# Evaluating Responsible AI Principles in Healthcare Application

**Investigating Fairness, Accountability, and Transparency**

Bezawit Lake Tilaye

A dissertation submitted to the University of Bristol in accordance with the requirements of the Foundation year of the Doctor of Philosophy in Practice-oriented AI in the Faculty of Engineering and Science.

School of Computer Science

August 2025

Word count: 10148

# Abstract

The availability of publicly accessible healthcare datasets has significantly accelerated the development of artificial intelligence (AI) systems, with advances in clinical decision-making and operational efficiency. However, evaluation studies done on MIMIC-III (Medical Information Mart for Intensive Care, version) database have reported concerns regarding the fairness of such systems. This project extends the evaluation to the MIMIC-IV database, focusing on in-hospital mortality prediction as a case study. Fairness, accountability, and transparency are assessed across the full machine learning pipeline, from data representation and preprocessing to model training and prediction, using the FAT Forensics toolkit.

The findings reveal fairness concerns affecting minority and disadvantaged subgroups and highlight the need for process transparency and accountability mechanisms throughout the pipeline. Over 3,900 matched patients exhibited label inconsistencies across insurance types. Additionally, true positive rates (TPRs) varied significantly across race and insurance subgroups, especially for Asian, Hispanic, and Medicaid patients. Linear and KNN models showed the high disparities, often exceeding a 10% threshold, whereas tree-based models like XGBoost demonstrated more equitable performance across subgroups. The results underscore that a full pipeline evaluation provides deeper insights, into risks in AI development. Therefore, the necessity of thorough and systematic auditing across the machine learning pipeline is emphasized to ensure the development of equitable, transparent, and trustworthy AI systems in healthcare.

# Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Taught Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, this work is my own work. Work done in collaboration with, or with the assistance of others, is indicated as such. I have identified all material in this dissertation which is not my own work through appropriate referencing and acknowledgement. Where I have quoted or otherwise incorporated material, which is the work of others, I have included the source in the references. Any views expressed in the dissertation, other than referenced material, are those of the author.

Bezawit Lake Tilaye
29 August 2025

# Contents

List of Tables iv

List of Figures v

1 Introduction 1

1.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 Aim and Objectives . . . . . . . . . . . . . . . . . . . . . . . . . 2

2 Literature Review 4

2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2.2 Responsible AI in Healthcare . . . . . . . . . . . . . . . . . . . . 4

2.2.1 Fairness . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

2.2.2 Transparency . . . . . . . . . . . . . . . . . . . . . . . . . 6

2.2.3 Accountability . . . . . . . . . . . . . . . . . . . . . . . . 7

2.3 MIMIC Dataset Evaluation . . . . . . . . . . . . . . . . . . . . . 9

2.4 Tools and Frameworks for Evaluating Responsible AI . . . . . . . . . . . . 10

3 Methodology 12

3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

3.2 Dataset and Preprocessing . . . . . . . . . . . . . . . . . . . . . . 12

3.2.1 Cohort Preparation . . . . . . . . . . . . . . . . . . . . . . 13

3.2.2 Feature Selection . . . . . . . . . . . . . . . . . . . . . . . 14

3.2.3 Feature Aggregation . . . . . . . . . . . . . . . . . . . . . . . . . . 15
3.3 Algorithms . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
3.4 Evaluation of the Dataset . . . . . . . . . . . . . . . . . . . . . . . . . 19
3.5 Evaluation of Model Performance and Predictions . . . . . . . . . . . . . . 21

**4 Results** **24**
4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
4.2 Cohort Preparation . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
4.3 Dataset Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26
4.3.1 Monitoring Frequency . . . . . . . . . . . . . . . . . . . . . . . . . 26
4.3.2 Missingness Rate . . . . . . . . . . . . . . . . . . . . . . . . . . . 26
4.3.3 Monitoring Ratio . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
4.3.4 Assessing Systemic Bias in the Dataset . . . . . . . . . . . . . . . . 30
4.4 Model and Prediction Evaluation . . . . . . . . . . . . . . . . . . . . . . 34
4.4.1 Area Under the ROC Curve (AUC-ROC) . . . . . . . . . . . . . . . . 34
4.4.2 Equal opportunity . . . . . . . . . . . . . . . . . . . . . . . . . . . 36
4.4.3 Demographic parity . . . . . . . . . . . . . . . . . . . . . . . . . . 38
4.4.4 Permutation Feature Importance . . . . . . . . . . . . . . . . . . . . 39
4.4.5 Counterfactual Fairness . . . . . . . . . . . . . . . . . . . . . . . . 39

**5 Discussion** **41**
5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41
5.2 Representation and Data Quality . . . . . . . . . . . . . . . . . . . . . . 41
5.3 Label Validity and Systemic Bias . . . . . . . . . . . . . . . . . . . . . 42
5.4 Model Performance and Subgroup Disparities . . . . . . . . . . . . . . . . 43
5.5 Permutation Feature Importance and Counterfactual Fairness . . . . . . . 44
5.6 Comparison with evaluation done on MIMIC-III . . . . . . . . . . . . . . 45

5.7 Implications for Responsible AI in Healthcare . . . . . . . . . . . . . . . . . 46

**6 Conclusion and Future Work 48**

6.1 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 48

6.2 Future Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 48

**Bibliography 50**

# List of Tables

1 Introduction 1

2 Literature Review 4

3 Methodology 12

4 Results 24

4.1 Cohort distribution across protected features. 25

4.2 Coverage of 17 features selected from MIMIC-IV database. Features with less than 40% coverage (highlighted in bold) were excluded. 27

4.3 Feature value ranges and corresponding binning intervals. 32

4.4 Model performance across demographic subgroups. Values represent AUC scores for each model. The *Death Rate* column reports the observed mortality in each subgroup. Best-performing models per group are shown in bold. 36

4.5 Comparison of True Positive Rates (TPRs) Across Insurance Groups and Models 38

5 Discussion 41

6 Conclusion and Future Work 48

Bibliography 50

# List of Figures

1 Introduction 1

2 Literature Review 4

3 Methodology 12

4 Results 24

4.1 Rate of missing vital sign data per insurance subgroup. . . . . . . . . . . . 28

4.2 Rate of missing vital sign data per race subgroup. . . . . . . . . . . . . . . 29

4.3 Monitoring Ratios of vital signs within the 48-Hour observation window across insurance subgroups. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30

4.4 Monitoring ratios of vital signs within the 48-Hour observation window across race subgroups. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31

4.5 Labelling bias related to insurance groups. . . . . . . . . . . . . . . . . . . . . 33

4.6 Cumulative distribution function (CDF) of Dynamic Time Warping (DTW) distances between matched patient pairs. Left: All features excluding PaO2 and Heart Rate. Middle: PaO2 only. Right: Heart rate only. . . . . 34

4.7 True Positive Rate (TPR) differences across racial subgroups for various models. Green cells indicate fairness within the 10% tolerance (fair), and red cells indicate potential bias where difference exceed 10%). . . . . . . . 37

5 Discussion 41

**6 Conclusion and Future Work** **48**

**Bibliography** **50**

# 1 Introduction

## 1.1 Introduction

The availability of publicly accessible digital healthcare datasets has significantly accelerated the development of Artificial Intelligent (AI) systems [15]. These systems have demonstrated the potential to support clinicians by improving diagnostic accuracy and facilitating timely interventions, ultimately leading to improved patient outcomes [11]. However, recent studies have shown concern that AI systems may fail to benefit all populations equally and have advocated for the development of responsible AI systems in healthcare [11, 13, 14, 15].

At the core of this call for responsibility are issues related to fairness, transparency, and accountability, which have been identified as persistent challenges in healthcare AI systems [19]. Fairness issues often arise due to various form of bias, including dataset, label, and measurement bias [8]. Transparency challenges may result from black-box models that are difficult to interpret [13], as well as from limited visibility into the entire model development life cycle [21]. Accountability is weakened when it is difficult to trace decisions and actions taken before, during and, after model development [19].

This highlights the importance of understanding how fairness, transparency, and accountability are interact and connected [2]. When bias is present, the principle of fairness is inherently violated, as certain populations may be systematically disadvantaged [14]. The identification and mitigation of such bias would be easy on transparency, which enables the examination of dataset patterns, feature importance and model behaviour [2]. Accountability complements these principles by ensuring that biased outcomes can be traced back to specific data, design choices, or deployment contexts [5]. They are mutually reinforcing components of responsible AI system [2].

In order to operationalize these principles in practice, evaluation and auditing of AI systems are essential [15]. In high-stakes domains like healthcare, where data often reflects deeply rooted societal disparities, the absence of through evaluation and auditing can lead to unintended and ethically problematic consequences [19]. Therefore, evaluating fairness, transparency, and accountability across the entire machine learning pipeline is critical to ensuring responsible AI in healthcare benefits all patient populations equitably [28].

Previous evaluations of responsible AI in healthcare have often considered fairness, accountability, and transparency separately, focusing on single principles or narrow components of the machine learning pipeline [1, 15]. This fragmented approach limits understanding of how these principles interact and reinforce one another. This project addresses these gaps by auditing fairness, accountability, and transparency across the full machine learning pipeline using the FAT Forensics toolkit [22], which provides an integrated framework for evaluating all three principles.

## 1.2 Aim and Objectives

This project aims to explore potential biases, risks, and limitations that may affect equitable and trustworthy decision-making in data-driven AI systems. It focuses on evaluating and auditing the AI model pipeline with respect to core responsible AI principles: fairness, accountability, and transparency. The study emphasizes evaluation rather than optimisation of model performance. It will give awareness on the challenges and opportunities to support responsible AI development in healthcare. The findings aim to inform future efforts in establishing ethical practices for AI in healthcare and identify areas where further research is needed to enhance responsible AI development. The objectives of the project are:

- To adapt and tailor [15] methodology to the MIMIC-IV dataset.
- To extend the evaluation methodology, using FAT Forensics toolkit for a comprehensive assessment of fairness, accountability, and transparency.
- To systematically evaluate and audit the machine learning (ML) pipeline, for alignment with core responsible AI principles fairness, accountability, and transparency.
- To investigate the interrelationships between fairness, accountability, and transparency, how reinforcing or conflicting effects emerge in practice.

The remainder of this document is structured as follows. Chapter 2 provides a review of relevant literature, including key principles of Responsible AI, existing frameworks for evaluating fairness, accountability, and transparency (FAT), and prior evaluations conducted on the MIMIC dataset. Chapter 3 describes the methodology employed for assessing FAT, including details on cohort selection, feature engineering, and the evaluation methods. Chapter 4 presents the results of the analysis across the machine learning pipeline. Finally, Chapter 5 concludes the study, summarizing key findings and outlining potential directions for future work.

# 2 Literature Review

## 2.1 Introduction

This section reviews existing literature on the principles of Responsible AI, with a particular emphasis on Fairness, Accountability, and Transparency (FAT). It reviews definitions, operationalisation and applications of these principles in healthcare. The review also highlights, evaluation done on MIMIC database, common methods and metrics used to assess FAT principles, providing a foundation for the subsequent analysis conducted in this project.

## 2.2 Responsible AI in Healthcare

With the rapid expansion of AI systems in healthcare, questions have been raised about the equality, trustworthiness and responsibility of such systems [5]. In response to these questions, the concept of Responsible AI has emerged as a framework for guiding the development and deployment of AI technologies in a manner that is safe, effective, and aligned with those concerns [19].

Responsible AI in healthcare is a multifaceted concept that encompasses the ethical, legal, and social implications of developing and deploying AI technologies [27]. It aims to ensure that AI systems are not only technically proficient but also promote equity and safeguard patient well-being [19]. The definition often revolves around a set of core principles that guide the design, development, and implementation of AI. Although, no universally accepted framework exists, the three commonly emphasized concepts are, Fairness, Transparency, and Accountability [19].

### 2.2.1 Fairness

Fairness refers to the equitable treatment of all individuals, ensuring that AI systems do not reinforce or exacerbate existing health disparities [19]. It has been emphasized that AI models should be trained on appropriately representative datasets to reduce algorithmic bias and support accurate clinical predictions across diverse patient populations [1]. Without such considerations, the risk of discriminatory outcomes, particularly for historically marginalized groups remains significant [4]. Fairness in AI is a complex and multifaceted concept, with various definitions and technical approaches to its evaluation. There are three prominent notions of fairness: group fairness, fairness through unawareness, and individual fairness [27].

**Fairness Through Unawareness**: is a straightforward approach to achieving fairness in AI. The core idea is that if an AI system is not explicitly given access to sensitive attributes, it cannot discriminate based on them [8]. The assumption is that by being unaware of these attributes, the model will treat all individuals equally, regardless of their group membership [27]. Despite its simplicity, fairness through unawareness is ineffective in practice. The primary limitation is the issue of redundant encodings or proxies. Sensitive attributes are often highly correlated with other, non-sensitive attributes in the data. Even if the sensitive attribute itself is removed, the AI model can still learn to discriminate based on these proxies, leading to biased outcomes [6]. Furthermore, in some cases, being aware of sensitive attributes can be necessary to correct for existing biases. For instance, in medical diagnoses, knowing a patient's sex or ethnicity might be crucial for accurate risk assessment. In such scenarios, enforcing unawareness could lead to less accurate and potentially harmful outcomes for certain groups [8].

**Group Fairness**: focus on ensuring that an AI system's outcomes are equitable across different subgroups. It involves partitioning the population into groups based on sensitive attributes and then requiring that certain statistical measures of performance are equal or similar across these groups [14]. Common group fairness metrics include equal accuracy, equal opportunity and demographic parity [8]. Group fairness metrics are statistically defined and can be directly measured and optimized, making them appealing for technical implementation and evaluation [4]. While group fairness metrics are easier to quantify, satisfying one often compromises another, particularly when outcome base rates differ significantly across groups [1].

**Individual Fairness**: focus on similar individuals should be treated similarly. If two individuals are similar with respect to a particular task or outcome, an AI system should produce similar outcomes for them, regardless of their membership in any protected group. Evaluating individual fairness is more challenging than group fairness because it requires defining a similarity metric between individuals [8]. The challenge lies in defining what constitutes similarity between individuals, which often requires a task-specific metric or distance function. It can be computationally intensive to enforce, and the definition of similarity can be subjective and difficult to operationalize in real-world scenarios [8]. Additionally, achieving perfect individual fairness might sometimes conflict with group fairness objectives, highlighting the inherent trade-offs in fairness definitions [6]. In clinical applications, balancing group and individual level fairness is challenging but essential for responsible AI system.

### 2.2.2 Transparency

Transparency can be defined from two aspects. Firstly, it refers to the interpretability of algorithmic decisions, which is critical for fostering trust in healthcare [19]. Secondly, it encompasses process transparency, which emphasizes visibility into the broader machine learning pipeline, including data preprocessing, feature engineering, and design decisions [21]. Together, these dimensions ensure that both the internal logic of models and the external steps of their development are open to validate. Model interpretability can be further categorized into global and local interpretability methods. Global methods aim to explain model behaviour across an entire dataset, while local methods focus on interpreting individual predictions [13].

**Global Explanation** methods provide an overall understanding of how a predictive model makes decisions across an entire dataset. They reveal the importance and influence of features on model predictions at the population level, helping to identify general patterns, key drivers, and potential biases embedded in the model [5]. These techniques summarize complex model behaviour into interpretable forms, which is critical for ensuring transparency, trust, and accountability in AI systems in healthcare. Global explanation techniques, such as Permutation Feature Importance and Partial Dependence Plots, offer population level insights into feature importance [13].

**Local Explanation** focus on interpreting individual predictions rather than global trends. They aim to identify which features most influenced a specific model decision, providing insight into why a particular prediction was made. Local methods like SHAP and LIME explain individual predictions, enabling clinicians to understand specific decision pathways [13].

### 2.2.3 Accountability

Accountability in AI involves the traceability of decision processes and the ability to attribute responsibility in cases of failure. This principle includes the requirement that systems be designed with mechanisms for tracing actions, identifying failures, and assigning responsibility [19]. In the context of healthcare, accountability must be implemented both technically and institutionally to ensure that AI systems operate reliably, ethically, and under meaningful oversight [27]. Practical approaches to accountability can be organized into three categories, ethical, technical, and societal accountability [5].

**Ethical accountability**: emphasizes alignment between the AI system behaviour and ethical values, such as fairness, safety, and respect for patient autonomy. Techniques in this category include ethical impact assessments, which evaluate potential harms and benefits before deployment, and stakeholder engagement, which integrates perspectives from clinicians, patients, and ethicists throughout the development lifecycle [21].

**Technical accountability**: involves building systems that are auditable and explainable. Common techniques include logging and audit trails, which capture decision pathways for post-hoc analysis, and algorithmic auditing, which systematically evaluates model behaviour against predefined fairness and performance criteria [5].

**Societal accountability**: focuses on external mechanisms that reinforce responsibility through regulation and institutional oversight. This includes compliance with legal frameworks, as well as public–private partnerships that promote shared standards, evaluation protocols, and cross-sectoral collaboration [21].

Together, these principles contribute to a responsible AI framework. Importantly, the principles should not be treated as a static requirement but as a dynamic process that evolves alongside technological, clinical, and regulatory developments [27]. In healthcare, where the consequences of AI errors can be significant, maintaining traceability, auditability, and clear lines of responsibility is essential to building trust and ensuring long-term system integrity [19]. However, the practical implementation of these responsible AI principles remains an ongoing challenge. Operationalizing these principles will require interdisciplinary collaboration, adaptive regulatory mechanisms, and context-sensitive system design [21].

## 2.3 MIMIC Dataset Evaluation

Multiple studies have reported performance disparities in models trained on MIMIC, particularly when evaluated across sensitive attributes such as ethnicity, gender, age, and insurance status (often used as a proxy for socioeconomic status) [1, 15]. For instance, [11] explored fairness in predicting ICU length of stay using MIMIC-IV. They conducted an ANOVA test to explore differences in ICU metrics among sensitive groups and evaluated model performance using Area under the curve (AUC)-based metrics.

Similarly, [13] conducted a study on interpretability and fairness in deep learning models using MIMIC-IV. They applied interpretability techniques such as DeepLIFT, feature ablation, and ArchDetect, while fairness was evaluated using AUC-based performance metrics across subgroups. In another study, [15] evaluated the generalizability and fairness of a benchmark LSTM model for in-hospital mortality prediction using MIMIC-III. Their analysis employed a multi-phase validation framework, incorporating both internal and external evaluations. Fairness was examined using AUC-based metrics and descriptive statistics across sensitive sub-populations.

These studies have reported statistically significant variations in model performance across gender, ethnicity, and insurance status [11, 13, 15]. Although MIMIC is one of the most extensively used datasets for AI applications in healthcare, ethical and responsible use of the dataset has not been widely explored [1]. Existing studies often focus on individual responsible AI dimensions, such as fairness, or transparency in isolation, with accountability comparatively less explored. Furthermore, the lack of standardisation in evaluation metrics adds another layer of complexity, as different studies adopt varied criteria and benchmarks [4]. This inconsistency not only hampers comparability across findings but also raises concerns about the reproducibility and reliability of ethical evaluations [19].

A deeper understanding of MIMIC is particularly important, as it serves as a representative use case for publicly available electronic health record (EHR) data and has gained substantial interest from the research community [15]. So, thoroughly evaluating MIMIC using a unified framework is essential for developing responsible AI models in healthcare [14]. Current studies vary widely in their evaluation methodologies, which can lead to inconsistent or biased conclusions [19]. Reproducible and standardized assessment is crucial for ensuring fairness, transparency, and accountability across machine learning pipelines [22]. Such evaluations can aid in identifying and mitigating ethical risks associated with the deployment of responsible AI in clinical practice [4].

## 2.4 Tools and Frameworks for Evaluating Responsible AI

Several methodologies and frameworks have been proposed to assess data quality and model performance with respect to responsible AI principles [25, 26, 3, 22]. These tools aim to make ethical evaluation more systematic and reproducible.

Fairness evaluations encompass a range of techniques, including subgroup performance comparison using AUROC, AUPRC, confusion matrices, F-statistics, accuracy, descriptive statistics, and visualizations [14]. To facilitate these evaluations, several tools and packages have been developed such as Fairlearn [25], AI Fairness 360 [3], and What-If-Tool [26]. Transparency or interpretability is typically evaluated through methods like SHAP, LIME, and various visualization techniques [13]. In contrast, accountability remains less well-quantified and is often fostered through more qualitative practices such as rigorous validation protocols, data documentation practices like datasheets, and the involvement of experts [19]. While these tools contribute significantly to ethical AI development, they often focus on only one aspect, such as fairness or interpretability, and tend to operate in isolation from the other concerns.

In contrast, FAT Forensics is a comprehensive Python toolbox that provides a unified evaluation framework for fairness, accountability, and transparency across the entire machine learning pipeline. It supports multiple methods including data description, disparate impact, sampling bias, feature influence, counterfactual and surrogate explainer [22].

FAT Forensics distinguishes itself by providing an integrated framework to evaluate fairness, accountability, and transparency throughout the machine learning pipeline. Unlike many existing tools that address these aspects in isolation, FAT Forensics provides a holistic approach, enabling more comprehensive and reproducible evaluations. Its modular API and support for different metrics make it a versatile and practical solution for assessing responsible AI development.

# 3 Methodology

## 3.1 Introduction

This section describes the methodology employed to evaluate responsible AI principles. The section begins with an explanation of the methods used for data selection and preprocessing. Feature engineering techniques are described, followed by an overview of the predictive algorithms selected to explore varying levels of model complexity and underlying assumptions. Finally, the evaluation methodology applied throughout the machine learning pipeline is presented.

## 3.2 Dataset and Preprocessing

The Medical Information Mart for Intensive Care IV (MIMIC-IV), a publicly available and de-identified electronic health record (EHR) dataset, was utilized in this study. MIMIC-IV contains comprehensive clinical data from over 60,000 intensive care unit (ICU) admissions at the Beth Israel Deaconess Medical Center, covering the period from 2008 to 2019 [10]. The dataset includes a wide range of structured data elements, such as patient demographics, vital signs, laboratory test results, medication administrations, procedures, and clinical outcomes [14]. This diversity of clinical information makes MIMIC-IV well-suited for machine learning tasks such as mortality prediction, length-of-stay estimation, and decompensation [9].

According to [14], in-hospital mortality is the most commonly studied prediction task using MIMIC database. In this study, **MIMIC-IV** was selected due to its more recent data, comprehensive documentation and public availability, all of which support reproducibility. All analyses were conducted using version 3.1 of MIMIC-IV, which was the latest version available at the time of access. The prediction task evaluated in this work was in-hospital mortality.

### 3.2.1 Cohort Preparation

The study cohort was constructed following methodologies established in prior research on fairness and interpretability using the MIMIC database [15, 13]. These inclusion criteria were ensure to minimize variability in both the quantity and quality of data across individuals. To ensure methodological consistency and to support subsequent analyses, the following inclusion criteria were applied to the MIMIC-IV dataset:

- **First ICU stay per patient:** only the first ICU admission for each patient was included to prevent information leakage resulting from multiple admissions [13]. This restriction was implemented to reduce potential confounding effects associated with prior ICU exposure, which may vary across demographic or socioeconomic groups.

- **ICU stay duration between 48 hours and 7 days:** ICU stays shorter than 48 hours were excluded, as they may represent atypical scenarios such as early discharges, transfers, or early mortality. Stays exceeding 7 days were also excluded, as they are often indicative of high clinical complexity or chronic conditions. The removal of these extremes was intended to reduce cohort heterogeneity and minimize potential bias stemming from the uneven distribution of such cases across subgroups.

- **Standardized observation window (first 48 hours):** for all selected patients, clinical data were restricted to the first 48 hours of the ICU stay. This standardization ensured consistent data availability across patients, thereby facilitating fair comparisons throughout the machine learning pipeline [15].

### 3.2.2 Feature Selection

Feature selection was guided by both clinical relevance and data completeness, with the goal of building fair, interpretable, and reproducible models [15]. Both time-series clinical variables and descriptive features from the MIMIC-IV was selected.

**Clinical Time-Series features**: this study adopted the feature set used in prior study [15], selecting 17 variables based on their availability and alignment with features in the MIMIC-IV database. When specific features from their study were not available in MIMIC-IV, similar substitute variables were selected from the PhysioNet 2012 Challenge feature set [20]. This approach preserved consistency with earlier studies while adapting to the structure and limitations of the MIMIC-IV dataset. To ensure data quality and reduce the risk of bias due to excessive missingness, a coverage threshold was applied; only features that have at least 40% record for a patient were retained. This threshold was chosen to balance data availability with clinical relevance. Features with incomplete data are more prone to imputation bias, which may disproportionately affect under-represented subgroups due to inconsistent data collection practices [4].

**Demographic Features**: in addition to physiological variables the following features were included; race, gender, age, marital status, insurance type, admission type, and language [13]. Among these **race**, **gender**, and **insurance** were treated as sensitive or protected features given their relevance to the evaluation [15].

Insurance data in the dataset primarily consisted of three major types, the public programs Medicare and Medicaid, and private insurance. These were recategorized into four distinct groups; Medicare, Medicaid, private, and other, to enable consistent subgroup analysis. The other category was included to account for less common or ambiguous entries such as self-pay, unknown, or government programs not clearly classified, which appeared infrequently but could not be reliably assigned to the main categories [15].

Race entries in the raw dataset were highly heterogeneous. To standardize these, the following encoding scheme was applied, patients identified as White and of non-Hispanic ethnicity were encoded as White; Asian and Black patients were encoded independently as Asian and Black, respectively; and all patients of Hispanic origin were grouped as Hispanic. Patients from other racial backgrounds, such as American Indian and Pacific Islander, were categorized under Other, following [15].

### 3.2.3 Feature Aggregation

To capture temporal patterns within the 48-hour observation window, statistical summaries were computed for each time-series feature across seven predefined temporal segments. These segments included the entire 48-hour window, as well as the first 10%, 25%, and 50%, and the last 50%, 25%, and 10% of the observation period [9]. This aggregation approach allowed the model to capture the full trends in patient stay.

For each continuous time-series feature, the minimum, maximum, mean, standard deviation, skewness, and the number of recorded measurements were computed within each segment [9]. For categorical variables, the minimum, maximum, mode, and count of observations were calculated. If a segment contained no recorded measurements for a given feature, all corresponding summary statistics for that segment were set to missing [9]. To address missingness, imputation was performed using summary statistics derived from the training set, mean values for continuous features and mode values for categorical features [9]. Following imputation, all features were standardized by subtracting the mean and dividing by the standard deviation from the training dataset [9]. All categorical features were then processed using one-hot encoding to convert them into a format suitable for machine learning models.

## 3.3 Algorithms

To predict in-hospital mortality, a set of machine learning algorithms was employed, to investigate how model architecture and underlying assumptions influences the behaviour of responsible AI systems.

**Logistic Regression (LR)**: is a linear classification model that estimates the probability of a binary outcome using the logistic sigmoid function [24]. It assumes a linear and additive relationship between the predictors and the outcome [23]. The simplicity and interpretability of LR make it a popular choice for healthcare applications [24]. However,its inability to capture complex, non-linear feature interactions may limit its performance, particularly in heterogeneous populations [24]. In scenarios where patient subgroups exhibit distinct risk patterns or interaction effects, LR may under-represent subgroup specific relationships, potentially resulting in disparities in the performance.

**Support Vector Machine**: is a supervised learning algorithm that identify an optimal decision boundary to separate outcome classes by maximizing the margin between them [24]. When the data is not linearly separable, kernel functions such as the radial basis function (RBF) are used to project the input data into a higher-dimensional space, where a linear separation becomes feasible [24]. As a result, SVMs are particularly well-suited for clinical datasets where the relationships between features and outcomes may be irregular or non-additive. However, it also introduces challenges, kernel transformations can amplify or suppress subgroup-specific patterns in unpredictable ways [23]. This may lead to variable model performance across demographic or clinical subgroups.

**Linear Discriminant Analysis (LDA)** is a statistical classification method that aims to separate outcome classes by modelling the distributional structure of the input data [24]. It assumes features for each class are drawn from multivariate Gaussian distributions with class specific means but a shared covariance matrix [28]. The model constructs a linear combination of input features that maximizes the ratio of between class variance to within-class variance. LDA performs well when these assumptions hold, deviations in subgroup variance or feature distribution can compromise model performance [24].

**Naive Bayes** is a probabilistic classification algorithm that applies Bayes theorem under the assumption that input features are conditionally independent given the class label [28]. While this independence assumption rarely holds in clinical data, it allows the model to efficiently compute the joint likelihood of feature values by multiplying individual feature likelihoods[23]. Naive Bayes is computationally efficient, scalable to high-dimensional data, and particularly useful in healthcare settings where many variables are recorded but may be sparsely observed [24]. However, its assumption of feature independence may overlook important interactions or correlations between variables [24]. If the model fails to capture these relationships, it may inadvertently favour patterns dominant in the majority population, leading to systematic misclassification or unequal performance across subgroups.

**Random Forest** is an ensemble learning method that constructs multiple decision trees and aggregates their outputs to produce a final prediction. Each individual tree is trained on a random subset of the data and selects a random subset of features at each split, a strategy that introduces variation across the trees and enhances generalization [23]. The forest then combines the predictions of all trees, to generate a more stable and accurate overall classification. This approach effectively captures complex, non-linear interactions and robust to missing data, making it well-suited for high-dimensional clinical datasets [24]. However, Random Forest directly learns patterns from the training data, any imbalances or under-representation of particular patient subgroups can bias the learned decision boundaries [24]. This can result in uneven performance if certain populations are not adequately reflected in the data.

**Extreme Gradient Boosting (XGBoost)** is a gradient boosting framework that builds decision trees sequentially, where each new tree is trained to reduce the residual errors of the ensemble built so far [23]. Unlike Random Forest's independently building trees, XGBoost iteratively optimizes a specified loss function via gradient descent, allowing it to learn complex, non-linear relationships and achieve high predictive accuracy [28]. While this iterative refinement enhances model expressiveness, it also increases the risk of over-fitting to subgroup-specific noise or disparities, particularly if some patient populations are under-represented or exhibit distinct feature interactions.

**k-Nearest Neighbors (kNN)**: is a non-parametric, instance-based learning algorithm that predict outcomes by identifying the k most similar patients in the training set, based on a chosen distance metric such as euclidean distance [28]. Because k-Nearest Neighbors (kNN) does not construct an explicit predictive function during training, its predictions are directly influenced by the distribution and structure of the training data [28]. This makes kNN intuitive and locally interpretable, as individual predictions can be traced to specific patient cases. However, its performance is sensitive to the choice of k, distance metric, and feature scaling [23]. In clinical contexts where data are often high-dimensional and unevenly distributed across subgroups, under-represented minority groups may lack sufficiently similar neighbours, resulting in less reliable predictions.

The selected models range from Linear classifiers, such as Logistic Regression, Linear Discriminant Analysis, Naive Bayes, and Support vector machines, to non-linear approaches, including Random Forest, Extreme Gradient Boosting, and k-Nearest Neighbors. The use of these diverse algorithms enables an examination of how varying model assumptions influence responsible AI development.

## 3.4 Evaluation of the Dataset

The raw measurement patterns within the selected cohort were assessed to identify differences in data availability, completeness, and distribution across patient subgroups. This allowed for the evaluation of disparities in the underlying data. Three complementary methods were employed; **Measurement Frequency**, **Missingness Rate**, and **Monitoring Ratio**. Measurement Frequency and Missingness Rate were calculated for the raw time series data before aggregation and preprocessing, while Monitoring ratio was computed using the final preprocessed dataset.

**Measurement Frequency**: was defined as the total number of non-missing observations for each feature during the first 48 hours of the ICU stay. For each patient, the number of recorded measurements per feature was counted and then aggregated across predefined subgroups. This metric reflected the intensity of clinical monitoring and the volume of information available for predictive models [18].

**Missingness rate**: was defined as the proportion of hourly time intervals during an ICU stay without recorded measurements. Each ICU stay was divided into hourly intervals and the missingness rate was calculated as the fraction of these intervals without any recorded data. This temporal approach highlights gaps in monitoring coverage that may disproportionately affect certain subgroups, potentially influencing the reliability of the data for modelling [18].

**Monitoring ratio**: to evaluate the frequency of clinical measurements across subgroups, the monitoring Ratio was computed as the ratio of the average number of observations per variable for a given subgroup relative to a baseline group. Baseline groups were defined based on the largest sample sizes in the dataset; Medicare for insurance, White for race, and male for gender.

For each subject, the number of recorded measurements for each clinical variable was counted, resulting in a total count of observations per subject per variable. Then, the mean total count of measurements across all subjects within each subgroup was calculated. The monitoring ratio was then obtained as the mean total count for a given subgroup divided by the mean total count for the baseline group. A ratio below one indicate less frequent monitoring compared to the reference group, whereas a ratio above one show more frequent monitoring. As a relative metric, the monitoring ratio helps in the identification of subgroup-level differences in clinical monitoring intensity.

**Systemic Bias**: a data record was considered to exhibit systemic bias, if it shared the same value for all non-protected features with another record but differed in the associated target (ground truth), solely due to a difference in the protected attribute [22]. It helps to identify biases stemming from data collection or labelling processes, rather than genuine clinical differences.

To evaluate the presence of systemic bias in the dataset, a two-step method was implemented: (1) detection of systemic bias using the FAT-Forensics framework [22] and (2) similarity validation via Dynamic Time Warping (DTW) [17]. The direct application of FAT-Forensics to continuous features posed challenges, as identifying similar pairs is less straightforward due to difficulties in defining similarity for continuous data. To address this, continuous variables were discretized into categorical bins, enabling feature similarity comparisons.

To ensure that the matched pairs identified in step (1) were not artifacts of binning or aggregation, similarity validation was performed using **Dynamic Time Warping (DTW)** on the raw time-series data. DTW enables the comparison of time series by stretching or compressing the time axis to align similar patterns [17]. DTW distances were computed between matched records to evaluate whether the pairs were indeed similar in their temporal pattern. This additional validation step strengthened the reliability of the systemic bias detection process.

## 3.5 Evaluation of Model Performance and Predictions

Following the training of predictive models for the in-hospital mortality task, a set of evaluation metrics was applied to assess performance across patient subgroups. These metrics were selected to quantify both overall predictive accuracy and the equitable distribution of performance across protected groups.

**Area Under the Receiver Operating Characteristic Curve (AUROC)**: was used as a global measure of model discrimination, summarizing the trade-off between sensitivity (true positive rate) and specificity (false positive rate) across all classification thresholds [6]. Although AUROC is not a fairness metric, it provides insight into the model's ability to rank patients according to risk, and can help identify disparities in predictive discrimination [15].

**Equal Accuracy**: was used to assess whether classification performance was balanced across different subgroups. For each subgroup g, accuracy was calculated as:

$$Accuracy_g = \frac{TP_g + TN_g}{TP_g + TN_g + FP_g + FN_g} \tag{3.1}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives respectively [8]. Large variations in accuracy across subgroups may indicate unequal error distributions, raising concerns about fairness, accountability, and the potential for disproportionate harm.

**Equal Opportunity**: requires the true positive rate (TPR), also known as sensitivity or recall, to be equal across different subgroups defined by a sensitive attribute [22]. Formally, for sensitive attribute A and outcome Y, equal opportunity is defined as:

$$\Pr(\hat{Y} = 1 \mid Y = 1, A = a) = \Pr(\hat{Y} = 1 \mid Y = 1, A = b) \tag{3.2}$$

This ensures that the probability of correctly identifying individuals with a positive outcome is consistent across demographic groups. In clinical settings, satisfying equal opportunity is crucial to avoid systematically overlooking high-risk patients in specific subgroups [6].

**Demographic Parity**: also known as statistical parity, requires that the probability of a positive prediction is equal across different demographic groups [8]. Formally, a predictive model satisfies demographic parity if:

$$P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b) \tag{3.3}$$

for all groups $a, b$, where $\hat{Y}$ is the predicted outcome and $A$ represents the sensitive attribute. This implies that the model's positive prediction rate should be independent of the sensitive attribute, ensuring that no group is systematically favoured or disadvantaged by the model's predictions [6]. While demographic parity is useful for identifying group-level prediction outcomes, it may conflict with clinical validity if baseline outcome rates (e.g., mortality rates) differ substantially across groups.

**Permutation feature importance**: is a model-agnostic technique used to identify the contribution of each feature to the predictive performance of a model [13]. The method involves randomly shuffling the values of a single feature across the dataset, thereby breaking the relationship between that feature and the target variable, while keeping other features intact [16]. The resulting decrease in model performance measured by metrics area under the curve is interpreted as the importance of the shuffled feature. Features that cause a significant drop in performance when permuted are considered more important for the model's predictions [16].

**Counterfactual fairness**: ensures that a model's prediction for an individual remains unchanged in a hypothetical scenario where only the protected attribute is altered, while all other relevant features remain constant. The core idea behind counterfactual fairness is that a model should produce the same prediction for an individual, even when sensitive attributes are hypothetically changed. A prediction is considered fair, if it would be the same in a counterfactual world where the individual belonged to a different demographic group [12]. Predictor $\hat{Y}$ is counter-factually fair if under any context $X = x$ and $A = a$,

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a), \tag{3.4}$$

for all $y$ and for any value $a'$ attainable by $A$.

# 4 Results

## 4.1 Introduction

This section describes the results of the evaluation carried out. The results are structured to align with the methodological steps outlined previously, beginning with the characteristics of the selected cohort and data preprocessing. The performance of various predictive models across different metrics is presented.

## 4.2 Cohort Preparation

The study cohort consisted of 44,421 patients after applying the cohort selection criteria 3.2.1. Table 4.1 summarizes patient distributions across different subgroups. Medicare patients represented the largest group (24,159) with a mortality rate of 12.0%, followed by Private (12,307; 6.5%), and Medicaid insurance (6,068; 8.7%). With respect to race, the cohort was predominantly White (29,864), with a mortality rate of (9.1%). Minority groups showed slightly different rates, Black patients (9.0%), Asian (11.4%), and Hispanic (8.3%). The gender distribution was relatively balanced, with 25,258 males (mortality rate: 9.5%) and 19,163 females (10.8%).

**Feature Selection** Based on the feature selection mentioned in 3.2.2, 14 clinical features were selected for the model development. Table 4.2 below presents the selected feature. Features with less than 40% coverage (highlighted) were excluded from the

Table 4.1: Cohort distribution across protected features.

| **Group** | **Category** | **Total Patients** | **Deaths** | **Survived** | **Mortality (%)** |
|---|---|---:|---:|---:|---:|
| Total | – | 44,421 | 4,460 | 39,961 | 10 |
| Insurance | Medicaid | 6,068 | 530 | 5,538 | 8.7 |
| | Medicare | 24,159 | 2,909 | 21,250 | 12.0 |
| | Other | 1,887 | 220 | 1,667 | 11.7 |
| | Private | 12,307 | 801 | 11,506 | 6.5 |
| Race | Asian | 1,332 | 152 | 1,180 | 11.4 |
| | Black | 3,904 | 353 | 3,551 | 9.0 |
| | Hispanic | 1,587 | 131 | 1,456 | 8.3 |
| | Other | 7,734 | 1,110 | 6,624 | 14.4 |
| | White | 29,864 | 2,714 | 27,150 | 9.1 |
| Gender | Female | 19,163 | 2,073 | 17,090 | 10.8 |
| | Male | 25,258 | 2,387 | 22,871 | 9.5 |

dataset.

## 4.3 Dataset Evaluation

In ICU, vital signs are consistently and frequently monitored, whereas laboratory measurements are less often and typically based on clinical necessity [18]. Therefore, the evaluation done on the dataset, was limited to vital sign features, including Heart Rate, Glasgow Coma Scale (GCS) components (Eye, Verbal, and Motor responses), Respiratory Rate, Temperature, Systolic Blood Pressure, and Fraction of Inspired Oxygen.

### 4.3.1 Monitoring Frequency

Patients covered by Medicare received more frequent monitoring compared to other insurance groups. Heart rate and respiratory rate measurements were the one with high monitoring frequency in all insurance groups. Patients identified as White had the highest measurement frequencies for nearly all vital signs, with heart rate, respiratory rate and systolic blood pressure being the most frequently recorded. In contrast, Asian and Hispanic patients generally have lower measurement frequencies. Both male and female patients showed high monitoring frequency, particularly for Heart Rate, Respiratory Rate and Systolic Blood Pressure. These variations may be influenced by differences in sample size among racial and insurance groups, as Medicare and White patients constituted the largest subgroups within the cohort.

### 4.3.2 Missingness Rate

Missingness rates were generally higher among patients who survived and decrease for those with poorer outcome. Private insurance patients who survived have highest missingness rates compared to other groups, 21% for heart rate, 23% for respiratory rate, and 47% for systolic blood pressure. However, among patients with poorer outcomes,

Table 4.2: Coverage of 17 features selected from MIMIC-IV database. Features with less than 40% coverage (highlighted in bold) were excluded.

| **Feature** | **Coverage (%)** |
|---|---:|
| Heart Rate | 99.94 |
| GCS Eye | 99.90 |
| GCS Verbal | 99.89 |
| GCS Motor | 99.89 |
| Respiratory Rate | 99.73 |
| Sodium | 98.53 |
| Creatinine | 98.53 |
| BUN | 98.50 |
| Hematocrit | 98.15 |
| Temperature (F) | 97.20 |
| NIBP Systolic | 89.80 |
| $FiO_2$ | 49.79 |
| pH (Arterial) | 44.24 |
| $PaO_2$ | 43.98 |
| **Arterial BP Diastolic** | **40.42** |
| **Arterial BP Systolic** | **40.40** |
| **Potassium** | **30.55** |
| **Glucose** | **30.20** |
| **ART BP Diastolic** | **5.60** |
| **Direct Bilirubin** | **3.38** |

these rates declined to levels comparable to, or slightly lower than, those observed in other insurance groups (see Figure 4.1). For example, the missingness rate for heart rate decreased from (21% to 14%), and for respiratory rate from (23% to 14%).

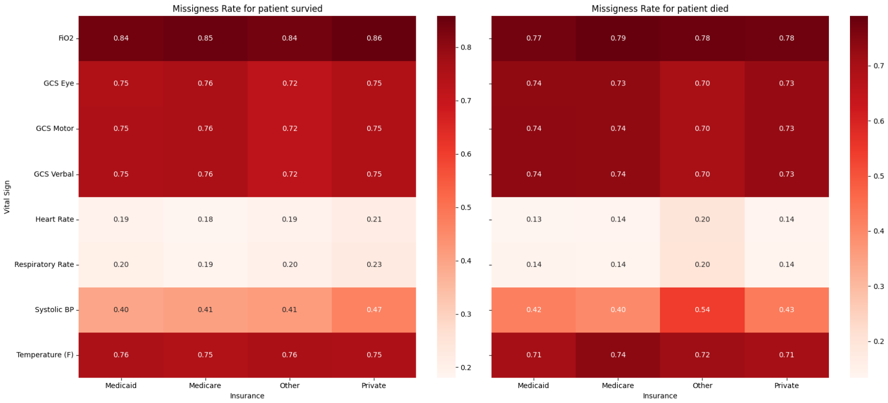

Figure 4.1: Rate of missing vital sign data per insurance subgroup.

A similar pattern was observed across racial groups, with missingness rates generally decreased among patients with poorer outcomes. For example, systolic blood pressure missingness decreased from (42% to 35%) for Asian and from (43% to 39%) for White patients. In contrast, Black and Hispanic patients showed a reverse pattern, with systolic blood pressure missingness increased from (36% to 38%) for Black and from (42% to 44%) for Hispanic patients. Also, no change in the missingness rate was observed for Glasgow Coma Scale (GCS) component features among Black patients, indicating consistent monitoring regardless of patient outcome (see Figure 4.2). For both male and female patients the missingness rate declined with poorer outcomes. Additionally, female patients consistently had lower missingness rate compared to males when outcomes were favourable, which may be related to their higher actual mortality rate (10.8%) compared to males (9.5%).

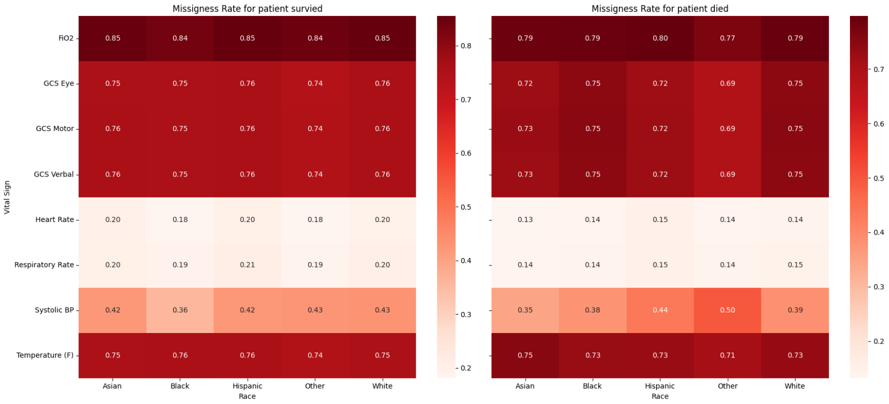

Figure 4.2: Rate of missing vital sign data per race subgroup.

### 4.3.3 Monitoring Ratio

Monitoring ratios for GCS component features were slightly higher among Medicaid and private insurance patients with favourable outcomes, approximately (1.03) compared to the baseline Medicare group. For private patients, this ratio sustained even among those with poorer outcomes. However, for other features, private and Medicaid patients exhibited lower monitoring ratios than the Medicare baseline when outcomes were favourable. Private insurance patients showed reduced monitoring ratios for Systolic Blood Pressure (0.90), while Medicaid patients had a lower ratio for temperature (0.95). When the outcome was poorer, private patients have higher monitoring ratios than both Medicare and Medicaid groups, although systolic blood pressure still remained the lowest at (0.97 (see Figure 4.3).

Hispanic, Asian and Black patients exhibited slightly higher monitoring ratios for the Glasgow Coma Scale (GCS) components compared to the baseline White group patients with favourable outcome. For other vital signs, their monitoring ratios were lower than the White group. However, as shown in Figure 4.4 Black patients showed higher monitoring ratios for Fraction of Inspired Oxygen (1.05) and Systolic Blood

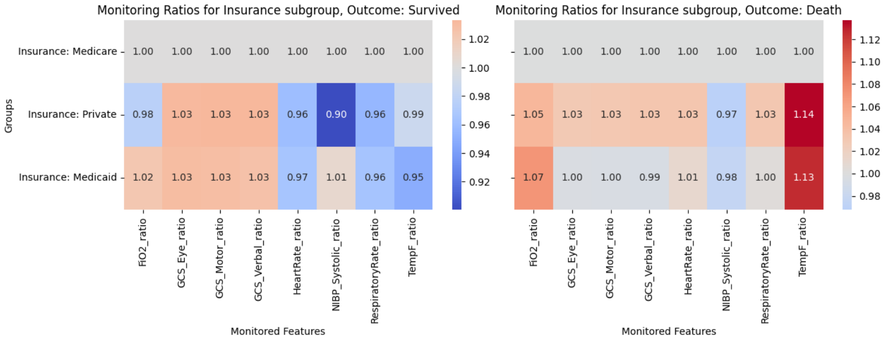

Figure 4.3: Monitoring Ratios of vital signs within the 48-Hour observation window across insurance subgroups.

Pressure (1.12) compared to both the baseline and other racial groups. Additionally, Hispanic patients had the second highest systolic blood pressure ratio (1.02) following black patients. When outcomes were poorer, monitoring ratio for GCS components feature remained higher for Asian and Hispanic patents compared to other groups. Asian patients also have higher systolic blood pressure than other groups (1.07). Conversely, Hispanic patients had the lowest monitoring ratio among racial groups, ranging from (0.91) for systolic blood pressure to (0.99) for respiratory rate. Black and white patients have slightly similar monitoring ratio when outcomes were poorer. Females generally had higher monitoring ratios than males for most vital signs and GCS features, except for Temperature monitoring which was slightly less frequent.

### 4.3.4 Assessing Systemic Bias in the Dataset

To enable comparisons across continuous clinical features with wide value ranges, each feature was discretized into clinically relevant bins. The binning approach was used to simplify the analysis by grouping continuous measurements into intervals. Initially, the minimum and maximum values for each feature were identified to capture their observed ranges within the dataset. For example, heart rate ranged from 40.0 to 162.6,

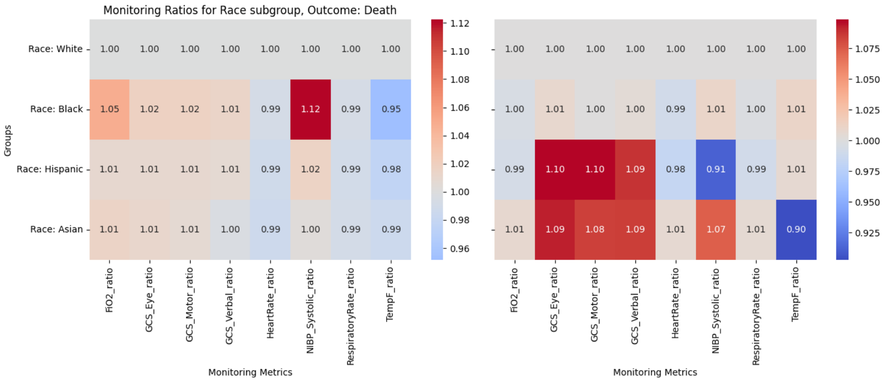

Figure 4.4: Monitoring ratios of vital signs within the 48-Hour observation window across race subgroups.

and respiratory rate from 5.0 to 43.2 (see Table 4.3). Based on these ranges, custom bin intervals were defined along with corresponding categorical labels, designed to reflect clinically meaningful thresholds. For instance, heart rate was divided into three bins, 40–60, 60–100, and 100–163 bpm, labelled as 1, 2, and 3, respectively. Through this process, comparisons across patient groups is possible.

A total of 3,995 subjects were identified with similar non-protected features but different observed outcomes across insurance types (see Figure 4.5). In examining the directionality of these observed outcome differences, 1,382 Medicare patients were matched with Private patients who shared similar non-protected features but had different outcome labels, Medicare patients were labelled as survived, while their Private counterparts were labelled as dead. Conversely, 2,013 Medicare patients were labelled as dead, whereas their matched Private counterparts were labelled as survived. Additionally, 1,695 Private patients were labelled as dead, while their matched Medicare counterparts were labelled as survived.

These differences suggest that insurance status alone may lead to substantial disparities in observed outcome labels, with the strongest disparities observed between Medicare and Private insurance groups. Although similar variations were observed involving

Table 4.3: Feature value ranges and corresponding binning intervals.

| Feature | Min – Max | Bins (Intervals) with Labels |
|---|---|---|
| Heart Rate | 40.0 – 162.6 | [40, 60), [60, 100), [100, 163] (1, 2, 3) |
| Respiratory Rate | 5.0 – 43.2 | [5, 12), [12, 20), [20, 44] (1, 2, 3) |
| Sodium | 120.0 – 160.0 | [120, 130), [130, 135), [135, 145), [145, 161] (1, 2, 3, 4) |
| Creatinine | 0.19 – 10.0 | [0.19, 0.6), [0.6, 1.2), [1.2, 10] (1, 2, 3) |
| BUN | 4.0 – 130.0 | [4, 7), [7, 20), [20, 130] (1, 2, 3) |
| Hematocrit | 15.0 – 55.0 | [15, 30), [30, 40), [40, 50), [50, 56] (1, 2, 3, 4) |
| Temperature | 95.0 – 103.7 | [95, 97), [97, 99), [99, 104] (1, 2, 3) |
| NIBP Systolic | 70.0 – 200.0 | [70, 90), [90, 120), [120, 140), [140, 201] (1, 2, 3, 4) |
| FiO2 | 10.0 – 100.0 | [10, 21), [21, 50), [50, 100] (1, 2, 3) |
| pH Arterial | 6.82 – 7.6 | [6.8, 7.35), [7.35, 7.45), [7.45, 7.7] (1, 2, 3) |
| PaO2 | 30.0 – 500.0 | [30, 60), [60, 75), [75, 100), [100, 501] (1, 2, 3, 4) |

Medicaid, the number of observed outcome differences was relatively smaller, ranging from 127 to 565 cases, indicating that the labelling bias is most significant within the Medicare–Private subgroups.

Overall, these results indicate a potential labelling bias in the dataset, where insurance status appears to influence outcome assignments independently of clinical features. This effect is most strong between Medicare and private insurance groups, with a tendency for labels to favour private patients. Specifically, switching to private insurance is more likely to result in a label change from dead to survived, while switching away from private is more likely to result in a change from survived to dead.

Potential labelling disparities were also observed across racial groups. A total of 833 subjects were identified with similar non-protected features but different observed outcomes across race. Most of these observed outcome differences occurred between Black and White patients, while the number of different observed outcomes involving

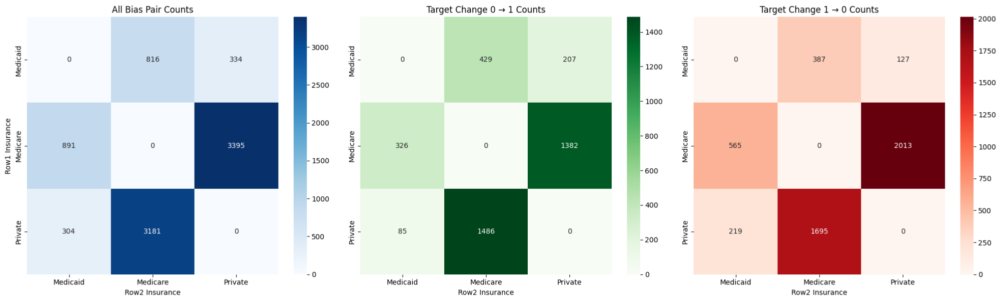

Figure 4.5: Labelling bias related to insurance groups.

Asian and Hispanic patients was smaller. Although, the magnitude of these differences is smaller than those observed for insurance, the result shows that outcomes may still vary across racial groups.

To ensure that the observed labelling disparities were not artifacts of data aggregation or binning, temporal similarity between matched pairs was computed at the raw time-series features using Dynamic Time Warping (DTW). The curve rises steeply (Figure 4.6), with over 80% of the distances falling below 50, indicating strong temporal similarity for the majority of matched pairs. This confirms that, across most features, matched patients had highly aligned clinical trajectories, strengthening confidence in the bias analysis.

The possible outliers identified in the dataset were Heart Rate (220045) and Partial pressure of arterial oxygen (PaO2) (220224), both of which exhibited higher DTW distances. Pa02, shows a much flatter curve, only about 20% of DTW distances are below 200, and some distances reach beyond 800. This wide spread suggests substantial variability in temporal patterns for PaO2, likely due to inconsistent measurement timing and broad value ranges observed in the dataset (30–500). The absence of a sharp rise in the CDF indicates a lack of consistent similarity in this feature across matched pairs. Heart Rate, lies between the two extremes. About 60% of DTW distances fall below 150,

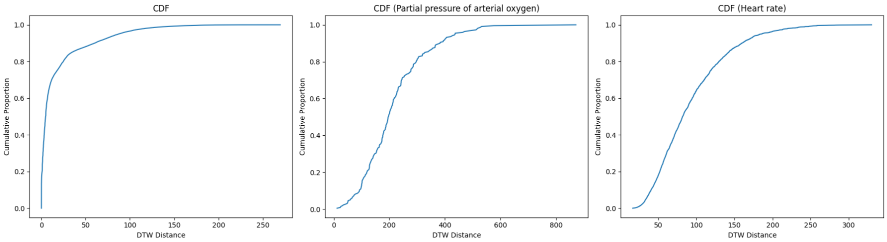

Figure 4.6: Cumulative distribution function (CDF) of Dynamic Time Warping (DTW) distances between matched patient pairs. Left: All features excluding PaO2 and Heart Rate. Middle: PaO2 only. Right: Heart rate only.

with a long tail extending past 300. This pattern reflects moderate similarity but with notable variation, possibly due to its wide clinical range (40–160 bpm).

Nonetheless, the overall analysis suggest that the majority of features preserved the temporal patterns after preprocessing. This finding strengthens confidence that the observed labelling bias is not an artifact of preprocessing but rather reflects genuine disparities in outcome assignment.

## 4.4 Model and Prediction Evaluation

All algorithms were trained using their default parameters; no hyper-parameter tuning was performed. The dataset was split into 85% training and 15% for testing.

### 4.4.1 Area Under the ROC Curve (AUC-ROC)

To assess model performance across subgroups, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was computed separately for each subgroups. This was done by filtering the test dataset according to each subgroup and calculating AUC-ROC using the predicted probabilities and true outcome labels within each group.

Across all subgroups and models, tree-based classifiers, Random Forest and XGBoost consistently outperformed other classifiers. XGBoost, in particular, had the most balanced performance, frequently achieving AUC scores above 0.93 in subgroups such as Male (0.945), Black (0.94), and Private insurance patients (0.956). In general performance of linear models was variable across subgroups. The Highest AUC scores for linear models were observed among patients with private insurance, where Logistic Regression achieved an AUC of (0.94), Linear Discriminant Analysis (0.93), and Support Vector Machines (0.92). In contrast, lower and inconsistent AUC scores were observed for Gaussian Naive Bayes for Asian subgroup (0.75) and for k-Nearest Neighbors (kNN) in the Black subgroup (0.78) (see Table 4.4).

The Private insurance group, despite being smaller than the dominant Medicare group and having a relatively low observed mortality rate (7.3%), achieved the highest AUC-ROC scores across nearly in all models (see Table 4.4. This suggests that patients with private insurance may exhibit more consistent or distinguishable clinical patterns, which are effectively captured by the predictive models. In contrast, larger subgroups, such as White patients (8.9% mortality) and those with Medicare (11.9% mortality), demonstrated moderately high, but not peak, AUC values. This may be attributed to greater clinical and demographic heterogeneity within these groups, which could complicate accurate classification.

In terms of racial subgroups, model performance varied even among groups with similar mortality rates. For example, both Asian and Hispanic patients had mortality rates near 10%, but XGBoost achieved higher AUC in the Hispanic group (0.928) than the Asian group (0.884). This indicates that performance differences are not solely due to class imbalance, but may stem from the distribution and separability of features within each group. Tree-based models may be better able to capture distinct patterns in the Hispanic group due to clearer feature boundaries. It is worth noting that, the high AUC observed in smaller groups, such as Private insurance, highlights that sample size alone does not determine model performance. Rather, the presence of distinctive and well-represented features within a subgroup may contribute more to predictive accuracy than raw number of samples.

Table 4.4: Model performance across demographic subgroups. Values represent AUC scores for each model. The *Death Rate* column reports the observed mortality in each subgroup. Best-performing models per group are shown in bold.

| Feature | Group | Total | Class 0 | Class 1 | Death Rate | LR | SVM | LDA | GNB | RF | XGBoost | kNN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gender | Female | 2854 | 2536 | 318 | 0.111 | 0.917 | 0.897 | 0.909 | 0.838 | 0.916 | **0.924** | 0.795 |
| Gender | Male | 3810 | 3459 | 351 | 0.092 | 0.934 | 0.916 | 0.924 | 0.855 | 0.939 | **0.945** | 0.811 |
| Insurance | Medicaid | 908 | 848 | 60 | 0.066 | 0.897 | 0.891 | 0.892 | 0.804 | 0.906 | **0.920** | 0.817 |
| Insurance | Medicare | 3669 | 3231 | 438 | 0.119 | 0.922 | 0.902 | 0.914 | 0.842 | 0.921 | **0.925** | 0.790 |
| Insurance | Other | 273 | 235 | 38 | 0.139 | 0.922 | 0.881 | 0.906 | 0.821 | 0.932 | **0.949** | 0.859 |
| Insurance | Private | 1814 | 1681 | 133 | 0.073 | 0.948 | 0.927 | 0.932 | 0.881 | 0.949 | **0.956** | 0.820 |
| Race | Asian | 220 | 199 | 21 | 0.095 | 0.884 | 0.839 | 0.861 | 0.760 | 0.854 | **0.884** | 0.804 |
| Race | Black | 576 | 515 | 61 | 0.106 | 0.908 | 0.880 | 0.906 | 0.808 | 0.910 | **0.943** | 0.783 |
| Race | Hispanic | 213 | 192 | 21 | 0.099 | 0.890 | 0.851 | 0.865 | 0.800 | 0.883 | **0.928** | 0.816 |
| Race | Other | 1165 | 1001 | 164 | 0.141 | **0.952** | 0.945 | 0.944 | 0.855 | 0.941 | 0.955 | 0.835 |
| Race | White | 4490 | 4088 | 402 | 0.090 | 0.923 | 0.904 | 0.914 | 0.853 | **0.931** | 0.930 | 0.793 |

### 4.4.2 Equal opportunity

To assess fairness in the identification of positive cases across subgroups, True Positive Rates (TPRs) were compared across each subgroup for all models. Equal opportunity fairness was considered satisfied when TPR differences between subgroups remained within a 10% threshold. Figure 4.7 presents TPR comparisons across racial subgroups. Green cells indicate that TPR differences are within the 10% tolerance (fair), whereas red cells highlight potential bias, where differences exceed 10%.

Across racial groups, Asian and Hispanic patients consistently exhibited lower TPRs compared to White and Black patients. The Black subgroup generally achieved the highest or near-highest TPRs, while Asian patients had the lowest ranging from 0.14 in KNN to

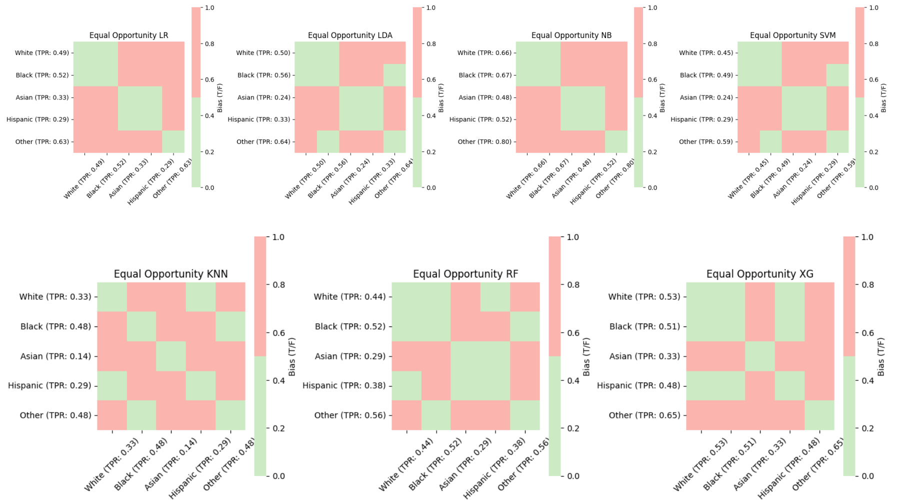

Figure 4.7: True Positive Rate (TPR) differences across racial subgroups for various models. Green cells indicate fairness within the 10% tolerance (fair), and red cells indicate potential bias where difference exceed 10%).

0.48 in GNB. Tree-based models (RF, XGBoost), maintained TPR disparities within the acceptable threshold. XGBoost, in particular, demonstrated the most consistent fairness across White (0.53), Black (0.51), and Hispanic (0.48) subgroups. In contrast, linear models (LR, LDA, SVM) and KNN showed larger disparities, often exceeding 10%.

As shown in Table 4.5, patients with Private insurance consistently received the highest TPRs, whereas Medicaid patients had the lowest or second-lowest TPRs. For example, in GNB the TPRs for Private insurance was 0.77 compared to 0.55 for Medicaid; in SVM, 0.56 vs 0.42; and in KNN: 0.49 vs. 0.32 for Medicare. These results indicate that both race and insurance status influence model performance. While linear and KNN models tend to amplify subgroup disparities, tree-based models had reduced bias and more equitable identification of positive cases.

Table 4.5: Comparison of True Positive Rates (TPRs) Across Insurance Groups and Models

| **Model** | **Medicare** | **Medicaid** | **Private** |
|---|---|---|---|
| Logistic Regression (LR) | 0.51 | 0.43 | 0.55 |
| Linear Discriminant Analysis (LDA) | 0.50 | 0.47 | 0.59 |
| Gaussian Naive Bayes (GNB) | 0.67 | 0.55 | 0.77 |
| Support Vector Machine (SVM) | 0.44 | 0.42 | 0.56 |
| K-Nearest Neighbors (KNN) | 0.32 | 0.40 | 0.49 |
| Random Forest (RF) | 0.42 | 0.45 | 0.56 |
| XGBoost | 0.51 | 0.47 | 0.56 |

### 4.4.3 Demographic parity

To assess demographic parity across subgroups, Positive Prediction Rates (PPRs) were compared for each model across all subgroup. Demographic parity was considered to be satisfied when the differences in PPRs between subgroups remained within a 10% threshold.

Across racial subgroups, no pair exhibited PPRs differences exceeding 10% threshold, indicating that demographic parity was generally maintained with respect to race across all models. In contrast, disparities were observed across insurance groups in Logistic Regression classifier. The PPRs were 0.26 for Medicare and 0.14 for both Medicaid and Private patients. The differences between Medicare and other two groups exceeded the threshold, indicating a violation of demographic parity. Medicare patients were substantially more likely to receive positive predictions, which may reflect demographic factors associated with this group (older age or higher clinical acuity) or potential issues with model calibration. For other models, no subgroup pairs exceeded the threshold, suggesting better adherence to demographic parity across insurance types.

### 4.4.4 Permutation Feature Importance

To understand the contribution of individual features to model performance, a Permutation Feature Importance (PFI) analysis was conducted on the Logistic Regression (LR) and XGBoost (XG) models. A total of 800 permutation repeats was performed to ensure stability and robustness in the importance scores. For both LR and XGBoost, sensitive features (e.g., race, insurance status, and gender) were not ranked highly in model prediction.

Permutation of these features caused little to no change in model performance, suggesting that they are not primary drivers of prediction. Although this may appear to contradict earlier fairness findings, such as disparities in true positive rates and demographic parity. An important consideration in interpreting these results is feature collinearity. When features are correlated, such as race and insurance status, the permutation of a single feature may not degrade performance, as similar information can still be retrieved from correlated counterparts [16].

### 4.4.5 Counterfactual Fairness

To evaluate the counterfactual fairness, changes in model predictions were analysed when sensitive attributes, were altered while all other features were held constant. A prediction flip caused solely by a change in a sensitive attribute indicates that the model may be influenced by that attribute.

For the Logistic Regression classifier, the predicted outcome were reversed for 55 out of 1,500 test samples by only changing sensitive attributes. Across these 55 samples, a total of 737 counterfactual instances were generated, each representing change in sensitive attributes that resulted in a prediction flip. A breakdown by original prediction label showed that among the 737 counterfactuals initially predicted as dead, 82.77% (610 flips) involved changes in race, 55.77% (411 flips) in gender, and 42.74% (316 flips) in insurance status. In contrast, only a single counterfactual instance was found for samples initially predicted as survived, and this flip was influenced by a change in insurance.

Counterfactual analysis of the XGBoost model revealed that predicted outcomes were reversed for only 5 out of 1,500 test samples. Although this represents a very small proportion of the overall test set, a total of 100 counterfactual instances were generated across these five samples, indicating that multiple distinct changes in sensitive attribute led to prediction flips. Among these counterfactuals, race was responsible for 100% of the flips, while changes in gender and insurance each contributed to 40% of the instances. All flips occurred in samples that were originally predicted as dead and no counterfactuals were found for survived predictions.

Overall, both models were found to be vulnerable to sensitive attributes. The predictions generated by the Logistic Regression model were more affected by changes in these attributes compared to the XGBoost classifier. Sensitivity was more noticed in dead predictions, suggesting that demographic attributes had a greater influence in high-risk cases.

# 5 Discussion

## 5.1 Introduction

The study investigated the principles of responsible AI in healthcare by evaluating model performance, data characteristics, and subgroup disparities across sensitive attributes such as race, gender, and insurance status. The findings highlight areas where prediction outcomes may be unevenly influenced by data representativeness, label validity, and model behaviour. In this section, the implications of these results are discussed, along with their limitations and potential causes.

## 5.2 Representation and Data Quality

The analysis revealed that, despite the dataset's large size and demographic diversity, significant imbalances exist within the underlying data. The patient population was heavily skewed toward White individuals (67%) and those insured by Medicare (54%), with minority racial groups and other insurance groups were under-represented. Such imbalances are likely to limit the generalizability of the predictive models trained on this data.

Monitoring patterns across subgroups exhibited clinically expected trends, patients with poorer outcomes generally subject to higher monitoring intensity. However, two notable exceptions were observed. First, Black and Hispanic patients with poorer outcomes experienced increased missingness rate, particularly Systolic Blood Pressure measurements, indicating potential disparities in care delivery or data recording practices. Second, privately insured patients showed a distinct separation in monitoring patterns; those with poorer outcomes received higher than baseline monitoring, whereas those with favourable outcomes had the lowest monitoring ratios relative to other insurance groups. This clear contrast in monitoring intensity may partially explain the consistently higher model performance observed within the privately insured subgroup across all classifiers.

It should be noted that these observed differences might arise from unobserved institutional protocols or clinician decision-making processes, which were not directly captured in the dataset. Nonetheless, such discrepancies have direct implications for the trustworthiness and fairness of AI models trained on this data. The presence of subgroup specific differences in monitoring could propagate into model predictions, potentially increasing inequalities in healthcare AI systems.

These findings emphasis the need for the development of more representative and high-quality training datasets. It is not solely the quantity of data that is critical, but the inclusion of descriptive, clinically relevant features that reflect diverse patient experiences. Furthermore, transparency in data collection pipelines is essential to identify and understand deviations arising from clinical, institutional, or socio-economic biases. This transparency forms a foundational element in building responsible AI systems within healthcare, promoting fairness, accountability, and equity.

## 5.3 Label Validity and Systemic Bias

Beyond data quality concerns, the analysis highlight potential labelling bias within the dataset, particularly across insurance and racial subgroups. There was variations in observed outcomes between Medicare and privately insured patients, even though they were similar in non-protected clinical features their ground truth labels was different.

These findings suggest that outcome labels may reflect underlying biases, whether stemming from subjective human judgment, institutional practices, or administrative policies, that differ by insurance status or racial group. When these biased labels used as in model training, AI systems will inherit or even amplifying existing inequities rather than providing objective clinical predictions.

It is important to note that the source of this labelling bias cannot be fully determined from the dataset alone, representing a key limitation. Nevertheless, the presence of label inconsistency highlights the need for transparent and auditable processes in label generation. To foster the development of responsible AI in healthcare, labelling procedures must be clearly documented and subjected to regular evaluations. Such transparency would enable the identification and mitigation of systemic biases originating from either human or administrative sources. Only by ensuring that training labels are ethically and clinically valid can AI models be trusted to support equitable decision-making and reduce rather than reinforce healthcare disparities.

## 5.4 Model Performance and Subgroup Disparities

Model evaluation revealed significant disparities in performance across demographic subgroups, even among classifiers with generally high accuracy. While tree-based models such as XGBoost and Random Forest consistently achieved strong AUC-ROC scores overall, disparities in true positive rates (TPRs) were still evident, particularly for Asian and Hispanic patients as well as Medicaid-insured individuals. This pattern suggests that while these models are better able to capture subgroup-specific feature interactions, performance gaps remain for under-represented populations. In contrast, linear models such as Logistic Regression and Support Vector Machines, which operate using global decision boundaries, demonstrated less consistent performance and larger disparities across minority groups. These findings imply that model architecture and flexibility play an important role in mitigating subgroup biases but are insufficient on their own to fully ensure fairness.

Also, the defined tolerance for demographic parity was met by most of the models except, Logistic Regression showed significant violations. Medicare patients were predicted positive at a substantially higher rate (0.26) compared to Medicaid and Private insurance groups (both 0.14). This suggests that LR, lacks flexibility and may disproportionately rely on dominant group patterns, resulting in unequal true and false positive rates and potentially unfair outcomes.

It should be acknowledged that all models were trained using default hyper-parameters without tuning, which likely limited their ability to generalize effectively across diverse sub-populations. Overall, these results show the importance of comprehensive fairness evaluations beyond traditional metrics like Accuracy, AUC-ROC and etc., which can mask disparities within subgroups. Transparent reporting of performance across sensitive attributes, combined with subgroup-specific evaluation, is essential to identify and address potential harms. Such practices are foundational for the responsible AI systems in high-stakes healthcare environments where equity and trustworthiness are necessary.

## 5.5 Permutation Feature Importance and Counterfactual Fairness

Permutation feature importance (PFI) analysis for both Logistic Regression (LR) and XGBoost models indicated that sensitive attributes had minimal impact on overall model performance. This finding might initially suggest that these models do not heavily rely on sensitive features, implying a degree of fairness at the global level.

However, counterfactual fairness evaluation had a different result. Predictions were observed to change when sensitive attributes were altered while keeping all other features remained constant. This indicates that certain demographic attributes can influence on individual predictions even when their overall importance appears low. This variations highlights how permutation feature importance might be affected by feature correlation [16]. Specifically, when two correlated features exist, permuting one feature may not degrade model performance substantially because the correlated feature still provides similar information, thereby underestimating their true importance. Based on these results, it is worth investigating future correlation.

Furthermore, it must be acknowledged that the PFI analysis was performed on a relatively limited test set (800 samples), which may have reduced the sensitivity and stability of the importance estimates. Larger and more representative samples are likely required to better capture the effects of sensitive features and provide more robust estimates.

## 5.6 Comparison with evaluation done on MIMIC-III

This study is built upon prior evaluations conducted using the MIMIC-III dataset [15] through an extension of the analysis to MIMIC-IV, the latest version. While core characteristics are shared between MIMIC-III and MIMIC-IV [7], shifts in clinical practice, increased dimensionality, changes in data structure related to the critical care focus, are reflected in MIMIC-IV. Correlations among data fields are also observed, all of which may influence model development and fairness assessments.

Consistent with findings from evaluations on MIMIC-III, class imbalance was observed to remain a persistent challenge, with models showing strong overall discrimination as measured by AUROC but notably lower performance for the minority and high-risk subgroups. The relatively good model performance for privately insured patients, despite their smaller representation, noted in prior studies, was also evident in this study.

Findings from [15] alongside, this study results, emphasize the necessity of multi-dimensional evaluations across the entire machine learning pipeline; from data to predictions; to promote responsible AI development. MIMIC serves as a prime example dataset, demonstrating that caution must be taken before deploying AI systems, especially in critical areas like healthcare. Responsible AI deployment in healthcare demands, detailed evaluations to ensure that technological advances translate into equitable improvements in patient outcomes.

## 5.7 Implications for Responsible AI in Healthcare

The findings from this study highlight several important considerations for the responsible development and deployment of AI systems in healthcare:

- Data is not neutral: patient representation, monitoring patterns, and missing-ness rate are influenced by systemic and institutional factors. These variations must be critically identified and addressed to avoid healthcare disparities.

- Labelling bias can propagate societal inequities: observed outcome labels are shaped by human judgment or administrative policies may encode existing biases, which predictive models can inherently learn and amplify. Transparent and auditable label generation processes are therefore essential.

- Performance metrics alone are insufficient: standard evaluation measures such as AUC-ROC may mask disparities in model performance across under-represented groups. Subgroup-specific analyses should be integrated throughout model development, rather than reserved as a final validation step.

- Multiple fairness metrics are necessary: no single metric fully captures the complexities of responsible AI principles. A comprehensive assessment framework is essential.

- Model selection should be informed by data characteristics: understanding the underlying data distributions, quality, and biases should guide algorithm choice. Models that better capture subgroup-specific patterns may reduce disparities and improve equitable performance.

# 6 Conclusion and Future Work

## 6.1 Conclusion

This study evaluated fairness, accountability, and transparency (FAT) across the full machine learning pipeline using the MIMIC-IV database for in-hospital mortality prediction. The results revealed multiple sources of potential disparity, including subgroup differences in data completeness, monitoring intensity, label validity, and model performance. These disparities, particularly along racial and insurance lines, underscore that bias can originate not only in model training but from upstream data collection and labelling processes. Transparent documentation of each pipeline step is therefore essential to building equitable and trustworthy AI systems in healthcare.

## 6.2 Future Work

Building on the findings of this study, the following areas need further research and development to support responsible AI in healthcare:

- **Bias Detection Methods:** the development of dedicated techniques to detect bias in both clinical data and model outputs is essential. Identifying specific biases will allow for more targeted and effective mitigation strategies.
- **Transparent Reporting and Reproducibility Protocols:** Adopting standardized guidelines for transparent reporting and reproducible methodologies will

enhance clarity, comparability, and accountability throughout the AI development pipeline.

- **Generalization to Broader Tasks and Modalities:** Extending this evaluation to other clinical prediction tasks and incorporating diverse data modalities and sources would provide a more comprehensive understanding of the challenges to develop responsible AI system across healthcare AI system.

# Bibliography

[1] Nikos Afxentis. “Predicting Mortality and Algorithmic Fairness of ICU Patients”. MA thesis. 2024.

[2] Stephanie Baker and Wei Xiang. “Explainable AI is responsible AI: How explainability creates trustworthy and socially responsible artificial intelligence”. In: *arXiv preprint arXiv:2312.01555* (2023).

[3] Rachel KE Bellamy et al. “AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias”. In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 4–1.

[4] Feng Chen et al. “Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models”. In: *Journal of the American Medical Informatics Association* 31.5 (2024), pp. 1172–1183.

[5] Ben Chester Cheong. “Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making”. In: *Frontiers in Human Dynamics* 6 (2024), p. 1421273.

[6] Sribala Vidyadhari Chinta et al. “AI-driven healthcare: Fairness in AI healthcare: A survey”. In: *PLOS Digital Health* 4.5 (2025), e0000864.

[7] Dillon Chrimes and Chanhee Kim. “Comparison of MIMIC-III and MIMIC-IV for big data analytics of health informatics”. In: *2023 IEEE International Conference on Big Data (BigData)*. IEEE. 2023, pp. 6128–6130.

[8] Jianhui Gao et al. “What is Fair? Defining Fairness in Machine Learning for Health”. In: *arXiv preprint arXiv:2406.09307* (2024).

[9] Hrayr Harutyunyan et al. “Multitask learning and benchmarking with clinical time series data”. In: *Scientific data* 6.1 (2019), p. 96.

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

[11] Alexandra Kakadiaris. “Evaluating the Fairness of the MIMIC-IV Dataset and a Baseline Algorithm: Application to the ICU Length of Stay Prediction”. In: *arXiv preprint arXiv:2401.00902* (2023).

[12] Matt J Kusner et al. “Counterfactual fairness”. In: *Advances in neural information processing systems* 30 (2017).

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

[14] Anand Murugan. “Implementing Fairness in Real-World Healthcare Machine Learning through Datasheet for Database”. MA thesis. University of Waterloo, 2024.

[15] Eliane Röösli, Selen Bozkurt, and Tina Hernandez-Boussard. “Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model”. In: *Scientific Data* 9.1 (2022), p. 24.

[16] scikit-learn developers. *Permutation Importance with Multicollinear or Correlated Features*. https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html. Accessed: 2025-08-19. 2025.

[17] Pavel Senin. "Dynamic time warping algorithm review". In: *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA* 855.1-23 (2008), p. 40.

[18] Junming Shi et al. "Implicit bias in ICU electronic health record data: measurement frequencies and missing data rates of clinical variables". In: *BMC Medical Informatics and Decision Making* 25 (2025), p. 241.

[19] Haytham Siala and Yichuan Wang. "SHIFTing artificial intelligence to be responsible in healthcare: A systematic review". In: *Social Science & Medicine* 296 (2022), p. 114782.

[20] I. Silva et al. "Predicting In-Hospital Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012". In: *Computing in Cardiology (2010)* 39 (2012), pp. 245–248.

[21] Aditya Singhal et al. "Toward fairness, accountability, transparency, and ethics in AI for social media and health care: scoping review". In: *JMIR Medical Informatics* 12.1 (2024), e50048.

[22] Kacper Sokol, Raul Santos-Rodriguez, and Peter Flach. "FAT Forensics: A Python toolbox for algorithmic fairness, accountability and transparency". In: *Software Impacts* 14 (2022), p. 100406.

[23] Herdiantri Sufriyana et al. "Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in pregnancy care: systematic review and meta-analysis". In: *JMIR medical informatics* 8.11 (2020), e16503.

[24] Shahadat Uddin et al. "Comparing different supervised machine learning algorithms for disease prediction". In: *BMC medical informatics and decision making* 19.1 (2019), pp. 1–16.

[25] Hilde Weerts et al. “Fairlearn: Assessing and improving fairness of AI systems”. In: *Journal of Machine Learning Research* 24.257 (2023), pp. 1–8.

[26] James Wexler et al. “The what-if tool: Interactive probing of machine learning models”. In: *IEEE transactions on visualization and computer graphics* 26.1 (2019), pp. 56–65.

[27] Nengfeng Zhou et al. “Bias, fairness, and accountability with AI and ML Algorithms”. In: *arXiv preprint arXiv:2105.06558* (2021).

[28] X. Zhou. “A study of machine learning applications in healthcare”. In: *Applied and Computational Engineering* 102 (2024), pp. 128–133.