXR

Table 4.4: Description of *CXR* module Tables with detailed information about the features.

Table	Description	Features
cxr_record_list	Lists all records in the MIMIC-CXR database	subject_id, study_id, dicom_id

MIMIC IV-Note

Table 4.5: Description of *Note* module Tables with detailed information about the features.

Table	Description	Features
discharge	Discharge summaries for hospitalizations	note_id, subject_id, hadm_id, note_type, note_seq, charttime, storetime, text
discharge_detail	Auxiliary information for discharge summaries	note_id, subject_id, field_name, field_value, field_ordinal
radiology	Radiology report	note_id, subject_id, hadm_id, note_type, note_seq, charttime, storetime, text
radiology_detail	Auxiliary information for radiology notes	note_id, subject_id, field_name, field_value, field_ordinal
cxr_record_list	Lists all records in the MIMIC-CXR database	subject_id, study_id, dicom_id

+ C5. Can/How can the dataset be/are created from the MIMIC database?

The MIMIC database is a comprehensive clinical research database that encompasses various types of data, such as patient admissions, ICU records, triage information, bedside health records, X-rays, and clinician medical notes. It offers researchers the flexibility to create custom datasets tailored to their specific research tasks.

For example, if the objective is to predict heart failure, relevant cohorts related to heart failure can be extracted from tables like admission, patient, diagnoses_icd, and d_icd_diagnoses in the hosp module. Additional features associated with heart failure

can be obtained by linking tables from the ICU module and ED module. Once the cohort and their corresponding heart-related features are extracted, they undergo pre-processing and cleaning before being represented in either a time series or non-time series format, depending on the prediction task. This allows for the creation of suitable datasets for predictive modeling. Similarly, researchers can curate a wide range of task-specific datasets based on their needs.

C6. What do the dataset's instances represent (e.g., documents, photos, people, countries)?

A dataset derived from the MIMIC contains patient health data. The data can be patient demography (Age, gender, ethnicity, language, etc.), ICU details, X-ray images, or even Clinician notes. It differs depending on the intended prediction task.

C7. How many instances are there in total (of each type, if appropriate)?

Dataset is extracted from the MIMIC database based on the intended task, and the count of instances depends on the dataset extracted.

For instance, If we intend to create a MIMIC IV ED dataset by linking ED, hosp, and ICU modules, then the dataset will have 425087 instances. Similarly, several complex datasets can be created, and the instance of the dataset varies depending on the prediction task/requirements.

C8. Does the dataset/database contain all possible instances, or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances because instances were withheld or unavailable).

The MIMIC-IV database is a subset of deidentified electronic health records (EHRs) obtained from patients admitted to BIDMC between 2008 and 2019. This curated collection has undergone validation and quality assurance by a team of interdisciplinary experts. The database includes diverse patients and diagnoses, making it suitable for various research purposes. However, it is important to acknowledge that the dataset is not comprehensive, as it represents a subset of the overall patient population. Researchers should be mindful of potential biases inherent in the dataset and employ appropriate methods to address them when conducting analyses or studies[1].

C9. What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

MIMIC IV (hosp and ICU module) and MIMIC-IV-ED (ED module) consist of raw unprocessed text, Date time, and number data in the comma-separated format of the patients admitted to the hospital, ICU, and ED. MIMIC-CXR and MIMIC-Note contain images of chest X-rays and free-text clinical notes for hospitalized patients, respectively. Table ?? provides detailed feature information of the data.

C10. Is there a label or target associated with each instance? If so, please provide a description.

The choice of target variable in the MIMIC dataset depends on the specific prediction task. For example, if the goal is to predict the length of stay in the ICU, the *los* attribute in the *icustay* table can serve as the target variable. On the other hand, if the objective is to predict *in-hospital mortality*, the *hospital_expire_flag* in the *admissions* table can be used as the target variable. The target variable selection is contingent upon the specific prediction task being undertaken.

C11. Are there recommended data splits (e.g., training, development/validation, testing)?

No

C12. Are there any errors, sources of noise, or redundancies in the database? If so, please provide a description.

Our analysis of the MIMIC IV dataset has revealed several biases and inconsistencies that researchers should be aware of,

1. Inconsistencies in patient details:Patient language is inconsistently recorded, with only English being specified while other languages are marked as '?' or unknown.
2. Inconsistencies in in-hospital expiry information: The admission table contains multiple reports of the same patient's death, leading to inconsistencies.
3. Vagueness in insurance coverage information: The dataset lacks definitive information about insurance coverage, limiting researchers' ability to draw conclusions on insurance choices.
4. Inconsistencies in hospital admit and discharge timestamps: The admission table exhibits inconsistencies in the recorded timestamps and missing values for death time. These might be data entry errors since there are instances where the ICU in time occurs before the hospital admission and ICU out time occurs after the hospital discharge time.

5. **Potential representation bias in the dataset**: Gender attribute has only Male and Female data recorded, and the Language attribute only has English-speaking patient's details. The database owners acknowledge the potential for bias, particularly since the data is derived from a single hospital system and may not be representative of the entire population.

The data [28] in the database is collected during routine clinical practice, reflecting the specific practices of the hospital. It is important to note that the database may have implausible values due to the archival process[1]. Therefore, caution should be exercised when using the data, and researchers should be mindful of the dataset limitations and potential biases.

We strongly recommend that researchers adhere to best practice guidelines [137] when analyzing the data.

C13. Does the database contain data that might be considered confidential (e.g., data protected by legal privilege or doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?

Yes, the MIMIC IV dataset includes medical records of patients, encompassing confidential personal and health-related information. However, the dataset is constructed with patient privacy as a priority, and all data within the database undergoes de-identification processes to comply with Health Insurance Portability and Accountability Act (HIPAA) regulations.

C14. Does the database identify subpopulations (e.g., by age or gender)?

Yes. Databases (specifically admission and patient tables) have patient demographic data such as age, gender, ethnicity, language, insurance, and marital status.

MIMIC IV Distribution statistics

C15. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the database? If so, please describe how.

No, all data in the database is de-identified by HIPAA regulations.

C16. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

Table 4.6: Admission distribution statistics.

Description	Value
Total records	180,733
Male	47%
Female	53%
Min Age	18
Max Age	91
Predominant Ethnicity	White (67.2%)

Table 4.7: Patient distribution statistics.

Description	Value
Total records	299,712
Male	47%
Female	53%
Min Age	18
Max Age	91

Table 4.8: ED table distribution statistics.

Description	Value
Total records	299712
Male	46%
Female	54%
Predominant Ethnicity	White (58%)
Predominant Disposition	Home

Yes, the database recorded demographic information like ethnicity, gender, age, marital status, language, and insurance.

+ C17. Do researchers have to take any important measures to handle the data with care?

To ensure patient privacy, researchers are required to comply with data usage agreements mandated in [28, 137] and obtain the necessary approvals and certifications before accessing the dataset. Researchers working with healthcare-related data are responsible for handling the data carefully and ethically, taking measures to prevent any potential harm or dissatisfaction. While the data is de-identified by HIPAA regulations, it is crucial to treat the data with respect and caution, following best practices. Additionally, the Institutional Review Board of the Beth Israel Deaconess Medical Center approved the collection of patient information and the creation of the research resource.

4.1.3 Collection Process

The question in this section provides a clear perspective of how the data is collected in MIMIC IV. This highlights potential data collection bias the researchers can be wary of while modeling.

CP1. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If subjects reported the data or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was extracted from the hospital databases of the Beth Israel Deaconess Medical Center (BIDMC) specifically for patients admitted to the intensive care units. A comprehensive patient list was compiled, including all medical record numbers associated with ICU or emergency department admissions from 2008 to 2019. To ensure the reliability of the database, a multidisciplinary team of scientists and clinicians thoroughly evaluated MIMIC-IV during its development, conducting code reviews and documenting any identified issues using a ticket system [28].

CP2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

MIMIC-IV is derived from two distinct database systems within the hospital setting: a customized *electronic health record (EHR)* used across the entire hospital and a special-

ized clinical information system called *MetaVision (iMDSoft)* specifically designed for the intensive care units at the Beth Israel Deaconess Medical Center (BIDMC).

To ensure the accuracy and reliability of the MIMIC-IV dataset, a diverse team of scientists and clinicians conducted a comprehensive evaluation during its development, which included code reviews and the systematic documentation of identified issues using a ticket system [28].

CP3. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., the recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Over *11 years, from 2008 - 2019*.

CP4. Were any ethical review processes conducted (e.g., by an institutional review board)?

Yes, the Institutional Review Board reviewed the collection of patient information and the creation of the research resource at the Beth Israel Deaconess Medical Center.

CP5. Did you collect the data directly from the individuals in question or obtain it via third parties or other sources (e.g., websites)?

Data is collected from hospital EHR and ICU-specific clinical information systems at the BIDMC called *CareVue and MetaVision (iMDSoft)*.

CP6. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please describe this analysis, including the outcomes and a link or other access point to any supporting documentation.

Unknown, however, the MIMIC data is deidentified[1], and patient identifiers were removed according to the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision [28].

4.1.4 Preprocessing/cleaning/labeling

Questions under this section detail the preprocessing steps the CRD owners took. Clear information about the data preprocessing steps will aid the HML practitioner during the duration of their task-specific, custom dataset.

P1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances,

processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

The data within the MIMIC-IV database underwent a reorganization to enhance its suitability for retrospective data analysis [28]. This involved denormalizing tables, eliminating audit trails, and consolidating the data into smaller tables. The primary objective of this process was to simplify the retrospective analysis of the database. Notably, no data cleaning procedures were applied to ensure that the dataset accurately represents real-world clinical data[1].

To protect patient privacy, patient identifiers were removed in compliance with HIPAA regulations. Random ciphers were used to replace patient identifiers, resulting in deidentified integer values for patients, hospitalizations, and ICU stays. Structured data underwent filtering using look-up tables and allow lists. Additionally, dates and times were randomly shifted into the future by specific days. Consequently, the data for each patient remains internally consistent [28].

P2. Is the software used to preprocess/clean/label the data available? If so, please provide a link or other access point.

Unknown. However, authors have stated that a free-text deidentification algorithm was used to remove personally identifiable information (PHI) from the free-text data if needed.

4.1.5 Uses

This section highlights the use case scenario of the database. By explicitly knowing the use case of the data, the researchers can make informed decisions, thereby avoiding any potential risks.

U1. Has the database/dataset been used for any tasks already? If so, please provide a description.

Yes, the MIMIC database is one of the most widely used CRD. It has been widely used for the below types of works,

1. Prediction tasks like,
   (a) Readmission [122, 123, 138–140] (30, 60, 90, 120 and custom days) - Predict patients at risk of readmission early in the health care process(helps to prioritize care towards such patients preventing mortality and readmission).
   (b) Mortality [37–39, 41, 43] - Predict the likelihood of patients dying.

i. In-hospital [76, 110, 141, 142] - Predict the likelihood of a patient dying in hospital while they are admitted (helpful to identify high-risk patients early on to provide medical interventions).
ii. Short term [42, 105, 111, 143, 144] - Predict short-term mortality (typically within 2-3 days) after ICU admission.
iii. Long term [37–39, 112] - Predict long-term mortality (typically within 30 days to 1 year) after hospital discharge.

(c) Length of stay (LOS) [77, 95, 113, 114] - Predict the length of stay of each admission(typically predicting $>$ 3 and 7 days stay. Custom days is also being predicted).

(d) Phenotype label and ICD-9/10 code grouping - Helpful in tasks like disease prediction, outcome analysis, treatment recommendation, and customized treatments.
i. Phenotype labeling [124, 125, 145, 146]classify patients into specific groups based on their diagnoses, procedures, medications, and other clinical variables.
ii. Grouping ICD 9/10 codes [126, 127, 147]into different categories based on patient diagnosis to classify the disease.

2. Prediction for specific health ailments like,
(a) Heart failure [48, 83]
(b) Chronic Kidney Disease (CKD) [128, 129]
(c) Chronic obstructive pulmonary disease (COPD) [130, 138]
(d) Coronary artery disease (CAD) [76, 79]
(e) Sepsis [128, 131]
(f) Cancer [132, 148]
(g) Ventilation failure [133, 149]

U2. Is there a repository that links to any or all papers or systems that use the database/dataset? If so, please provide a link or other access point.

No, however, the owners [28] have provided the repository, where the code and other discussions related to the database are hosted.

U3. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there

anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

The data available in the database reflects the idiosyncrasies of routine clinical practice, as stated by the owners. The archival process may have introduced implausible values and potential bias into the data. Therefore, researchers need to follow best practice guidelines when using the data for analysis or other purposes [28].

U4. Are there tasks for which the dataset should not be used? If so, please provide a description.

Unknown, the owners of the database did not provide clear information in the documentation.

4.1.6 Distribution

This section provides an overview of how data is distributed within the database and how it can be distributed to the public.

+ D1. Is the data publicly available? How and where can it be accessed (e.g., website, GitHub)?

Yes. The MIMIC-IV data is accessible to the public through the PhysioNet[2]. To gain access, individuals need to become PhysioNet-certified users and agree to the data use agreement. Once granted access, users can download the complete set of files or select specific subsets that align with their requirements.

D2. Will the dataset be distributed under a copyright or other intellectual property (IP) license and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU and provide a link or access point to, or otherwise reproduce, any relevant licensing terms or ToU and any fees associated with these restrictions.

Access to the MIMIC-IV data is granted through a license agreement called the *Data Use Agreement (DUA)*, which outlines the terms and conditions for data usage. To obtain access, users are required to complete an online course on the ethical use of human subject's research data and obtain a certificate of completion. Users can apply for dataset access through the PhysioNet [2] with the certificate. The application process involves agreeing to the DUA terms and providing details about the intended use of the data.

[2] https://physionet.org/content/mimiciv/2.2/

4.1.7 Maintenance

This section highlights how the database is maintained and how frequently it's being updated. This gives users a clear picture of how to keep up with the different versions of the data.

MA1. Is the database maintained? Who will be supporting/hosting/maintaining the database?

Yes, MIMIC-IV is maintained by the Laboratory for Computational Physiology at the Massachusetts Institute of Technology (MIT) and BIDMC. They provide ongoing support and maintenance for the database.

MA2. How can the database's owner/curator/manager be contacted (e.g., email address)?

For private issues, they can be contacted at `mimic-support@physionet.org`, and for issues related to patient health information (PHI), `phi-report@physionet.org` is being used[1].

MA3 Will the database be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

Yes, the MIT Laboratory regularly updates the MIMIC-IV database for the Computational Physiology team. The latest version, v2.2, has been released, which includes updates from the previous version, v1.0. The frequency of future updates is unknown, but any information regarding updates can be found on the official website and GitHub.

MA4. Will older versions of the database continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

Previous versions of the database will continue to be supported and maintained. However, it is not explicitly stated whether they might have any further updates by the owners.

MA5. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description.

Content for the MIMIC website and documentation is hosted publicly on GitHub. To raise a problem or to suggest an improvement, a new issue can be created. Users can also take part in the discussion channel.

+ M6. What is the Data life cycle of the MIMIC database?

1. Data Acquisition

2. Data Archive
3. Data Preparation
4. Data loading

constitutes the life cycle of MIMIC CRD.

Data Acquisition.

Data is collected from the source, which may be internal, external, or both.

Table 4.9: Data Acquisition

Internal (In-hospital) data	External data source
ICU/MICU/SICU/CCU /CVICU/NICU data (vitals, trends, anomalies) Chart details (Fluids, medications, etc.) Demographics (age, gender, ethnicity, language, marital status, religion, insurance, etc.) Lab reports Billing details Physician notes Provider order entries, etc.	Social Security Death Index, etc.

Data Archive Data collected from the source is Archived before proceeding with data preparation for later use.

Data Preparation

To ensure compliance with HIPAA regulations, deidentification, date shifts, and format conversions are applied to the archival data. The data is then reorganized into a more suitable format for retrospective analysis, which involves consolidating tables, denormalizing data, and removing audit trails. It's important to note that no data cleaning procedures were performed to maintain the authenticity of the real-world clinical dataset. Feedback from users will be considered for further iterations, and the final version of the data will be loaded into the database.

Data loading to database

The final version of the data is then loaded to the database built on a PostgreSQL relational database management system and hosted on a secure server infrastructure. The data can be downloaded locally or accessed on the cloud via BigQuery, AWS, or GCS.

4.2 How Datasheet can be used for HML Modelling

For any machine learning (ML) model, especially in healthcare, the integrity and bias-free nature of the data are critical. The composition section of the Datasheet meticulously documents the MIMIC IV Clinical Research Database (CRD) and identifies potential biases present. It guides ML practitioners in refining their datasets for specific healthcare tasks—emergency department readmission, ICU mortality, or length of hospital stay—while being mindful of existing data inconsistencies.

Following dataset preparation, the next step involves training the model for the chosen task. Post-training, rigorous evaluation using relevant metrics is necessary to assess the model's efficacy. In Healthcare Machine Learning (HML), scrutinizing fairness alongside performance is imperative. The analysis section of the Datasheet offers insights into selecting sensitive attributes for the analysis and aids in choosing appropriate fairness metrics. This Datasheet will be a comprehensive resource for practitioners working with MIMIC data and will foster the development of equitable HML models.

Chapter 5

Role of Datasheet for Database in modeling - ICU Mortality prediction task

To explore the intricate link between clinical data and fairness in healthcare prediction tasks, we benchmarked static baseline models such as Logistic Regression (LR) and XG Boost, as well as the time-series LSTM model, against the SOTA model for ICU mortality.

5.1 MIMIC Database

The datasheet provides detailed information on the motivation, collection, composition, and usage of MIMIC CRD, while Table 5.1 summarizes its general demographic characteristics for ICU Mortality. This study analyzes the latest publicly available MIMIC-IV v2.0 for its up-to-date and comprehensive information, despite both MIMIC-III and MIMIC-IV being popular in HML research.

Sensitive attribute statistics of MIMIC IV ICU Mortality data: We followed the work of Meng et al. [20] and grouped Ethnic demography into 5 categories based on the geographic origin. Data contains 68% of the White population followed by the Other (13.8%) subgroup. Age is categorized into 6 buckets 5.1 following the work of Röösli et al. [18]. 24% are of 30 to 49 and 50 to 69 age bins followed by 70 to 80 (23.6%). Children below 17 are not part of this CRD. Gender-wise, only Males and Females are recorded, while 56% of the patients are Males. Medicaid, Medicare, and Other were the 3 types of

Table 5.1: Breakdown of Sensitive Attributes in MIMIC IV v2.0 ICU Mortality dataset: Distribution of patient demographics across gender, age, ethnicity, language, and insurance type, highlighting the diversity and potential biases inherent in the database.

Gender			Age			Ethnicity			Language			Insurance		
Patient	Count	%	Range	Count	%	Group	Count	%	Type	Count	%	Type	Count	%
Female	25671	44.3	0-17	0	0	Asian	1701	2.9	English	52077	89.8	Other	27262	47
Male	32328	55.7	18-29	6315	10.9	Black	6565	11.3	?	5922	10.2	Medicare	26385	45.5
			30-49	13868	23.9	Hispanic/Latino	2228	3.8				Medicaid	4352	7.5
			50-69	13702	23.6	White	39522	68.1						
			70-89	11264	19.4	Other	7983	13.8						
			90+	12859	22.2									

Insurance listed, with English being the only language recorded, whereas others are left as '?'.

There are several inconsistencies like in-hospital expiry information [1] and hospital admit and discharge timestamps. Without careful consideration, there might be a risk of inadvertently overlooking potential target-feature associations or data inconsistencies which can result in models that perpetuate bias [18].

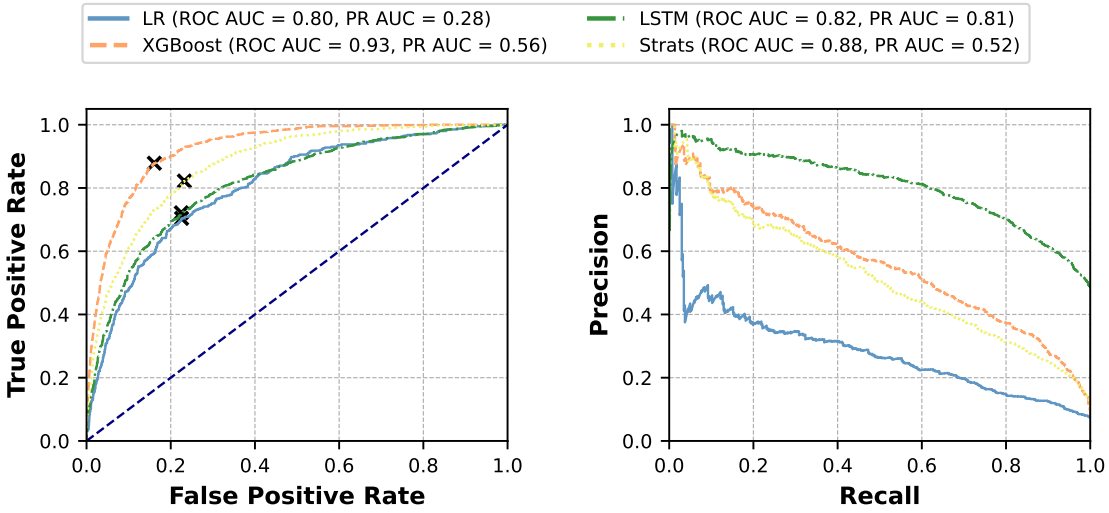
5.2 Experimental setup

To examine how fair the predictions of SOTA ICU mortality models are, we conducted a series of experiments against several baselines on the MIMIC IV dataset. We adopted an 80:20 split for the training and test sets. The models were trained for up to 1000 epochs until the validation loss stopped improving for 10 continuous epochs, applying a 10-fold cross-validation for 3 different random seeds. The target of the model is to predict the probability of mortality ($\not{0}$) following ICU admission of the patient. The experiments were conducted on NVIDIA GeForce RTX 2080 Ti GPU and the entire code of the implementation are available on this GitHub page.

5.3 Experimental Results

In this study, we rigorously evaluated the performance and fairness of HML models, specifically focusing on ICU mortality prediction. Our evaluation comprised widely recognized

[1] The admission table contains multiple reports of the same patient's death, leading to inconsistencies. However, owners of MIMIC have mentioned the potential of bias within the CRD



(a) ROC Curve (b) PR Curve

Figure 5.1: Prediction performance analysis across models. Panels (a) and (b) show the ROC-AUC and PR-AUC of the models, respectively, and the operating points in (a).

baselines, including Logistic Regression (LR), XG Boost, LSTM, and the SOTA model described by Tipirneni et al. in [29].

5.3.1 Model Performance Evaluation

We employed the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) and Precision-Recall Area Under the Curve (PR-AUC) metrics for a comprehensive comparison of model performance. These comparisons, derived from 3 Monte Carlo simulations to ensure statistical robustness, are visually represented in Figure 5.1. Surprisingly, the XG Boost model demonstrated a significant performance uplift, outperforming the SOTA model by 6.44% in terms of ROC-AUC, as detailed in Table 5.2.

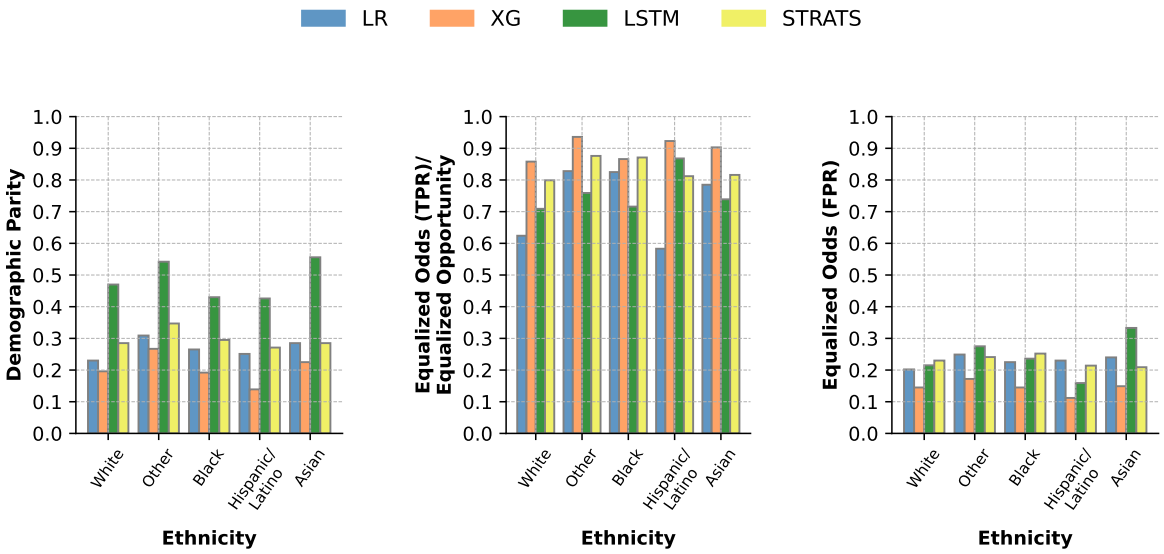
5.3.2 Model Fairness Assessment

Further, to scrutinize the ethical aspect of the model application, we analyzed each model using established fairness metrics. This assessment aimed to identify any persistent biases or disparities in prediction accuracy across different ethnic groups. The results, summarized in Table 5.3, revealed notable variations in model fairness. Figures 5.2(b) and 5.2(c)

Table 5.2: Comparative Analysis of ICU Mortality Prediction: Predictive performance of models trained on MIMIC IV v2.0, contrasting static and time-series models as reflected by the ROC-AUC and PR-AUC curves for 3 Monte Carlo runs.

Type	Model	ROC-AUC		PR-AUC	
Static	LR	0.805	0.012	0.276	0.005
	XG Boost	**0.926**	**0.006**	0.551	0.007
Time series	LSTM	0.886	0.003	**0.807**	**0.011**
	STraTS (SOTA)	0.870	0.002	0.520	0.006

elucidate how XG Boost consistently surpassed the SOTA model in both Equalized Opportunity (EOp) and Equalized Odds (EO) metrics, highlighting its superior fairness profile. Additionally, Figure 5.3 illustrates the disparities in Demographic Parity (DP) across ethnic categories, notably where the SOTA model still manifests considerable bias despite achieving higher DP scores than XG Boost.



(a) Demographic Parity (b) Equalized Opportunity (c) Equalized Odds

Figure 5.2: Analysis of fairness metrics for each model across ethnic subgroups. Panels (a), (b), and (c) show the comparative analysis for demographic parity (DP), equalized odds for TPR (also known as EOp), and FPR, respectively. Inconsistency drawn from these metrics reveals the extent of fairness exhibited by each model, highlighting the interplay between algorithmic performance and demographic impact.

These findings underscore the intricate challenges of embedding fairness into HML models. The observed discrepancies across models, especially in the context of fairness metrics,

Table 5.3: Evaluation of Fairness Metrics (Demographic Parity (DP), Equalized Odds for True Positive Rate (TPR)/Equalized Opportunity (EOp), and Equalized Odds for False Positive Rate (FPR)) Across Model Types: This table presents a comparative analysis of fairness metrics for di□erent models, stratified by ethnic groups.

Type	**Models**	**Ethnicity**	**DP** $	**EO(TPR)/ EOp** " $	**EO(FPR)** # $
Static	Logistic Regression	White	0.230	0.624	0.202
		Other	0.309	0.828	0.249
		Black	0.265	0.825	0.225
		Hispanic/Latino	0.251	0.583	0.230
		Asian	0.285	0.785	0.240
	XG Boost	White	0.196	0.858	0.145
		Other	**0.267**	**0.936**	**0.171**
		Black	0.192	0.866	0.145
		Hispanic/Latino	0.139	0.923	0.112
		Asian	0.225	0.903	0.149
Time Series	LSTM	White	0.460	0.709	0.215
		Other	0.534	0.759	0.275
		Black	0.421	0.716	0.236
		Hispanic/Latino	0.426	0.868	0.159
		Asian	0.545	0.739	0.333
	STraTS (SOTA)	White	0.289	0.799	0.230
		Other	**0.352**	**0.876**	**0.241**
		Black	0.304	0.871	0.252
		Hispanic/Latino	0.271	0.812	0.214
		Asian	0.299	0.816	0.209

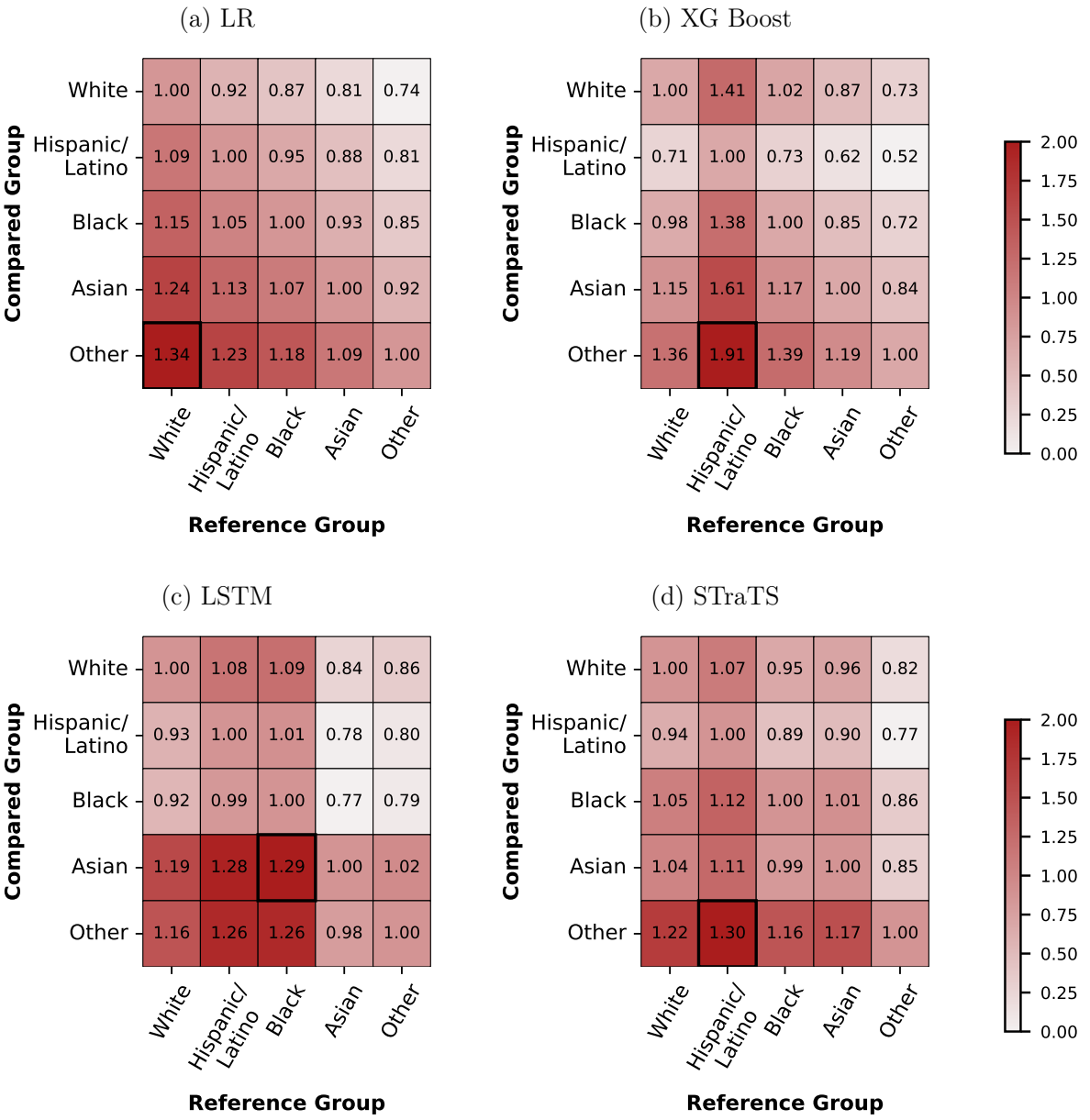


Figure 5.3: Heatmaps Illustrating disparate model impact across Ethnic subgroups Groups. Red < 1.0 indicates the disparity of the model against the compared group, Red > 1.0 reflects the model's disparity against the reference group and the value of 1.0 shows Parity against the comparison. Highlighted cells represent the most disparate subgroup comparison.

highlight the necessity of looking at model development from a fairness perspective. It is imperative to focus closely on the data we use, how we use it, and the inherent biases/CRD data inconsistencies while building models. The 'Datasheet for database' resource will serve as an essential tool for all the ML practitioners working with CRD to develop equitable HML models. Doing this isn't just about making models more accurate; it's about making sure everyone gets fair treatment from these HML systems.

In light of our results, we advocate for an integrated strategy that considers both the technical and ethical aspects of HML modeling. This strategy should encompass rigorous data analysis, conscious model selection, and continuous fairness evaluation to mitigate bias and promote equity in healthcare analytics.

Chapter 6

Discussions and Conclusion

6.1 Discussion

The field of fair HML actively explores ways to prevent model discrimination. Given that, the data fed into models, particularly in healthcare, can reflect real-world biases, making it vital to analyze fairness in consideration of the data while modeling [15]. So, it is essential to understand the inherent biases in the data. MIMIC is the go-to data source for HML models and is used across the globe. Meng et al. work [20] on MIMIC states the importance of demographic features for model prediction and highlights fairness concerns. Given that, our analysis reveals that SOTA HML models are currently not reporting fairness metrics. This is concerning since downstream application of these models can significantly disadvantage subpopulations. On the other hand, practitioners may find it difficult to navigate large databases such as MIMIC [28], which makes it harder to effectively track a model's fairness properties.

To initiate a shift towards data-centric fairness, we introduce 'Datasheet for CRD' for MIMIC IV v2.0, a comprehensive resource custom-designed for CRD. A resource crafted to assist practitioners in identifying data anomalies and evaluating feature-target relationships for fairness assessments thus facilitating the development of robust, data-informed, and equitable HML models. It serves as a blueprint for researchers to analyze (i) real-world data inconsistency and (ii) task-specific feature-target association essential for fairness evaluations.

Synthetic Data Generation Enhanced by Datasheets:

As we advance toward synthetic data generation—a necessary step in light of the difficulties in gathering diverse and all-encompassing health data—the 'Datasheet for CRD'

becomes an indispensable instrument. It is imperative to recognize that synthetic data, derived solely from CRDs like MIMIC, does not completely capture the global patient population as addressed in1. Combining these synthetic datasets with real patient data collected from diverse geographical locations is necessary to ensure global representation. By using the datasheet to identify and address data gaps in addition to it, we facilitate the creation of enriched synthetic datasets that more accurately mirror the global patient population, advancing the scope and impact of HML research.

Enhancing Generative AI in Healthcare with Datasheets: Datasheets for Clinical Research Databases like MIMIC IV are vital tools for enhancing Generative AI, particularly Large Language Models (LLMs) used in healthcare. They offer a way to refine these models, ensuring the medical knowledge they generate is free from the inherent data biases present in their training datasets. This fusion of LLM's capabilities with the bias-aware data from datasheets helps develop fair and generalizable AI.

Enhancing Trustworthy HML predictive modeling: The 'Datasheet for CRD' coupled with a comprehensive Model Card [150] offers a robust framework for integrating the expertise of healthcare professionals directly into the AI development process. Encapsulating clinician's nuanced insights and observations within the datasheet and detailing the operational aspects and performance/fairness metrics in the Model Card enhances the ability to create models that resonate with the realities of clinical practice. This human-in-the-loop methodology ensures patient and practitioner trust in evolving HML models, making AI both cutting-edge and deeply trusted.

Future Work: Future work will include developing datasheets for other popular databases such as eICU [151] and HiRID [152] as this study focuses solely on model fairness concerning ethnic attributes. However, it does not explore other correlated variables, such as socio-economic status, access to healthcare, pre-existing health conditions, and treatment quality, which can also impact the prediction outcome. Future work could include obtaining and analyzing these details from CRD and developing a datasheet for those to offer a more comprehensive understanding of the factors influencing model fairness.

6.2 Conclusion

Our study highlights the critical need for a renewed focus on fairness in HML models. By demonstrating how simple baselines outperform SOTA HML models on MIMIC, we reveal the complexities involved in achieving fairness and the need for the community to move towards reporting fairness metrics in HML by default. Moreover, for successful real-world

use of HML, there is a need to adopt a data-centric approach to fairness, which entails a thorough examination of the data, its contextual use, stakeholder discussions, and potential model biases.

References

[1] G. Montcho et al. “Population Aging and Work Life Duration in Canada”. In: *Canadian Public Policy* 49.S1 (2023), pp. 32–47.

[2] K. Hayes et al. “Factors influencing the mental health consequences of climate change in Canada”. In: *International journal of environmental research and public health* 16.9 (2019), p. 1583.

[3] D. Duong et al. *Overworked health workers are “past the point of exhaustion”*. 2023.

[4] *InfoGraphic*. https://www.cihi.ca/en/infographic-canadas-seniors-population-outlook-uncharted-territory. [Accessed 01-04-2024].

[5] K. Belsher. “From shortage to solution: A study of nursing retention policies in British Columbia”. In: (2023).

[6] A. Rajkomar et al. “Machine learning in medicine”. In: *New England Journal of Medicine* 380.14 (2019), pp. 1347–1358.

[7] E. J. Topol. “High-performance medicine: the convergence of human and artificial intelligence”. In: *Nature medicine* 25.1 (2019), pp. 44–56.

[8] A. Rajkomar et al. “Ensuring fairness in machine learning to advance health equity”. In: *Annals of internal medicine* 169.12 (2018), pp. 866–872.

[9] Q. Feng et al. “Fair machine learning in healthcare: A review”. In: *arXiv preprint arXiv:2206.14397* (2022).

[10] J. Buolamwini et al. “Gender shades: Intersectional accuracy disparities in commercial gender classification”. In: *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 77–91.

[11] S. Corbett-Davies et al. “The measure and mismeasure of fairness: A critical review of fair machine learning”. In: *arXiv preprint arXiv:1808.00023* (2018).

[12] N. Mehrabi et al. “A survey on bias and fairness in machine learning”. In: *ACM computing surveys (CSUR)* 54.6 (2021), pp. 1–35.

[13] F. Li et al. “Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction”. In: *Journal of Biomedical Informatics* 138 (2023), p. 104294.

[14] J. Zou et al. “Design AI so that it’s fair”. In: *Nature* 559.7714 (2018), pp. 324–326.

[15] I. Chen et al. “Why is my classifier discriminatory?” In: *Advances in neural information processing systems* 31 (2018).

[16] I. Y. Chen et al. “Can AI help reduce disparities in general medical and mental health care?” In: *AMA journal of ethics* 21.2 (2019), pp. 167–179.

[17] E. Röösli et al. “Bias at warp speed: how AI may contribute to the disparities gap in the time of COVID-19”. In: *Journal of the American Medical Informatics Association* 28.1 (2021), pp. 190–192.

[18] E. Röösli et al. “Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model”. In: *Scientific Data* 9.1 (2022), p. 24.

[19] B. Hsu et al. “Pushing the limits of fairness impossibility: Who’s the fairest of them all?” In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 32749–32761.

[20] C. Meng et al. “Mimic-if: Interpretability and fairness evaluation of deep learning models on mimic-iv dataset”. In: *arXiv preprint arXiv:2102.06761* (2021).

[21] M. B. Zafar et al. “Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment”. In: *Proceedings of the 26th international conference on world wide web*. 2017, pp. 1171–1180.

[22] C. Meng et al. “Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset”. In: *Scientific Reports* 12.1 (2022), p. 7166.

[23] S. Liu et al. “Accuracy and fairness trade-o□s in machine learning: A stochastic multi-objective approach”. In: *Computational Management Science* 19.3 (2022), pp. 513–537.

[24] A. L. Layer et al. “DuETT: Dual Event Time Transformer for Electronic Health Records”. In: (2023).

[25] Z. Deng et al. “Fifa: Making fairness more generalizable in classifiers trained on imbalanced data”. In: *arXiv preprint arXiv:2206.02792* (2022).

[26] M. Buyl et al. “Inherent Limitations of AI Fairness”. In: *arXiv:2212.06495* (2022).

[27] A. Johnson et al. “MIMIC-III, a freely accessible critical care database Sci”. In: *Data* 3.160035 (2016), pp. 10–1038.

[28] A. E. Johnson et al. “MIMIC-IV, a freely accessible electronic health record dataset”. In: *Scientific data* 10.1 (2023), p. 1.

[29] S. Tipirneni et al. “Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16.6 (2022), pp. 1–17.

[30] T. Gebru et al. “Datasheets for datasets”. In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[31] K. K. Saravanakumar. *The Impossibility Theorem of Machine Fairness – A Causal Perspective*. 2021. arXiv: 2007.06024 [cs.LG].

[32] M. Raghavan. “What Should We Do when Our Ideas of Fairness Conflict?” In: *Communications of the ACM* 67.1 (2023), pp. 88–97.

[33] URL: https://mimic.mit.edu/.

[34] G.-J. Geersing et al. “Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews”. In: *PloS one* 7.2 (2012), e32844.

[35] M. J. Page et al. “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews”. In: *International journal of surgery* 88 (2021), p. 105906.

[36] K. G. Moons et al. “Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration”. In: *Annals of internal medicine* 162.1 (2015), W1–W73.

[37] G. Kong et al. “Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU”. In: *BMC medical informatics and decision making* 20 (2020), pp. 1–10.

[38] W. Caicedo-Torres et al. “ISeeU: Visually interpretable deep learning for mortality prediction inside the ICU”. In: *Journal of biomedical informatics* 98 (2019), p. 103269.

[39] F. S. Ahmad et al. “A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (EHRs)”. In: *Journal of Ambient Intelligence and Humanized Computing* 12 (2021), pp. 3283–3293.

[40] S. Purushotham et al. “Benchmarking deep learning models on large healthcare datasets”. In: *Journal of biomedical informatics* 83 (2018), pp. 112–134.

[41] M. Feng et al. “Transthoracic echocardiography and mortality in sepsis: analysis of the MIMIC-III database”. In: *Intensive care medicine* 44 (2018), pp. 884–892.

[42] N. Hou et al. “Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost”. In: *Journal of translational medicine* 18.1 (2020), pp. 1–14.

[43] K. Lin et al. “Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model”. In: *International journal of medical informatics* 125 (2019), pp. 55–61.

[44] J. Xu et al. “Association of sex with clinical outcome in critically ill sepsis patients: a retrospective analysis of the large clinical database MIMIC-III”. In: *Shock* 52.2 (2019), pp. 146–151.

[45] B. Cheng et al. “Serum anion gap predicts all-cause mortality in critically ill patients with acute kidney injury: analysis of the MIMIC-III database”. In: *Disease markers* 2020 (2020).

[46] V. Sandfort et al. “Prolonged elevated heart rate and 90-day survival in acutely ill patients: data from the MIMIC-III database”. In: *Journal of intensive care medicine* 34.8 (2019), pp. 622–629.

[47] S. Zhou et al. “Early combination of albumin with crystalloids administration might be beneficial for the survival of septic patients: a retrospective analysis from MIMIC-IV database”. In: *Annals of intensive care* 11 (2021), pp. 1–10.

[48] F. Li et al. “Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database”. In: *BMJ open* 11.7 (2021), e044779.

[49] Z. Dai et al. “Analysis of adult disease characteristics and mortality on MIMIC-III”. In: *PloS one* 15.4 (2020), e0232176.

[50] W. Ye et al. “The association between neutrophil-to-lymphocyte count ratio and mortality in septic patients: a retrospective analysis of the MIMIC-III database”. In: *Journal of Thoracic Disease* 12.5 (2020), p. 1843.

[51] F. Gong et al. “The relationship between the serum anion gap and all-cause mortality in acute pancreatitis: an analysis of the MIMIC-III database”. In: *International Journal of General Medicine* (2021), pp. 531–538.

[52] X. Liu et al. “Serum anion gap at admission predicts all-cause mortality in critically ill patients with cerebral infarction: evidence from the MIMIC-III database”. In: *Biomarkers* 25.8 (2020), pp. 725–732.

[53] Y. Yu et al. “Admission oxygen saturation and all-cause in-hospital mortality in acute myocardial infarction patients: data from the MIMIC-III database”. In: *Annals of translational medicine* 8.21 (2020).

[54] Z. Nowroozilarki et al. “Real-time mortality prediction using MIMIC-IV ICU data via boosted nonparametric hazards”. In: *2021 IEEE EMBS international conference on biomedical and health informatics (BHI)*. IEEE. 2021, pp. 1–4.

[55] B. Huang et al. “Mortality prediction for patients with acute respiratory distress syndrome based on machine learning: a population-based study”. In: *Annals of translational medicine* 9.9 (2021).

[56] E.-q. Liu et al. “Blood urea nitrogen and in-hospital mortality in critically Ill patients with cardiogenic shock: analysis of the MIMIC-III database”. In: *BioMed Research International* 2021 (2021), pp. 1–7.

[57] H.-J. Jhou et al. “Plasma anion gap and risk of in-hospital mortality in patients with acute ischemic stroke: analysis from the MIMIC-IV database”. In: *Journal of Personalized Medicine* 11.10 (2021), p. 1004.

[58] Z. Zeng et al. “Development and validation of a novel blending machine learning model for hospital mortality prediction in ICU patients with Sepsis”. In: *BioData mining* 14 (2021), pp. 1–15.

[59] T. Zhang et al. “Association of acidemia with short-term mortality of acute myocardial infarction: a retrospective study base on MIMIC-III database”. In: *Clinical and Applied Thrombosis/Hemostasis* 26 (2020), p. 1076029620950837.

[60] S. Maheshwari et al. “A comprehensive evaluation for the prediction of mortality in intensive care units with LSTM networks: patients with cardiovascular disease”. In: *Biomedical Engineering/Biomedizinische Technik* 65.4 (2020), pp. 435–446.

[61] I. Bendavid et al. “A novel machine learning model to predict respiratory failure and invasive mechanical ventilation in critically ill patients su□ering from COVID-19”. In: *Scientific Reports* 12.1 (2022), p. 10573.

[62] W. Caicedo-Torres et al. “ISeeU2: Visually interpretable mortality prediction inside the ICU using deep learning and free-text medical notes”. In: *Expert Systems with Applications* 202 (2022), p. 117190.

[63] Y. Ji et al. “Lower serum chloride concentrations are associated with increased risk of mortality in critically ill cirrhotic patients: an analysis of the MIMIC-III database”. In: *BMC gastroenterology* 21.1 (2021), p. 200.

[64] V. Danilatou et al. “Outcome prediction in critically-ill patients with venous thromboembolism and/or cancer using machine learning algorithms: external validation and comparison with scoring systems”. In: *International Journal of Molecular Sciences* 23.13 (2022), p. 7132.

[65] Y. Zhu et al. “Serum anion gap level predicts all-cause mortality in septic patients: A retrospective study based on the MIMIC III database”. In: *Journal of Intensive Care Medicine* 38.4 (2023), pp. 349–357.

[66] K. Pang et al. “Establishment of ICU mortality risk prediction models with machine learning algorithm using MIMIC-IV database”. In: *Diagnostics* 12.5 (2022), p. 1068.

[67] C. Li et al. “Developing and verifying a multivariate model to predict the survival probability after coronary artery bypass grafting in patients with coronary atherosclerosis based on the MIMIC-III database”. In: *Heart & Lung* 52 (2022), pp. 61–70.

[68] C. Wang et al. “Cox-LASSO analysis for hospital mortality in patients with sepsis received continuous renal replacement therapy: a MIMIC-III database study”. In: *Frontiers in Medicine* 8 (2022), p. 778536.

[69] C. Liu et al. “E□ect of transthoracic echocardiography on short-term outcomes in patients with acute kidney injury in the intensive care unit: a retrospective cohort study based on the MIMIC-III database”. In: *Annals of translational medicine* 10.15 (2022).

[70] H. Fu et al. “The relationship between transthoracic echocardiography and mortality in adult patients with multiple organ dysfunction syndrome: analysis of the MIMIC-III database”. In: *Annals of Translational Medicine* 10.6 (2022).

[71] W. Liu et al. “Identification of key predictors of hospital mortality in critically ill patients with embolic stroke using machine learning”. In: *Bioscience Reports* 42.9 (2022), BSR20220995.

[72] Y.-Q. Liu et al. “Relationship between the red cell distribution width-to-platelet ratio and in-hospital mortality among critically ill patients with acute myocardial infarction: a retrospective analysis of the MIMIC-IV database”. In: *BMJ open* 12.9 (2022), e062384.

[73] Z. Xia et al. “Survival Prediction in Patients with Hypertensive Chronic Kidney Disease in Intensive Care Unit: A Retrospective Analysis Based on the MIMIC-III Database”. In: *Journal of Immunology Research* 2022 (2022).

[74] W. Xie et al. “Machine learning prediction models and nomogram to predict the risk of in-hospital death for severe DKA: A clinical study based on MIMIC-IV, eICU databases, and a college hospital ICU”. In: *International Journal of Medical Informatics* 174 (2023), p. 105049.

[75] J. Xu et al. “Timing of vasopressin initiation and mortality in patients with septic shock: analysis of the MIMIC-III and MIMIC-IV databases”. In: *BMC Infectious Diseases* 23.1 (2023), p. 199.

[76] W. Yang et al. “Mortality prediction among ICU inpatients based on MIMIC-III database results from the conditional medical generative adversarial network”. In: *Heliyon* 9.2 (2023).

[77] T. Liu et al. “The association between serum albumin and long length of stay of patients with acute heart failure: A retrospective study based on the MIMIC-IV database”. In: *Plos one* 18.2 (2023), e0282289.

[78] R. Zhang et al. “Independent e□ects of the triglyceride-glucose index on all-cause mortality in critically ill patients with coronary heart disease: analysis of the MIMIC-III database”. In: *Cardiovascular Diabetology* 22.1 (2023), p. 10.

[79] Z. Ye et al. “The prediction of in-hospital mortality in chronic kidney disease patients with coronary artery disease using machine learning models”. In: *European Journal of Medical Research* 28.1 (2023), pp. 1–13.

[80] H. Shi et al. “Association between early vasopressor administration and in-hospital mortality in critically ill patients with acute pancreatitis: A cohort study from the MIMIC-IV database.” In: *European Review for Medical & Pharmacological Sciences* 27.2 (2023).

[81] W. Jiang et al. “Development and validation of a nomogram for predicting in-hospital mortality of elderly patients with persistent sepsis-associated acute kidney injury in intensive care units: a retrospective cohort study using the MIMIC-IV database”. In: *BMJ open* 13.3 (2023), e069824.

[82] D. Bo et al. “Survival benefits of oral anticoagulation therapy in acute kidney injury patients with atrial fibrillation: a retrospective study from the MIMIC-IV database”. In: *BMJ open* 13.1 (2023), e069333.

[83] A. Ali et al. “Prediction of In-Hospital Mortality Among Heart Failure Patients: An Automated Machine Learning Analysis of Mimic-III Database”. In: *American Heart Journal* 254 (2022), p. 261.

[84] L. Zhao et al. “A novel prognostic model for predicting the mortality risk of patients with sepsis-related acute respiratory failure: a cohort study using the MIMIC-IV database”. In: *Current Medical Research and Opinion* 38.4 (2022), pp. 629–636.

[85] X. Jiang et al. “Comparison of machine learning algorithms to SAPS II in predicting in-hospital mortality of fractures of the pelvis and acetabulum: analyzes based on MIMIC-III database”. In: *All Life* 15.1 (2022), pp. 1000–1012.

[86] J. Sanii et al. “Explainable Machine Learning Models for Pneumonia Mortality Risk Prediction Using MIMIC-III Data”. In: *2022 9th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE. 2022, pp. 68–73.

[87] Y. Su et al. “Early predicting 30-day mortality in sepsis in MIMIC-III by an artificial neural networks model”. In: *European Journal of Medical Research* 27.1 (2022), p. 294.

[88] R. Tang et al. “Predictive value of machine learning for in-hospital mortality for trauma-induced acute respiratory distress syndrome patients: an analysis using the data from MIMIC III”. In: *Zhonghua wei Zhong Bing ji jiu yi xue* 34.3 (2022), pp. 260–264.

[89] R. Liu et al. “Predicting in-hospital mortality for MIMIC-III patients: A nomogram combined with SOFA score”. In: *Medicine* 101.42 (2022), e31251.

[90] S. Gu et al. “Development and validation of a RASS-related nomogram to predict the in-hospital mortality of neurocritical patients: a retrospective analysis based on the MIMIC-IV clinical database”. In: *Current Medical Research and Opinion* 38.11 (2022), pp. 1923–1933.

[91] X. Huang et al. “The hemoglobin-to-red cell distribution width ratio to predict all-cause mortality in patients with sepsis-associated encephalopathy in the MIMIC-IV database”. In: *International Journal of Clinical Practice* 2022 (2022).

[92] H. Tang et al. “Development and validation of a deep learning model to predict the survival of patients in ICU”. In: *Journal of the American Medical Informatics Association* 29.9 (2022), pp. 1567–1576.

[93] G. Salillari et al. “Comparison of Classification with Reject Option Approaches on MIMIC-IV Dataset”. In: *International Conference on Artificial Intelligence in Medicine*. Springer. 2022, pp. 210–219.

[94] X. Liu et al. “A machine learning predictive model of in-hospital mortality in patients with sepsis complicated by anemia: a retrospective study based on the MIMIC-III database”. In: (2021).

[95] T. Shu et al. “Development and assessment of scoring model for ICU stay and mortality prediction after emergency admissions in ischemic heart disease: a retrospective study of MIMIC-IV databases”. In: *Internal and Emergency Medicine* 18.2 (2023), pp. 487–497.

[96] Y. Fang et al. “Association between early ondansetron administration and in-hospital mortality in critically ill patients: analysis of the MIMIC-IV database”. In: *Journal of Translational Medicine* 20.1 (2022), p. 223.

[97] Y. Huo et al. “Impact of central venous pressure on the mortality of patients with sepsis-related acute kidney injury: a propensity score-matched analysis based on the MIMIC IV database”. In: *Annals of Translational Medicine* 10.4 (2022).

[98] W. Sun et al. “The e□ects of midazolam or propofol plus fentanyl on ICU mortality: a retrospective study based on the MIMIC-IV database”. In: *Annals of Translational Medicine* 10.4 (2022).

[99] Z.-y. Zou et al. “Early prophylactic anticoagulation with heparin alleviates mortality in critically ill patients with sepsis: a retrospective analysis from the MIMIC-IV database”. In: *Burns & Trauma* 10 (2022), tkac029.

[100] X.-D. Li et al. “A novel nomogram to predict mortality in patients with stroke: a survival analysis based on the MIMIC-III clinical database”. In: *BMC medical informatics and decision making* 22.1 (2022), p. 92.

[101] D. Han et al. “A Novel Nomogram for predicting survival in patients with severe acute pancreatitis: an analysis based on the large MIMIC-III Clinical Database”. In: *Emergency Medicine International* 2021 (2021).

[102] E. J. Tsiklidis et al. “Predicting risk for trauma patients using static and dynamic information from the MIMIC III database”. In: *Plos one* 17.1 (2022), e0262523.

[103] J. Wang et al. “Minimum heart rate and mortality in critically ill myocardial infarction patients: an analysis of the MIMIC-III database”. In: *Annals of translational medicine* 9.6 (2021).

[104] Y. Liu et al. “A time-incorporated SOFA score-based machine learning model for predicting mortality in critically ill patients: a multicenter, real-world study”. In: *International Journal of Medical Informatics* 163 (2022), p. 104776.

[105] H. Zhang et al. “The value of anion gap for predicting the short-term all-cause mortality of critically ill patients with cardiac diseases, based on MIMIC-III database”. In: *Heart & Lung* 55 (2022), pp. 59–67.

[106] Z. Qin et al. “Relationship between the hemoglobin-to-red cell distribution width ratio and all-cause mortality in ischemic stroke patients with atrial fibrillation: an analysis from the MIMIC-IV database”. In: *Neuropsychiatric Disease and Treatment* 18 (2022), p. 341.

[107] J.-C. Peng et al. “Favorable outcomes of anticoagulation with unfractioned heparin in sepsis-induced coagulopathy: a retrospective analysis of MIMIC-III database”. In: *Frontiers in Medicine* 8 (2022), p. 773339.

[108] Y. Zhao et al. “Statistical analysis and machine learning prediction of disease outcomes for COVID-19 and pneumonia patients”. In: *Frontiers in cellular and infection microbiology* 12 (2022), p. 838749.

[109] S. Wu et al. “The association between systemic immune-inflammation index and all-cause mortality in acute ischemic stroke patients: analysis from the MIMIC-IV database”. In: *Emergency medicine international* 2022 (2022), pp. 1–10.

[110] N. Ding et al. “An artificial neural networks model for early predicting in-hospital mortality in acute pancreatitis in MIMIC-III”. In: *BioMed research international* 2021 (2021).

[111] M. Luo et al. “Association between hematocrit and the 30-day mortality of patients with sepsis: A retrospective analysis based on the large-scale clinical database MIMIC-IV”. In: *PloS one* 17.3 (2022), e0265758.

[112] D. Liu et al. “Admission hyperglycemia predicts long-term mortality in critically Ill patients with subarachnoid hemorrhage: a retrospective analysis of the MIMIC-III database”. In: *Frontiers in Neurology* 12 (2021), p. 678998.

[113] D. Wang et al. “E□ect of Admission Serum Calcium Levels and Length of Stay in Patients with Acute Pancreatitis: Data from the MIMIC-III Database”. In: *Emergency Medicine International* 2022 (2022).

[114] D. Geethamani et al. “Heterogeneous Multi-Model Ensemble based Length of Stay Prediction on MIMIC III”. In: (2022).

[115] D. Wang et al. “Relationship between blood glucose levels and length of hospital stay in patients with acute pancreatitis: An analysis of MIMIC-III database”. In: *Clinical and Translational Science* 16.2 (2023), pp. 246–257.

[116] S. Gokhale et al. “Hospital length of stay prediction for general surgery and total knee arthroplasty admissions: Systematic review and meta-analysis of published prediction models”. In: *Digital Health* 9 (2023), p. 20552076231177497.

[117] C.-C. Yang et al. “Risk factor identification and prediction models for prolonged length of stay in hospital after acute ischemic stroke using artificial neural networks”. In: *Frontiers in Neurology* 14 (2023), p. 1085178.

[118] J. L. Tully et al. “Machine learning prediction models to reduce length of stay at ambulatory surgery centers through case resequencing”. In: *Journal of Medical Systems* 47.1 (2023), p. 71.

[119] W.-T. Chiu et al. “Identifying Risk Factors for Prolonged Length of Stay in Hospital and Developing Prediction Models for Patients with Cardiac Arrest Receiving Targeted Temperature Management”. In: *Reviews in Cardiovascular Medicine* 24.2 (2023), p. 55.

[120] Y. Deng et al. “Explainable time-series deep learning models for the prediction of mortality, prolonged length of stay and 30-day readmission in intensive care patients”. In: *Frontiers in Medicine* 9 (2022), p. 933037.

[121] Y. Selim et al. 2022.

[122] J. Thacker. “A Machine Learning Pipeline for Readmission Prediction with MIMIC-III”. PhD thesis. Auburn University, 2023.

[123] Q. Chen et al. “Outcome-Oriented Predictive Process Monitoring to Predict Unplanned ICU Readmission in MIMIC-IV Database”. In: (2022).

[124] J. Zhang et al. “Clinical utility of automatic phenotype annotation in unstructured clinical notes: intensive care unit use”. In: *BMJ Health & Care Informatics* 29.1 (2022), e100519.

[125] S. Yang et al. “Machine learning approaches for electronic health records phenotyping: a methodical review”. In: *Journal of the American Medical Informatics Association* 30.2 (2023), pp. 367–381.

[126] F. Li et al. “ICD coding from clinical text using multi-filter residual convolutional neural network”. In: *proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 05. 2020, pp. 8180–8187.

[127] J. Huang et al. “An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes”. In: *Computer methods and programs in biomedicine* 177 (2019), pp. 141–153.

[128] S. Yue et al. “Machine learning for the prediction of acute kidney injury in patients with sepsis”. In: *Journal of translational medicine* 20.1 (2022), pp. 1–12.

[129] C. Sun et al. “Association between acute kidney injury and prognoses of cardiac surgery patients: Analysis of the MIMIC-III database”. In: *Frontiers in Surgery* 9 (2022).

[130] T. Liu et al. “E□ects of high-flow oxygen therapy on patients with hypoxemia after extubation and predictors of reintubation: a retrospective study based on the MIMIC-IV database”. In: *BMC Pulmonary Medicine* 21.1 (2021), pp. 1–15.

[131] M. Böck et al. “Superhuman performance on sepsis MIMIC-III data by distributional reinforcement learning”. In: *PLoS One* 17.11 (2022), e0275358.

[132] A. A. R. Magna et al. “Application of machine learning and word embeddings in the classification of cancer diagnosis using patient anamnesis”. In: *Ieee Access* 8 (2020), pp. 106198–106213.

[133] M. Sayed et al. “Predicting duration of mechanical ventilation in acute respiratory distress syndrome using supervised machine learning”. In: *Journal of Clinical Medicine* 10.17 (2021), p. 3824.

[134] S. Goethals et al. “Beyond Accuracy-Fairness: Stop evaluating bias mitigation methods solely on between-group metrics”. In: *arXiv preprint arXiv:2401.13391* (2024).

[135] C. Denis et al. *Fairness guarantee in multi-class classification*. 2023. arXiv: 2109.13642 [math.ST].

[136] S. Caton et al. *Fairness in Machine Learning: A Survey*. 2020. arXiv: 2010.04053 [cs.LG].

[137] A. L. Goldberger et al. “PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals”. In: *circulation* 101.23 (2000), e215–e220.

[138] J. C. Rojas et al. “Predicting intensive care unit readmission with machine learning using electronic health record data”. In: *Annals of the American Thoracic Society* 15.7 (2018), pp. 846–853.

[139] R. Assaf et al. “30-day hospital readmission prediction using MIMIC data”. In: *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*. IEEE. 2020, pp. 1–6.

[140] A. Partovi et al. “MiPy: A Framework for Benchmarking Machine Learning Prediction of Unplanned Hospital and ICU Readmission in the MIMIC-IV Database”. In: (2022).

[141] J. Theis et al. “Improving the in-hospital mortality prediction of diabetes ICU patients using a process mining/deep learning architecture”. In: *IEEE Journal of Biomedical and Health Informatics* 26.1 (2021), pp. 388–399.

[142] H. Chen et al. “Association between normalized lactate load and mortality in patients with septic shock: an analysis of the MIMIC-III database”. In: *BMC anesthesiology* 21.1 (2021), pp. 1–8.

[143] Z. Lu et al. “Development of a nomogram to predict 28-day mortality of patients with sepsis-induced coagulopathy: an analysis of the MIMIC-III database”. In: *Frontiers in medicine* 8 (2021), p. 661710.

[144] Q. Gao et al. “Sentiment analysis based on the nursing notes on in-hospital 28-day mortality of sepsis patients utilizing the MIMIC-III database”. In: *Computational and Mathematical Methods in Medicine* 2021 (2021).

[145] A. Singh et al. “Multi-label natural language processing to identify diagnosis and procedure codes from MIMIC-III inpatient notes”. In: *arXiv preprint arXiv:2003.07507* (2020).

[146] H. Dong et al. “Ontology-based and weakly supervised rare disease phenotyping from clinical notes”. In: *arXiv preprint arXiv:2205.05656* (2022).

[147] M. Li et al. “Automated ICD-9 coding via a deep learning approach”. In: *IEEE/ACM transactions on computational biology and bioinformatics* 16.4 (2018), pp. 1193–1202.

[148] A. P. Kurniati et al. “Process mining in oncology using the MIMIC-III dataset”. In: *Journal of Physics: Conference Series*. Vol. 971. 1. IOP Publishing. 2018, p. 012008.

[149] G. Geri et al. “Cardio-pulmonary-renal interactions in ICU patients. Role of mechanical ventilation, venous congestion and perfusion deficit on worsening of renal function: Insights from the MIMIC-III database”. In: *Journal of critical care* 64 (2021), pp. 100–107.

[150] M. Mitchell et al. “Model cards for model reporting”. In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 220–229.

[151] T. J. Pollard et al. “The eICU Collaborative Research Database, a freely available multi-center database for critical care research”. In: *Scientific data* 5.1 (2018), pp. 1–13.

[152] S. L. Hyland et al. “Early prediction of circulatory failure in the intensive care unit using machine learning”. In: *Nature medicine* 26.3 (2020), pp. 364–373.

[153] M. Goossens et al. *The LaTeX Companion*. Reading, Massachusetts: Addison-Wesley, 1994.

[154] D. Knuth. *The TeXbook*. Reading, Massachusetts: Addison-Wesley, 1986.

[155] L. Lamport. *LaTeX — A Document Preparation System*. Second. Reading, Massachusetts: Addison-Wesley, 1994.

[156] N. Japkowicz. “The class imbalance problem: Significance and strategies”. In: *Proc. of the Int’l Conf. on artificial intelligence*. Vol. 56. 2000, pp. 111–117.

[157] H. He et al. “Learning from imbalanced data”. In: *IEEE Transactions on knowledge and data engineering* 21.9 (2009), pp. 1263–1284.

[158] Y. Sun et al. “Classification of imbalanced data: A review”. In: *International journal of pattern recognition and artificial intelligence* 23.04 (2009), pp. 687–719.

[159] J. Luo et al. “Big data application in biomedical research and health care: a literature review”. In: *Biomedical informatics insights* 8 (2016), BII–S31559.

[160] W. Raghupathi et al. “Big data analytics in healthcare: promise and potential”. In: *Health information science and systems* 2 (2014), pp. 1–10.

[161] Z. Obermeyer et al. “Dissecting racial bias in an algorithm used to manage the health of populations”. In: *Science* 366.6464 (2019), pp. 447–453.

[162] R. B. Parikh et al. “Addressing bias in artificial intelligence in health care”. In: *Jama* 322.24 (2019), pp. 2377–2378.

[163] K.-H. Yu et al. “Artificial intelligence in healthcare”. In: *Nature biomedical engineering* 2.10 (2018), pp. 719–731.

[164] E. Strelcenia et al. “A Survey on GAN Techniques for Data Augmentation to Address the Imbalanced Data Issues in Credit Card Fraud Detection”. In: *Machine Learning and Knowledge Extraction* 5.1 (2023), pp. 304–329.

[165] G. Chandrashekar et al. “A survey on feature selection methods”. In: *Computers & Electrical Engineering* 40.1 (2014), pp. 16–28.

[166] J. Sen et al. “A time series analysis-based forecasting framework for the Indian healthcare sector”. In: *arXiv preprint arXiv:1705.01144* (2017).

[167] M. Gupta et al. “An extensive data processing pipeline for mimic-iv”. In: *Machine Learning for Health*. PMLR. 2022, pp. 311–325.

[168] S. Wang et al. “Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii”. In: *Proceedings of the ACM conference on health, inference, and learning*. 2020, pp. 222–235.

[169] A. Johnson et al. “MIMIC-IV-ED”. In: *PhysioNet* (2021).

[170] I. Y. Chen et al. “Ethical machine learning in healthcare”. In: *Annual review of biomedical data science* 4 (2021), pp. 123–144.

[171] J. K. Paulus et al. “Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities”. In: *NPJ digital medicine* 3.1 (2020), p. 99.

[172] D. R. Williams et al. “Racism and health I: Pathways and scientific evidence”. In: *American behavioral scientist* 57.8 (2013), pp. 1152–1173.

[173] B. D. Sommers et al. “Health insurance coverage and health—what the recent evidence tells us”. In: *N Engl J Med* 377.6 (2017), pp. 586–593.

[174] O. Baclic et al. “Artificial intelligence in public health: Challenges and opportunities for public health made possible by advances in natural language processing”. In: *Canada Communicable Disease Report* 46.6 (2020), p. 161.

[175] E. Racine et al. “Healthcare uses of artificial intelligence: Challenges and opportunities for growth”. In: *Healthcare management forum*. Vol. 32. 5. SAGE Publications Sage CA: Los Angeles, CA. 2019, pp. 272–275.

[176] M. Millman et al. “Access to health care in America”. In: (1993).

[177] E. Tasci et al. “Bias and Class Imbalance in Oncologic Data—Towards Inclusive and Transferrable AI in Large Scale Oncology Data Sets”. In: *Cancers* 14.12 (2022), p. 2897.

[178] R. M. Weinick et al. “Racial and ethnic di□erences in access to and use of health care services, 1977 to 1996”. In: *Medical Care Research and Review* 57.1_suppl (2000), pp. 36–54.

[179] R. M. Mayberry et al. “Racial and ethnic di□erences in access to medical care”. In: *Medical Care Research and Review* 57.1_suppl (2000), pp. 108–145.

[180] S. Cruz-Flores et al. “Racial-ethnic disparities in stroke care: the American experience: a statement for healthcare professionals from the American Heart Association/American Stroke Association”. In: *Stroke* 42.7 (2011), pp. 2091–2116.

[181] R. K. Bailey et al. “Racial and ethnic di□erences in depression: current perspectives”. In: *Neuropsychiatric disease and treatment* (2019), pp. 603–609.

[182] T. Janevic et al. ““Just because you have ears doesn’t mean you can hear”—perception of racial-ethnic discrimination during childbirth”. In: *Ethnicity & Disease* 30.4 (2020), p. 533.

[183] T. A. LaVeist. “Beyond dummy variables and sample selection: what health services researchers ought to know about race as a variable.” In: *Health services research* 29.1 (1994), p. 1.

[184] D. A. Vyas et al. *Hidden in plain sight—reconsidering the use of race correction in clinical algorithms*. 2020.

[185] R. Cooper et al. “The biological concept of race and its application to public health and epidemiology”. In: *Journal of Health Politics, Policy and Law* 11.1 (1986), pp. 97–116.

[186] D. E. Crews et al. “Ethnicity as a taxonomic tool in biomedical and biosocial research”. In: *Ethnicity & Disease* (1991), pp. 42–49.

[187] N. Krieger. “Shades of di□erence: theoretical underpinnings of the medical controversy on black/white di□erences in the United States, 1830–1870”. In: *International Journal of Health Services* 17.2 (1987), pp. 259–278.

[188] S. S. Anand. “Using ethnicity as a classification variable in health research: perpetuating the myth of biological determinism, serving socio-political agendas, or making valuable contributions to medical sciences?” In: *Ethnicity and Health* 4.4 (1999), pp. 241–244.

[189] M. Chen et al. “Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review”. In: *Journal of the American Medical Informatics Association* 27.11 (2020), pp. 1764–1773.

[190] C. C. Gravlee. “How race becomes biology: embodiment of social inequality”. In: *American journal of physical anthropology* 139.1 (2009), pp. 47–57.

[191] J. L. Alhusen et al. “Racial discrimination and adverse birth outcomes: an integrative review”. In: *Journal of midwifery & women’s health* 61.6 (2016), pp. 707–720.

[192] K. R. van Daalen et al. “Racial discrimination and adverse pregnancy outcomes: a systematic review and meta-analysis”. In: *BMJ Global Health* 7.8 (2022), e009227.

[193] P. Skrabanek. “The emptiness of the black box”. In: *Epidemiology* (1994), pp. 553–555.

[194] A. Menotti et al. “Comparison of multivariate predictive power of major risk factors for coronary heart diseases in di□erent countries: results from eight nations of the Seven Countries Study, 25-year follow-up”. In: *Journal of cardiovascular risk* 3.1 (1996), pp. 69–75.

[195] P. A. Senior et al. “Ethnicity as a variable in epidemiological research”. In: *Bmj* 309.6950 (1994), pp. 327–330.

[196] A. Avati et al. “Improving palliative care with deep learning”. In: *BMC medical informatics and decision making* 18.4 (2018), pp. 55–64.

[197] B. Butcher et al. *Feature Engineering and Selection: A Practical Approach for Predictive Models: by Max Kuhn and Kjell Johnson. Boca Raton, FL: Chapman & Hall/CRC Press, 2019, xv+ 297 pp., $79.95 (H), ISBN: 978-1-13-807922-9.* 2020.

[198] W. Boag et al. “EHR Safari: Data is Contextual”. In: (2022).

[199] S. Dash et al. “Big data in healthcare: management, analysis and future prospects”. In: *Journal of Big Data* 6.1 (2019), pp. 1–25.

[200] J. A. Doshi et al. “Data, data everywhere, but access remains a big issue for researchers: a review of access policies for publicly-funded patient-level health care data in the United States”. In: *eGEMs* 4.2 (2016).

[201] A. Makady et al. “Policies for use of real-world data in health technology assessment (HTA): a comparative study of six HTA agencies”. In: *Value in Health* 20.4 (2017), pp. 520–532.

[202] C. L. Brown. “Health-Care Data Protection and Biometric Authentication Policies: Comparative Culture and Technology Acceptance in China and in the United States”. In: *Review of Policy Research* 29.1 (2012), pp. 141–159.

[203] L. Rasmy et al. “Representation of EHR data for predictive modeling: a comparison between UMLS and other terminologies”. In: *Journal of the American Medical Informatics Association* 27.10 (2020), pp. 1593–1599.

[204] X.-Q. Luo et al. “Development and validation of machine learning models for real-time mortality prediction in critically ill patients with sepsis-associated acute kidney injury”. In: *Frontiers in Medicine* 9 (2022), p. 853102.

[205] K. Hur et al. “UniHPF: Universal Healthcare Predictive Framework with Zero Domain Knowledge”. In: *arXiv preprint arXiv:2211.08082* (2022).

[206] A. K. Menon et al. “The cost of fairness in binary classification”. In: *Conference on Fairness, accountability and transparency*. PMLR. 2018, pp. 107–118.

[207] J. Chai et al. “Fairness with adaptive weights”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 2853–2866.

[208] G. Zhang et al. “Fairness reprogramming”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 34347–34362.

[209] P. Team. “LaTeX Compatibility”. In: *PubPub Help* (Jan. 2021). https://help.pubpub.org/pub/latex-compatibility.

[210] M. Abramowitz et al. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Vol. 55. US Government printing o□ce, 1948.

[211] G. Demartini et al. “Data Bias Management”. In: *arXiv preprint arXiv:2305.09686* (2023).

[212] C. Dwork et al. “Fairness through awareness”. In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012, pp. 214–226.

APPENDICES

Appendix A

Datasheet for MIMIC IV v2.0

This webpage provides access to the entire datasheet. The section below provides a closer look at the sensitive attribute correlation analysis of HML risk prediction tasks.

MIMIC-III and IV have been pivotal in healthcare machine learning (HML) prediction tasks. This analysis zeroes in on MIMIC-IV version 2.0, the most recent release available to the public. The table 5.1 provides a complete overview of the sensitive attributes of MIMIC IV ICU data.

A.1 Sensitive attribute correlation analysis with risk prediction outcomes

Our study employed the *Chi-Square* statistical test to discern the attribute most strongly associated with prediction outcomes. The chi-square test results illuminate the relationships between patient-sensitive characteristics and mortality, guiding the feature selection for predictive modeling and fairness evaluation. Gender, with a chi-square statistic of 3.33 and a p-value of approximately 0.068, shows a marginal association with mortality; however, it falls just outside the conventional alpha level of 0.05 for statistical significance. Despite a high chi-square statistic of 151.77, age yields a p-value of 0.467, suggesting that the observed variations across different ages could be due to chance, thus making it less reliable for predicting mortality in this context.

Language shows a similar pattern to gender, with a chi-square statistic of 3.37 and a p-value of about 0.066, hovering near the boundary of significance but not conclusively so. This

marginal association suggests that while there might be some relationship with mortality, it is not as strong or as clear-cut as one would prefer for a predictive model.

In contrast, *ethnicity*, and *insurance* status demonstrate a much stronger association with mortality outcomes. Ethnicity, in particular, stands out with a significant chi-square statistic and a p-value of $2.704e^{-15}$ that effectively rejects the null hypothesis of no association. While insurance status also exhibits a significant p-value ($2.630e^{-17}$), the decision to focus on ethnicity over insurance (potential ambiguity in Other Insurance information) is driven by the detailed and consistent recording of ethnic data in the MIMIC dataset. This level of detail in the ethnicity data provides a solid foundation for predictive analysis, ensuring that the model's outcomes are reliable and interpretable.

From the survey conducted in the chapter 2, we identified the below as the most widely researched predictive modeling tasks. So we procured datasets for,

- In-hospital mortality concerning
  - Heart failure,
  - Chronic kidney disease (CKD),
  - Sepsis.
- 30-day readmission, and
- Length of stay (LOS) for heart failure by adhering to the established pipeline except for sepsis mortality.

For sepsis mortality, we directly extracted patient data affected by sepsis, omitting the use of a specific pipeline, to validate and compare outcomes derived from both methodologies.

A.1.1 In-Hospital Mortality Prediction

Sepsis-Related Ailments

Cohort distribution insights - The cohort's median age was 63, evenly distributed between male (53%) and female (47%) patients. The dataset primarily comprised the White demographic (69%), with subsequent representation from Black, Other, Hispanic/Latino, and Asian subgroups.

Cohort insurance utilization - Medicare emerged as the most utilized insurance across all ethnicities (51%), with Other insurance closely behind at 41%. Medicaid saw the least

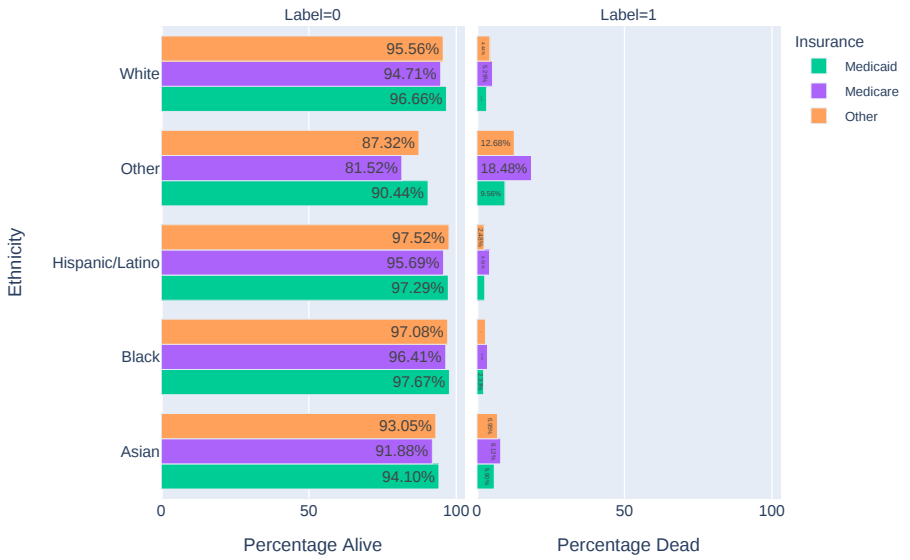


Figure A.1: Sepsis Mortality in Relation to Ethnicity/Insurance: The Figure illustrates the % mortality among sepsis patients by depicting the variations in mortality across di□erent ethnic groups and insurance categories, with Label 0 denoting the alive patients and Label 1 indicating the deceased.

utilization at 8%. Notably, White and Black patients had the highest Medicare utilization, indicative of an older population within these communities. Conversely, minority groups showed a preference for Other insurance, suggesting a relatively younger demographic or the presence of private insurance coverage. Despite being the largest demographic, only 5.5% of Caucasians utilized Medicaid, whereas Black patients had a higher Medicaid utilization, highlighting socio-economic disparities.

Mortality rate analysis - Analysis starkly illustrates that individuals from other ethnic groups and Asians had consistently higher death proportions, irrespective of their insurance status. The Chi-square test ($^2 = 975.185$, $p < 0.001$) validated a strong association between ethnicity and sepsis mortality, underlining the necessity of considering demographic variables in predicting sepsis mortality.

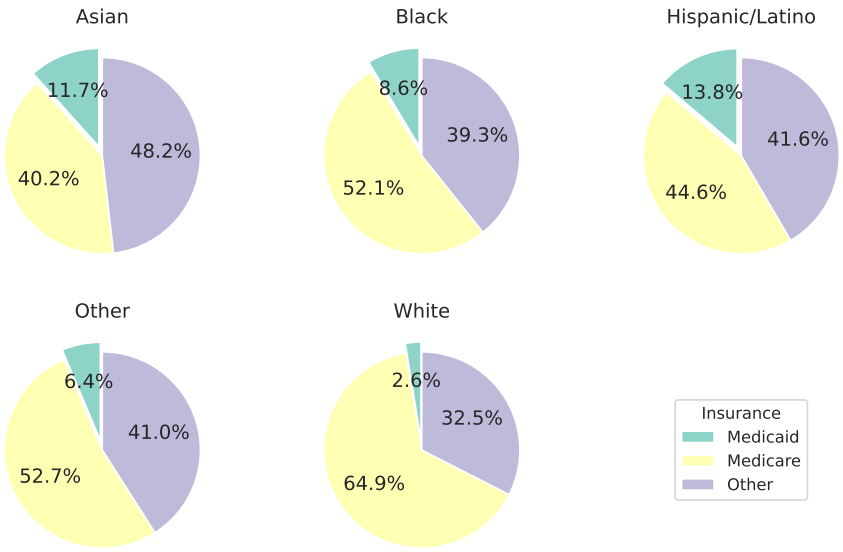


Figure A.2: Insurance Utilization among Heart Failure Cohorts based on Ethnicity: The figure showcases the % insurance utilization among heart failure patients, categorized by their respective ethnicity.

Heart failure

Cohort distribution insights - The heart failure cohort consisted of patients with a median age of 72, with a nearly equal distribution of males (55%) and females (45%). Most patients (70%) belonged to the White demographic, followed by the Black, Other, Hispanic/Latino, and Asian subgroups, reflecting the specific geographical location of data collection.

Cohort insurance utilization - White and Black individuals had the highest patient counts, followed by the Other subgroup. Medicare was the most commonly used insurance among all the subgroups, except for Asians, as indicated by the figure. White patients exhibited the highest utilization of Medicare, followed by the Other-ethnic subgroup. Medicaid utilization was lower across all subgroups except for Hispanic/Latino, Asian, and Black patients.

Mortality rate analysis - Individuals from other ethnic groups consistently exhibited higher mortality rates, irrespective of their insurance. Asians insured with Other and Medicaid also demonstrated a relatively higher proportion of deaths. Caucasian death rates were significantly lower in comparison to other subgroups. The chi-square test statistic yielded a significant result of 105.107 ($p < 0.001$), providing robust evidence of an

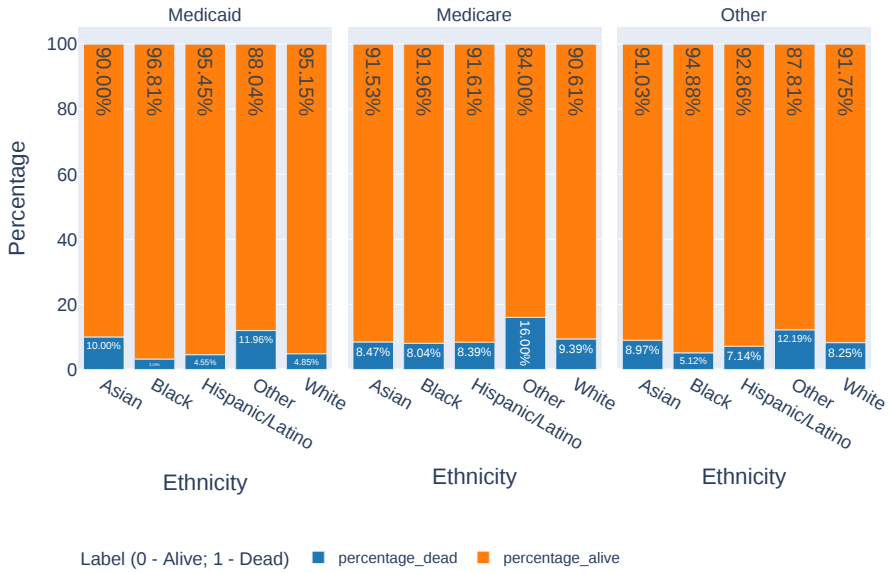


Figure A.3: % CKD mortality rate with respect to their Ethnicity/Insurance.

association between the variables.

Chronic Kidney Disease

Cohort distribution insights - The CKD cohort primarily consists of older male patients with a median age of 71. White individuals represent the majority (64%), followed by Black, Other, Hispanic/Latino, and Asian subgroups. White and Black patients have the highest patient counts, along with the Other subgroup.

Cohort insurance utilization - Medicare insurance is widely utilized, with 60.5% of patients across all ethnicities opting for it. Medicaid records the lowest utilization, with only 4.7% of patients using it. Among the race subgroups, White and Black individuals demonstrate the highest usage of Medicare insurance, while Asians show a preference for other insurance types. 2.3% of Caucasians utilize Medicaid insurance, whereas Black patients have a higher utilization of Medicaid insurance compared to different subgroups.

Mortality rate analysis - The figure shows individuals from other-ethnic groups consistently exhibit higher proportions of deaths, regardless of their insurance type. Medicare-insured Blacks experience the second-highest mortality rates, preceded by Caucasians.

Figure A.4: Relationship between Readmission rates and patient's Ethnicity/Insurance.

The chi-square ($^2 = 106.578$, $p < 0.001$) statistical test shows a significant association, suggesting a potential relationship between ethnicity and mortality outcomes.

A.1.2 30-day ICU Readmission Analysis

Cohort distribution insights - The median age for patients readmitted to the ICU within 30 days was 64, with a distribution of 56% male and 44% female. Predominantly, the White demographic represented 69% of the cohort, followed by Black, Other, Hispanic/Latino, and Asian subgroups, reflecting the geographical context of data collection.

Cohort insurance utilization - A significant portion of the cohort primarily relied on Other insurance (48%), with Medicare following closely at 44%. The analysis indicates that Other-subgroup and Hispanic/Latino patients predominantly utilized Other insurance, with Black and Asian patients following suit. Medicaid utilization was notably lower across all ethnic subgroups, at 8%. Interestingly, only a small fraction (5.4%) of Caucasian patients utilized Medicaid, with Black and Other ethnic patients showing higher Medicaid utilization, underscoring socio-economic disparities.

Table A.1: Comparative Analysis of % Length of ICU stay (</> 7 days) of di□erent demographic groups based on their insurance status in the heart disease prediction task.

LOS		%	
Ethnicity	Insurance	<7 days	>7 days
Asian	Medicaid	75.926	24.074
	Medicare	87.634	12.366
	Other	82.511	17.489
Black	Medicaid	88.938	11.062
	Medicare	88.377	11.623
	Other	87.197	12.803
Hispanic/Latino	Medicaid	91.304	8.696
	Medicare	85.185	14.815
	Other	86.282	13.718
Other	Medicaid	78.947	21.053
	Medicare	82.698	17.302
	Other	82.653	17.347
White	Medicaid	83.898	16.102
	Medicare	87.341	12.659
	Other	85.873	14.127

Readmission Rates analysis - The analysis reveals that the Black subgroup, along with Medicaid-insured Asian patients, displayed higher readmission rates across various ailments. The chi-square ($^2 = 141.89$, $p < 0.001$) test further corroborates the significant association between ethnicity concerning readmission rates.

A.1.3 ICU Length of Stay Prediction

The ICU LOS > 7 days cohort had a median age of 72, balanced between males (55%) and females (45%). The dataset predominantly consisted of Whites (69%), followed by other demographic groups.

Furthermore, the Chi-square ($p < 0.001$) statistical test confirms an association between LOS and ethnicity, reinforcing the influence of these demographic factors on ICU outcomes for heart failure patients.

This analysis underscores the importance of considering the feature association and data quality in predictive modeling.