# Advancing Trust in Real-World Healthcare AI: A Framework for Enhanced Transparency and Fairness

**Anand Murugan**
University of Waterloo
Ontario, Canada N2L5S7
a7muruga@uwaterloo.ca

**Sirisha Rambhatla**
University of Waterloo
Ontario, Canada N2L5S7
srambhat@uwaterloo.ca

**Alexander Wong**
University of Waterloo
Ontario, Canada N2L5S7
alexander.wong@uwaterloo.ca

## Abstract

Healthcare Machine Learning (HML) models are revolutionizing the healthcare industry, promising improved patient outcomes and enhanced public health. However, the black-box nature of HML models raises concerns about bias and trust. To ensure these models are reliable and trustworthy, examining the data they use and evaluating their transparency, fairness, and accountability alongside their predictive abilities is critical. This study presents a six-year systematic review of risk prediction HML models using the MIMIC clinical research database (CRD). The results were striking: for the popular MIMIC IV – Intensive Care Unit (ICU) mortality task, a naive baseline outperformed the state-of-the-art (SOTA) model, showing not only better predictive performance but also improved fairness among different patient groups, though some unfairness remains. These results highlight the urgent need for a robust framework to enhance the trustworthiness of HML models. We propose a three-step framework: (i) a 'Datasheet for CRD' to promote data transparency and identify inherent data bias, inspired by recent recommendations for datasheets in datasets; (ii) a 'Model Card' detailing model characteristics, providing interpretable results and evaluating performance across demographic groups, inspired by model cards for model reporting; and (iii) a 'Fairness Report' to thoroughly assess and address any biases related to sensitive attributes, promoting fairness across all groups. This framework will assist practitioners in identifying data anomalies using the datasheet, model structure and its usage using the model card and fairness results and its impact from the fairness report, thereby fostering the development of trustworthy HML models.

## 1 Introduction

Machine learning (ML) has become a powerful tool in the healthcare industry, enabling the development of accurate prediction models that can significantly improve public health outcomes [85, 103]. However, in the high-stakes context of healthcare, ensuring trustworthy and fair predictive outcomes is crucial [84]. Despite significant progress in HML, unfairness still exists due to biases in data and algorithms [6, 16, 71]. These biases can perpetuate through the model's complex, black-box nature, resulting in unfair results and decreased stakeholder trust. It is essential to address these concerns

Preprint. Under review.

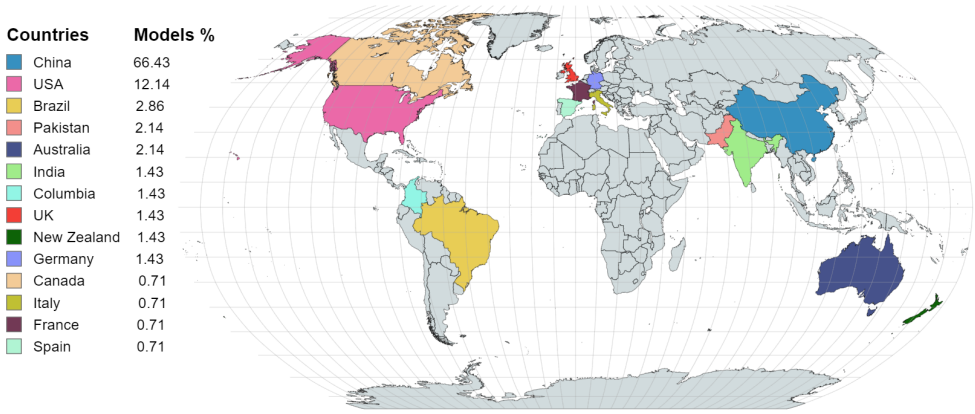



Figure 1: Distribution of HML models using MIMIC data worldwide from 2018 to 2024. The country list represents the widespread usage of MIMIC data for healthcare predictions globally, and the % value indicates the HML model researched by the respective country.

to ensure that the benefits of HML models are accessible to everyone, regardless of race, gender, or socio-economic status.

Fair ML is an active area of research focusing on how models can perpetuate inequalities and the ways to address them [11, 12, 35, 52, 72, 86, 87, 132]. Popular ML fairness strategies aim to assess and reduce algorithmic discrimination by imposing fairness constraints during model training [20, 49, 60, 73]. However, one of its inherent limitations is the data itself, which encodes real-world biases and inconsistencies, making it difficult to achieve fairness in healthcare settings [7]. This insight can be traced back to 2018, when Chen et al. [11] proposed that the predictive fairness evaluation must consider model training in the context of the input data.

To investigate the progress towards fairness in HML, we analyze the state of fair HML research through the lens of Medical Information Mart for Intensive Care (MIMIC) [45, 46] database, one of the most popular CRD [87]; see Figure 1 for its popularity across the world. Our survey identified ICU mortality prediction as the most widely researched task in MIMIC III/IV. To analyze the ICU mortality task, we compared the performance and fairness properties (across ethnicity) of the SOTA: DuETT [1] [48] with simple baseline models (such as XGBoost) on MIMIC IV. Surprisingly, we find that the SOTA model is outperformed by XGBoost not only in performance but also in terms of its fairness properties, with disparate impact (DI) fairness metric revealing substantial inequalities across ethnic subgroups [2]. This demonstrates that despite more than half a decade since [11], it has been challenging to incorporate fairness in HML modeling, highlighting a need to reconsider it in the context of (i) the task and (ii) the data driving the model, and ways to make it a central part of any HML analysis.

The aforementioned example is not to undermine the work conducted by SOTA; it is to highlight the complexities involved in accomplishing HML fairness. For instance, MIMIC IV is a large database, which makes it especially challenging for practitioners to focus on modeling and fairness concurrently since they may only be interested in a small section of it. With recent calls in the community to develop 'Datasheets for Dataset', spearheaded by Gebru et al. [27], we develop a blueprint to enable the next generation of HML that is trustworthy and fair. Specifically, we introduce the 'Datasheet for CRD (MIMIC IV v2.0)', a comprehensive data-centric resource for practitioners tailored for complex CRDs. Here, we call for the data and tasks to take center stage in any HML modeling to accomplish fairness. To this end, our datasheet is designed to support practitioners in detecting data anomalies across the database and analyzing task-specific associations between features and the target, enabling equitable decisions in HML modeling.

[1] DuETT is considered SOTA in this study for its reported superior performance metrics with MIMIC IV for ICU Mortality task, comprehensive evaluation in comparison with a varied baseline for different tasks and its innovative technique. However, no fairness metrics were reported by the paper.

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

For trustworthy HML modeling, model techniques and considerations play a crucial role in addition to the data. It is essential to know how a model was trained, the data it was exposed to, and its expected behavior in various scenarios [74]. This information is vital for identifying potential biases and understanding the model's limitations. Our 'Model card' provides detailed documentation of ML models, including their intended use, performance metrics, and evaluation results across different demographic groups. It facilitates informed decision-making by providing a comprehensive overview of a model's strengths and weaknesses, promoting transparency. It enables practitioners to assess whether a model is suitable for deployment in sensitive applications like healthcare, where the implications of biased or inaccurate predictions can be life-altering. We followed the 'Model Cards for Model Reporting' [74] framework to report on the analyzed HML models.

So, we advocate for the following three-step framework to build a trustworthy HML model. (i) a 'Datasheet for CRD' to identify inherent data bias and build a fairness-centric HML; (ii) a 'Model Card' detailing model characteristics and evaluating performance across demographic groups; and (iii) a 'Fairness Report' to thoroughly assess and address any biases related to sensitive attributes. This approach is especially significant in the light of recent works in fairness theory, which establish that it is not feasible to satisfy all fairness criteria at once [91]. There have also been calls to include end-users in determining key fairness goals [83]. Yet, the lack of fairness in the analysis (by SOTA) is alarming. Therefore, using our framework together promotes data transparency, model reliability, and fair outcomes irrespective of their sensitive attributes. It also enhances the trustworthiness of HML models by ensuring that all relevant factors have been considered and communicated effectively. Our contributions to this study are as follows,

1. **Comprehensive Review of HML Progress:** Systematic survey spanning over half a decade of HML models trained on MIMIC III/IV.
2. **Investigation of the challenges faced by fair HML** via an analysis of ICU Mortality prediction in MIMIC. Here, we rigorously compare the performance of the SOTA on this task with naive baselines to demonstrate that the current HML is struggling to incorporate fairness evaluation [3].
3. **Comprehensive 'Datasheet for CRD (MIMIC IV v2.0)':** To spark advances in fair HML and to aid practitioners, we present a comprehensive datasheet for MIMIC IV v2.0 CRD, highlighting data inconsistencies across the database and includes task-specific feature-target analysis to streamline fairness evaluations for equitable decision-making in HML models. The datasheet is made available at `https://anonymous.4open.science/r/DatasheetCRD-1F00/README.md`
4. **Model card for the HML models:** We provide comprehensive model cards for HML risk prediction models, which report on model details, performance metrics, train/val/test information, and its intended use.
5. **HML model fairness report:** We provide a detailed report on the fairness of the analyzed HML models concerning the highly correlated sensitive attribute to evaluate the model's performance across demographics.

## 2 Systematic Survey on MIMIC trained HML models

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

### 2.1 Survey Method

PubMed and Google Scholar databases were extensively searched because of their medical focus, broad coverage, and potential to uncover emerging trends beyond specialized databases. We adopted a broad search term (see Appendix A.1) approach following [28] to capture extensive research on MIMIC-trained HML models. Only 220 studies were included after the initial abstract and title screening, as shown in Table 3 in Appendix A.1.

Following PRISMA guideline [77], studies were screened to exclusively consider those utilizing MIMIC III/IV—the recent publicly available database versions—as the primary data source. A total

[3]We surprisingly find that naive baselines outperform the SOTA on this task.

of 140 papers from both databases, including recent research up until February 2024, were selected after analyzing the citation count, methodologies used, and clarity of the work. The selected scientific works span conferences like NeurIPS, IJCAI, ICML, ICLR, AAAI, ACM FAccT, and journals like Nature, JAMA, JAMIA, BMC, PLoS One, etc.

We meticulously adhered to the TRIPOD guidelines, as outlined by Moons et al. in [75]. This adherence was pivotal in guiding our data abstraction, evaluation, and synthesis methodology (see Appendix A.1), specifically identifying and addressing data bias. The TRIPOD guidelines provided a comprehensive framework that ensured our approach was systematic, methodical, and transparent, facilitating the development of robust, fair, and reliable HML risk prediction models.

### 2.2 Healthcare Risk Prediction Task

We grouped the selected research works based on the predicted task and observed that ICU mortality is the most extensively studied HML prediction task worldwide ( 70%). It is closely followed by ICU readmission and ICU length of stay (LOS). The comprehensive list of the HML prediction tasks is included in Appendix A.2. Our survey found that about 67% of HML models trained on the MIMIC CRD came from China [4], with significant contributions from the USA, Brazil, Pakistan, Australia, and India, as shown in Figure 1. These findings highlight that ICU mortality prediction is a key area within HML models, making it the main focus of the study.

## 3 Problem Formulation

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

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

$$\widehat{y} = f(\mathcal{D}) \tag{1}$$

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

Given these predictions, we then use the fairness metrics defined in section 3.2 to evaluate the models.

### 3.1 Healthcare ML Fairness Measurements

Fairness in HML ensures equitable model performance and decision-making across diverse patient groups, avoiding bias based on sensitive attributes like ethnicity, gender, or age [5].

As mentioned above, for each patient $i$, $\mathbf{s}_i \in \mathbb{R}^S$ is a sensitive attribute vector. We performed a comprehensive statistical analysis to identify the most closely associated sensitive attribute with the target variable. Chi-square analysis revealed a significant association between ethnicity and ICU mortality, identifying it $s_i^{eth}$ as the sensitive attribute. This informed our selection of fairness group metrics [25] to evaluate disparities across ethnic groups. In the subsequent sections, we omit patient index $i$ for notational simplicity.

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

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

The ethnic attribute $s^{eth}$ is categorized into five groups: Asian, Black, Hispanic/Latino, White, and Other. For fairness assessments, we compare pairs of these groups, and the function priv$(\cdot)$ assigns status for the sub-groups, $a$ for `'privileged'` and $b$ for `'protected'` based on the ethnic groups being compared. The ground truth is denoted by $y$, and the model $f(\cdot)$ prediction is $\widehat{y}$.

### 3.2 Fairness Metrics

Fairness metrics in machine learning serve as quantitative benchmarks to evaluate and ensure that algorithms perform equitably across all user groups, particularly when decisions impact individual's lives. Various fairness metrics exist, each with different implications for model assessment, and based on our analysis results, group fairness metrics are used in this study. It provides unbiased risk assessments across all demographic groups, preventing systematic over- or under-prediction for any group.

**Demographic Parity**, which advocates for equal opportunity allocation, stipulates that the probability of a favorable outcome should be independent of the sensitive attribute [21], i.e. the probability of a positive prediction should be equal across different ethnic groups.

$$\mathbb{P}(\widehat{y} = 1|\text{priv}(s^{eth}) = a)) = \mathbb{P}(\widehat{y} = 1|\text{priv}(s^{eth}) = b)) \tag{2}$$

**Equalized Odds**, a metric that promotes equal performance, requires equal decision rates for privileged and unprivileged groups and is defined as [10], i.e. the model's decision rates for predicting an outcome should be the same across different demographic groups, given the actual outcome.

$$\mathbb{P}(\widehat{y} = y|\text{priv}(s^{eth}) = a, y) = \mathbb{P}(\widehat{y} = y|\text{priv}(s^{eth}) = b, y), \quad \forall y \in \{0, 1\} \tag{3}$$

**Equal Opportunity** advocates equal true positive rates across different ethnic groups [122], aiming for fairness in model sensitivity.

$$\mathbb{P}(\widehat{y} = 1|\text{priv}(s^{eth}) = a, y = 1) = \mathbb{P}(\widehat{y} = 1|\text{priv}(s^{eth}) = b, y = 1) \tag{4}$$

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

$$\frac{\sum(\text{priv}(s^{eth}) = a, \widehat{y} = 1)/\sum(\text{priv}(s^{eth}) = a)}{\sum(\text{priv}(s^{eth}) = b, \widehat{y} = 1)/\sum(\text{priv}(s^{eth}) = b)} \tag{5}$$

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

## 4 Datasheet for Clinical Research Database

Understanding the inherent data and its inconsistencies during modeling is essential to achieve data-centric fairness. Practitioners may find it challenging to navigate the documentation while concentrating on modeling and fairness due to the breadth of the database [46], as they might only be interested in datasets relevant to their task. So, we present the 'Datasheet for CRD'[6] for MIMIC IV v2.0 as shown in Appendix A.4 Figure 5.

The datasheet for MIMIC IV v2.0 includes:

1. An expanded overview of the MIMIC IV database, adapted from the [27] template to cover all CRD intricacies.

[6]Our comprehensive 'Datasheet for MIMIC IV v2.0' can be accessed at `https://anonymous.4open.science/r/DatasheetCRD-1F00/`

2. A detailed composition section highlighting the CRDs unique structure and potential biases, incorporating MIMIC IV modules like 'Hosp', 'ICU', MIMIC IV-ED, MIMIC IV Notes, and MIMIC-CXR.
3. Custom queries tailored to the CRD, providing a deep understanding of data collection, composition, and task-specific usage, prefixed with '+' in the datasheet. For instance, *'Can/How the dataset be/are constructed from the MIMIC database?'* replaces general questions like *'What is the composition of the dataset?'*.
4. An analysis section offering insights into selecting sensitive attributes for HML prediction tasks (ICU Mortality, Length of Stay, Readmission) and appropriate fairness metrics.

More than just an inventory, this datasheet provides essential transparency, detailing the database structure, data collection methodologies, management practices, data inconsistencies, and key fairness considerations.

# 5 Utilizing Datasheet for Database in modeling - ICU Mortality prediction task

To explore the link between clinical data and fairness in healthcare predictions, we benchmarked static models like Logistic Regression (LR) and XGBoost, as well as time-series models like LSTM and STraTS, against the SOTA DuETT model for ICU mortality prediction.

## 5.1 MIMIC Database

The datasheet provides detailed information on the motivation, collection, composition, and usage of MIMIC CRD. Table 1 summarizes the general demographic characteristics for ICU Mortality of MIMIC IV. The sensitive attribute statistics and the complete feature lists (see Table 4) are provided in Appendix A.3.

Table 1: Breakdown of Sensitive Attributes in MIMIC IV v2.0 ICU Mortality dataset: Distribution of patient demographics across gender, age, ethnicity, language, and insurance type, highlighting the diversity and potential biases inherent in the database.

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

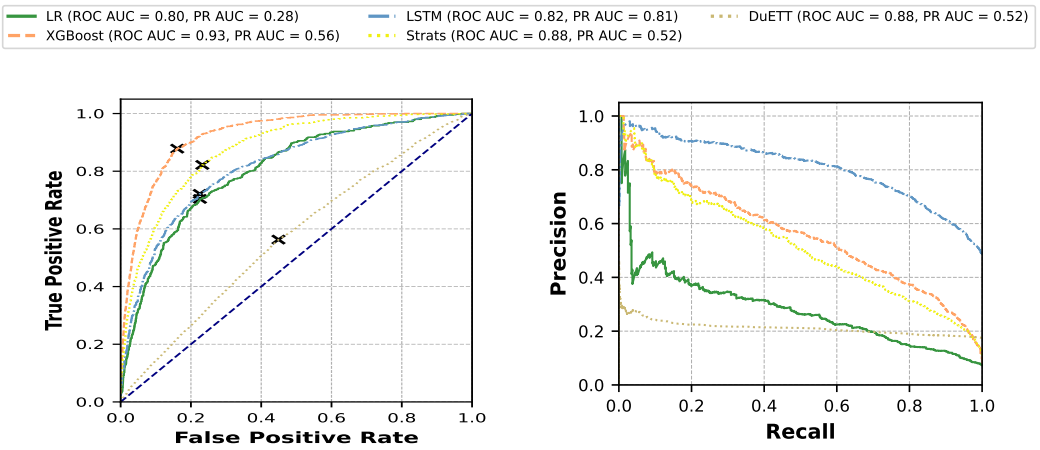
## 5.2 Experimental setup

To examine how fair the predictions of the SOTA model are, we conducted a series of experiments against several baselines on the MIMIC IV dataset. We adopted an 80:20 split for the training and test sets. The models were trained for up to 1000 epochs until the validation loss stopped improving for 10 continuous epochs, applying a 10-fold cross-validation for 3 Monte Carlo runs. The target of the model is to predict the probability of mortality ($\widehat{Y}$) following ICU admission of the patient. The experiments were conducted on NVIDIA GeForce RTX 2080 Ti GPU and the entire implementation code are available on this GitHub page.

## 5.3 Experimental Results

We highlight our results in Table 2, which reports the mean and standard deviation of the performance metric over 3 Monte Carlo runs. Our evaluation comprised static baselines like LR, XG Boost, and TS baselines like LSTM, STrats [102] with the SOTA DuETT [48] [7].

[7] Physionet 2012 Challenge code of the DuETT is only made online. We used the DuETT architecture provided in [48] and replicated it for the ICU Mortality task for evaluation




(a) ROC Curve

(b) PR Curve

Figure 2: Prediction performance analysis across models. Panels (a) and (b) show the ROC-AUC and PR-AUC of the models, respectively, and the operating points in (a).

### 5.3.1 Model Performance Evaluation

We used the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) and Precision-Recall Area Under the Curve (PR-AUC) metrics to compare model performance comprehensively. These comparisons, derived from 3 Monte Carlo simulations to ensure statistical robustness, are visually represented in Figure 2. Surprisingly, the XG Boost model demonstrated a significant performance uplift, outperforming STrats by 6.44% and SOTA by 59.6% in ROC-AUC, as detailed in Table 2.

Table 2: Comparative Analysis of ICU Mortality Prediction: Predictive performance of models trained on MIMIC IV v2.0, contrasting static and time-series models as reflected by the ROC-AUC and PR-AUC curves for 3 Monte Carlo runs.

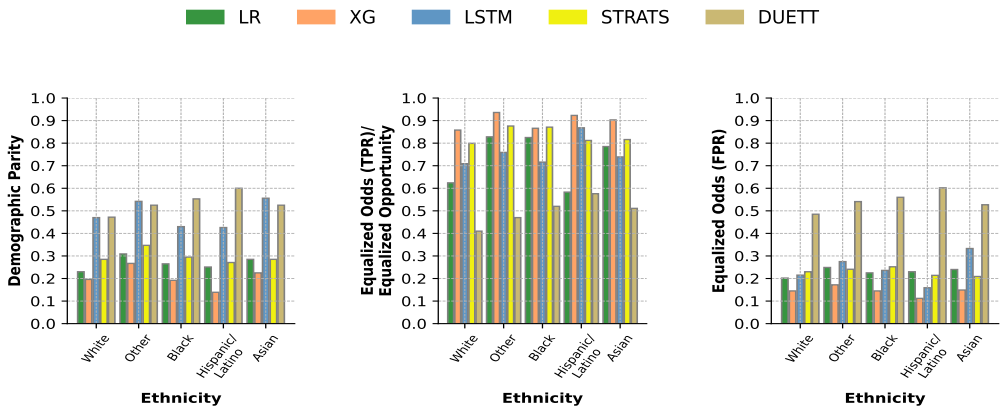
| **Type** | **Model** | **ROC-AUC** | **PR-AUC** |
|---|---|---|---|
| Static | LR | 0.805±0.012 | 0.276±0.005 |
| | **XG Boost** | **0.926±0.006** | 0.551±0.007 |
| Time series | **LSTM** | 0.886±0.003 | **0.807±0.011** |
| | STraTS | 0.870±0.002 | 0.520±0.006 |
| | **DuETT (SOTA)** | **0.580±0.028** | **0.220±0.005** |

**Model Card:** A model card detailing the model specifications, training/testing process, data usage, its expected behavior in various scenarios, and its intended use in addition to the above performance results will add another layer of transparency most needed in building trustworthy HML models. This information can help ML practitioners identify potential algorithmic bias quickly and understand model limitations. Detailed model cards for all the analyzed HML models are available here.

### 5.3.2 Model Fairness Assessment

To evaluate the ethical aspects, we assessed each model using established fairness metrics to identify the biases and disparities across ethnic groups. The results, summarized in the Appendix A.5 Table 5, revealed notable variations in the HML model's fairness. Figures 3(b) and 3(c) show how XG Boost consistently surpassed the SOTA model in both Equalized Opportunity (EOp) and Equalized Odds (EO) metrics, highlighting its superior fairness profile. Additionally, Figure 4 illustrates Demographic Parity (DP) disparities across ethnic categories, notably where the SOTA model still manifests considerable bias despite achieving higher DP scores than XG Boost.

**Fairness Report:** In addition to the 'Datasheet for database' and the 'Model card', the 'Fairness report' is also imperative for building a trustworthy HML model. It will list all information related to




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

Figure 3: Analysis of fairness metrics for each model across ethnic subgroups. Panels (a), (b), and (c) show the comparative analysis for demographic parity (DP), equalized odds for TPR (also known as EOp), and FPR, respectively. Inconsistency drawn from these metrics reveals the extent of fairness exhibited by each model, highlighting the interplay between algorithmic performance and demographic impact.

(a) LR (b) XG Boost (c) LSTM

(d) STraTS (e) DuETT (SOTA)

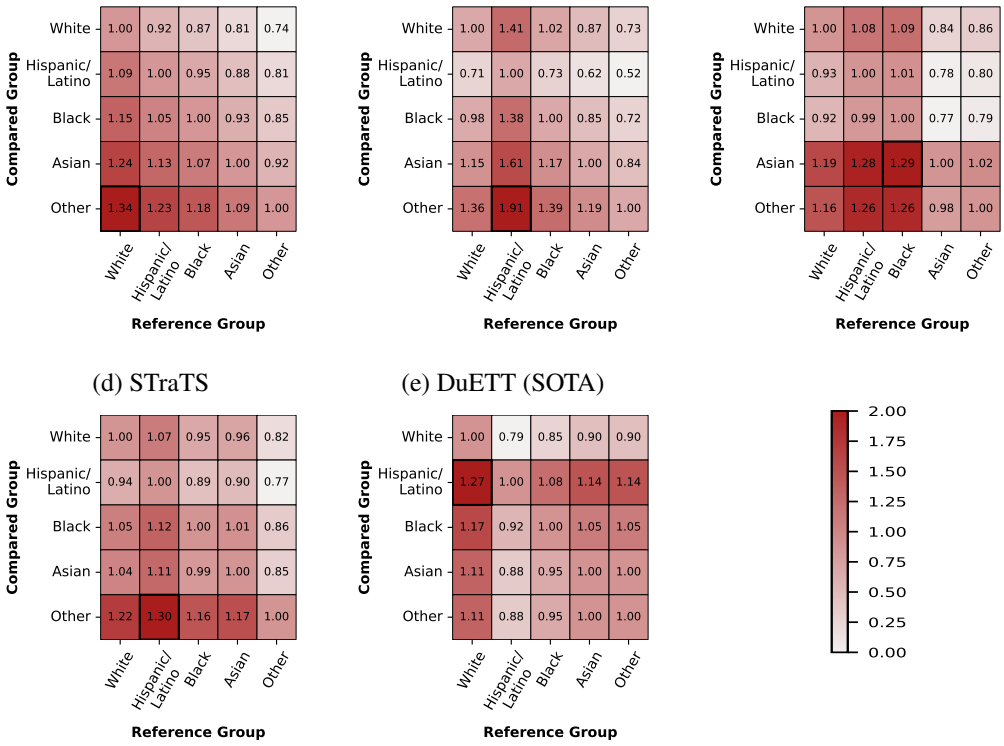



Figure 4: Heatmaps illustrating disparate model impact across ethnic subgroup groups. Red < 1.0 indicates the disparity of the model against the compared group, Red > 1.0 reflects the model's disparity against the reference group and the value of 1.0 shows parity against the comparison. Highlighted cells represent the most disparate subgroup comparison.

HML fairness, evaluation metrics, fairness performance, bias mitigation steps, and recommendations. A detailed fairness report is provided in Supplementary Section C.

## 6 Discussions, Limitations, and Future Work

Trustworthy healthcare machine learning models are essential for ensuring accurate, transparent, and unbiased predictions, improving patient outcomes, and fostering confidence among healthcare providers and patients. Our findings highlight the complex challenges of embedding fairness into HML models, emphasizing the need to address data quality, usage practices, and inherent biases in CRD data. The observed discrepancies in fairness metrics [8] across models, underscoring our framework's necessity, including a 'Datasheet for Database', 'Model Card', and 'Fairness Report'. This approach enhances model accuracy and ensures equitable treatment from HML systems.

**Synthetic Data Generation Enhanced by Datasheets:**

As we advance toward synthetic data generation, the Datasheet for CRD is crucial. It is imperative to recognize that synthetic data, derived solely from CRDs like MIMIC, do not fully capture the global patient population as addressed in Section 1. Combining synthetic datasets with real patient data from diverse locations ensures global representation. By using the datasheet to identify and address data gaps in addition to it, we facilitate the creation of enriched synthetic datasets that better mirror the global patient population.

**Limitations and Future Work:** This study focuses mainly on HML model fairness concerning sensitive attributes, especially ethnicity. Yes, it is necessary to evaluate the fairness of the model based on the sensitive attributes to analyze the bias, but it does not fully explore other correlated variables, such as socio-economic status, access to healthcare, pre-existing health conditions, and treatment quality, which also impact the model's prediction outcome. Our future work will include obtaining and analyzing these details from other CRDs such as eICU [80] and HiRID [40] [9] and develop a datasheet for those to offer a more comprehensive understanding of the factors influencing model fairness. All information provided in this study and the 'Datasheet for database' adhere to the strict policies of handling medical data, safeguarding patient privacy, and maintaining research integrity. We strongly discourage any misuse of the provided datasheet or survey/analysis results, including but not limited to developing algorithms/models with negative societal impacts.

## References


[1] Fahad Shabbir Ahmad, Liaqat Ali, Hasan Ali Khattak, Tahir Hameed, Iram Wajahat, Seifedine Kadry, and Syed Ahmad Chan Bukhari. A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (ehrs). *Journal of Ambient Intelligence and Humanized Computing*, 12:3283–3293, 2021.

[2] Aliya Ali, Saleha Yurf Asghar, Ali Danish Khan Yousafzai, Ali Haider Bangash, Rabia Mohsin, Arshiya Fatima, Saiqa Zehra, Ayesha Khalid Khan, Ali Haider Shah, Syed Mohammad Mehmood Abbas, et al. Prediction of in-hospital mortality among heart failure patients: An automated machine learning analysis of mimic-iii database. *American Heart Journal*, 254:261, 2022.

[3] Itai Bendavid, Liran Statlender, Leonid Shvartser, Shmuel Teppler, Roy Azullay, Rotem Sapir, and Pierre Singer. A novel machine learning model to predict respiratory failure and invasive mechanical ventilation in critically ill patients suffering from covid-19. *Scientific Reports*, 12(1):10573, 2022.

[4] Dan Bo, Xinchun Wang, and Yu Wang. Survival benefits of oral anticoagulation therapy in acute kidney injury patients with atrial fibrillation: a retrospective study from the mimic-iv database. *BMJ open*, 13(1): e069333, 2023.

[5] Markus Böck, Julien Malle, Daniel Pasterk, Hrvoje Kukina, Ramin Hasani, and Clemens Heitzinger. Superhuman performance on sepsis mimic-iii data by distributional reinforcement learning. *PLoS One*, 17(11):e0275358, 2022.

[6] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.


[8] Röösli et al. [87] studied the pervasive challenges around bias and fairness in HML models using MIMIC but only with fairness diagnostic tool. Our study provides a framework for trustworthy HML model by analyzing the SOTA and baseline performance on MIMIC IV

[9] MIMIC doesn't fully cover the other correlated variables like socio-economic status, access to healthcare, pre-existing health conditions

[7] Maarten Buyl and Tijl De Bie. Inherent limitations of ai fairness. *arXiv:2212.06495*, 2022.

[8] William Caicedo-Torres and Jairo Gutierrez. Iseeu: Visually interpretable deep learning for mortality prediction inside the icu. *Journal of biomedical informatics*, 98:103269, 2019.

[9] William Caicedo-Torres and Jairo Gutierrez. Iseeu2: Visually interpretable mortality prediction inside the icu using deep learning and free-text medical notes. *Expert Systems with Applications*, 202:117190, 2022.

[10] Simon Caton and Christian Haas. Fairness in machine learning: A survey, 2020.

[11] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

[12] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.

[13] Qifan Chen, Yang Lu, Charmaine Tam, and Simon Poon. Outcome-oriented predictive process monitoring to predict unplanned icu readmission in mimic-iv database. 2022.

[14] Bihuan Cheng, Diwen Li, Yuqiang Gong, Binyu Ying, Benji Wang, et al. Serum anion gap predicts all-cause mortality in critically ill patients with acute kidney injury: analysis of the mimic-iii database. *Disease markers*, 2020, 2020.

[15] Wei-Ting Chiu, Lung Chan, Jakir Hossain Bhuiyan Masud, Chien-Tai Hong, Yu-San Chien, Chih-Hsin Hsu, Cheng-Hsueh Wu, Chen-Hsu Wang, Shennie Tan, and Chen-Chih Chung. Identifying risk factors for prolonged length of stay in hospital and developing prediction models for patients with cardiac arrest receiving targeted temperature management. *Reviews in Cardiovascular Medicine*, 24(2):55, 2023.

[16] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[17] Zheng Dai, Siru Liu, Jinfa Wu, Mengdie Li, Jialin Liu, and Ke Li. Analysis of adult disease characteristics and mortality on mimic-iii. *PloS one*, 15(4):e0232176, 2020.

[18] Vasiliki Danilatou, Stylianos Nikolakakis, Despoina Antonakaki, Christos Tzagkarakis, Dimitrios Mavroidis, Theodoros Kostoulas, and Sotirios Ioannidis. Outcome prediction in critically-ill patients with venous thromboembolism and/or cancer using machine learning algorithms: external validation and comparison with scoring systems. *International Journal of Molecular Sciences*, 23(13):7132, 2022.

[19] Yuhan Deng, Shuang Liu, Ziyao Wang, Yuxin Wang, Yong Jiang, and Baohua Liu. Explainable time-series deep learning models for the prediction of mortality, prolonged length of stay and 30-day readmission in intensive care patients. *Frontiers in Medicine*, 9:933037, 2022.

[20] Zhun Deng, Jiayao Zhang, Linjun Zhang, Ting Ye, Yates Coley, Weijie J Su, and James Zou. Fifa: Making fairness more generalizable in classifiers trained on imbalanced data. *arXiv preprint arXiv:2206.02792*, 2022.

[21] Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. Fairness guarantee in multi-class classification, 2023.

[22] Ning Ding, Cuirong Guo, Changluo Li, Yang Zhou, and Xiangping Chai. An artificial neural networks model for early predicting in-hospital mortality in acute pancreatitis in mimic-iii. *BioMed research international*, 2021, 2021.

[23] Yingying Fang, Chao Xiong, and Xinghe Wang. Association between early ondansetron administration and in-hospital mortality in critically ill patients: analysis of the mimic-iv database. *Journal of Translational Medicine*, 20(1):223, 2022.

[24] Mengling Feng, Jakob I McSparron, Dang Trung Kien, David J Stone, David H Roberts, Richard M Schwartzstein, Antoine Vieillard-Baron, and Leo Anthony Celi. Transthoracic echocardiography and mortality in sepsis: analysis of the mimic-iii database. *Intensive care medicine*, 44:884–892, 2018.

[25] Qizhang Feng, Mengnan Du, Na Zou, and Xia Hu. Fair machine learning in healthcare: A review. *arXiv preprint arXiv:2206.14397*, 2022.

[26] Haiyan Fu, Zhansheng Hu, Jianing Gong, Nan Li, Liu Na, Qiang Zhang, Shuying Wang, and Hongyang Du. The relationship between transthoracic echocardiography and mortality in adult patients with multiple organ dysfunction syndrome: analysis of the mimic-iii database. *Annals of Translational Medicine*, 10(6), 2022.

[27] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[28] Geert-Jan Geersing, Walter Bouwmeester, Peter Zuithoff, Rene Spijker, Mariska Leeflang, and Karel Moons. Search filters for finding prognostic and diagnostic prediction studies in medline to enhance systematic reviews. *PloS one*, 7(2):e32844, 2012.

[29] D Geethamani and R Rangaraj. Heterogeneous multi-model ensemble based length of stay prediction on mimic iii. 2022.

[30] Swapna Gokhale, David Taylor, Jaskirath Gill, Yanan Hu, Nikolajs Zeps, Vincent Lequertier, Helena Teede, and Joanne Enticott. Hospital length of stay prediction for general surgery and total knee arthroplasty admissions: Systematic review and meta-analysis of published prediction models. *Digital Health*, 9:20552076231177497, 2023.

[31] Fang Gong, Quan Zhou, Chunmei Gui, Shaohua Huang, and Zuoan Qin. The relationship between the serum anion gap and all-cause mortality in acute pancreatitis: an analysis of the mimic-iii database. *International Journal of General Medicine*, pages 531–538, 2021.

[32] Shenyan Gu, Yuqin Wang, Kaifu Ke, Xin Tong, Jiahui Gu, and Yuanyuan Zhang. Development and validation of a rass-related nomogram to predict the in-hospital mortality of neurocritical patients: a retrospective analysis based on the mimic-iv clinical database. *Current Medical Research and Opinion*, 38(11):1923–1933, 2022.

[33] Didi Han, Fengshuo Xu, Chengzhuo Li, Luming Zhang, Rui Yang, Shuai Zheng, Zichen Wang, Jun Lyu, et al. A novel nomogram for predicting survival in patients with severe acute pancreatitis: an analysis based on the large mimic-iii clinical database. *Emergency Medicine International*, 2021, 2021.

[34] Nianzong Hou, Mingzhe Li, Lu He, Bing Xie, Lin Wang, Rumin Zhang, Yong Yu, Xiaodong Sun, Zhengsheng Pan, and Kai Wang. Predicting 30-days mortality for mimic-iii patients with sepsis-3: a machine learning approach using xgboost. *Journal of translational medicine*, 18(1):1–14, 2020.

[35] Brian Hsu, Rahul Mazumder, Preetam Nandy, and Kinjal Basu. Pushing the limits of fairness impossibility: Who's the fairest of them all? *Advances in Neural Information Processing Systems*, 35:32749–32761, 2022.

[36] Bingsheng Huang, Dong Liang, Rushi Zou, Xiaxia Yu, Guo Dan, Haofan Huang, Heng Liu, and Yong Liu. Mortality prediction for patients with acute respiratory distress syndrome based on machine learning: a population-based study. *Annals of translational medicine*, 9(9), 2021.

[37] Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. *Computer methods and programs in biomedicine*, 177: 141–153, 2019.

[38] Xiaxuan Huang, Shiqi Yuan, Yitong Ling, Shanyuan Tan, Tao Huang, Hongtao Cheng, Jun Lyu, et al. The hemoglobin-to-red cell distribution width ratio to predict all-cause mortality in patients with sepsis-associated encephalopathy in the mimic-iv database. *International Journal of Clinical Practice*, 2022, 2022.

[39] Yan Huo, Xinrui Wang, Bo Li, Jordi Rello, Won Young Kim, Xiaoting Wang, and Zhenjie Hu. Impact of central venous pressure on the mortality of patients with sepsis-related acute kidney injury: a propensity score-matched analysis based on the mimic iv database. *Annals of Translational Medicine*, 10(4), 2022.

[40] Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3):364–373, 2020.

[41] Hong-Jie Jhou, Po-Huang Chen, Li-Yu Yang, Shu-Hao Chang, and Cho-Hao Lee. Plasma anion gap and risk of in-hospital mortality in patients with acute ischemic stroke: analysis from the mimic-iv database. *Journal of Personalized Medicine*, 11(10):1004, 2021.

[42] Yun Ji and Libin Li. Lower serum chloride concentrations are associated with increased risk of mortality in critically ill cirrhotic patients: an analysis of the mimic-iii database. *BMC gastroenterology*, 21(1):200, 2021.

[43] Wei Jiang, Chuanqing Zhang, Jiangquan Yu, Jun Shao, and Ruiqiang Zheng. Development and validation of a nomogram for predicting in-hospital mortality of elderly patients with persistent sepsis-associated acute kidney injury in intensive care units: a retrospective cohort study using the mimic-iv database. *BMJ open*, 13(3):e069824, 2023.

[44] Xiang Jiang, Weifan Dai, and Yanrong Cai. Comparison of machine learning algorithms to saps ii in predicting in-hospital mortality of fractures of the pelvis and acetabulum: analyzes based on mimic-iii database. *All Life*, 15(1):1000–1012, 2022.

[45] AE Johnson, Tom J Pollard, Lu Shen, L-w H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, L Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database sci. *Data*, 3(160035):10–1038, 2016.

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

[47] Guilan Kong, Ke Lin, and Yonghua Hu. Using machine learning methods to predict in-hospital mortality of sepsis patients in the icu. *BMC medical informatics and decision making*, 20:1–10, 2020.

[48] Alex Labach, Aslesha Pokhrel, Xiao Shi Huang, Saba Zuberi, Seung Eun Yi, Maksims Volkovs, Tomi Poutanen, and Rahul G Krishnan. Duett: Dual event time transformer for electronic health records. In *Machine Learning for Healthcare Conference*, pages 403–422. PMLR, 2023.

[49] Alex Labach Layer, Aslesha Pokhrel Layer, Xiao Shi Huang Layer, Saba Zuberi Layer, Seung Eun Yi Meta, Maksims Volkovs Layer, and Tomi Poutanen Signal. Duett: Dual event time transformer for electronic health records. 2023.

[50] Chengzhuo Li, Fengshuo Xu, Didi Han, Shuai Zheng, Wen Ma, Rui Yang, Zichen Wang, Yue Liu, and Jun Lyu. Developing and verifying a multivariate model to predict the survival probability after coronary artery bypass grafting in patients with coronary atherosclerosis based on the mimic-iii database. *Heart & Lung*, 52:61–70, 2022.

[51] Fei Li and Hong Yu. Icd coding from clinical text using multi-filter residual convolutional neural network. In *proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8180–8187, 2020.

[52] Fuchen Li, Patrick Wu, Henry H Ong, Josh F Peterson, Wei-Qi Wei, and Juan Zhao. Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction. *Journal of Biomedical Informatics*, 138:104294, 2023.

[53] Fuhai Li, Hui Xin, Jidong Zhang, Mingqiang Fu, Jingmin Zhou, and Zhexun Lian. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the mimic-iii database. *BMJ open*, 11(7):e044779, 2021.

[54] Xiao-Dan Li and Min-Min Li. A novel nomogram to predict mortality in patients with stroke: a survival analysis based on the mimic-iii clinical database. *BMC medical informatics and decision making*, 22(1): 92, 2022.

[55] Ke Lin, Yonghua Hu, and Guilan Kong. Predicting in-hospital mortality of patients with acute kidney injury in the icu using random forest model. *International journal of medical informatics*, 125:55–61, 2019.

[56] Caijie Liu, Shuying Wang, and Xiuzhen Wang. Effect of transthoracic echocardiography on short-term outcomes in patients with acute kidney injury in the intensive care unit: a retrospective cohort study based on the mimic-iii database. *Annals of translational medicine*, 10(15), 2022.

[57] Dongliang Liu, Yiyang Tang, and Qian Zhang. Admission hyperglycemia predicts long-term mortality in critically ill patients with subarachnoid hemorrhage: a retrospective analysis of the mimic-iii database. *Frontiers in Neurology*, 12:678998, 2021.

[58] En-qian Liu and Chun-lai Zeng. Blood urea nitrogen and in-hospital mortality in critically ill patients with cardiogenic shock: analysis of the mimic-iii database. *BioMed Research International*, 2021:1–7, 2021.

[59] Ran Liu, Haiwang Liu, Ling Li, Zhixue Wang, and Yan Li. Predicting in-hospital mortality for mimic-iii patients: A nomogram combined with sofa score. *Medicine*, 101(42):e31251, 2022.

[60] Suyun Liu and Luis Nunes Vicente. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science*, 19(3):513–537, 2022.

[61] Tao Liu, Haochen Xuan, Lili Wang, Xiaoqun Li, Zhihao Lu, Zhaoxuan Tian, Junhong Chen, Chaofan Wang, Dongye Li, and Tongda Xu. The association between serum albumin and long length of stay of patients with acute heart failure: A retrospective study based on the mimic-iv database. *Plos one*, 18(2): e0282289, 2023.

[62] Taotao Liu, Qinyu Zhao, and Bin Du. Effects of high-flow oxygen therapy on patients with hypoxemia after extubation and predictors of reintubation: a retrospective study based on the mimic-iv database. *BMC Pulmonary Medicine*, 21(1):1–15, 2021.

[63] Wei Liu, Wei Ma, Na Bai, Chunyan Li, Kuangpin Liu, Jinwei Yang, Sijia Zhang, Kewei Zhu, Qiang Zhou, Hua Liu, et al. Identification of key predictors of hospital mortality in critically ill patients with embolic stroke using machine learning. *Bioscience Reports*, 42(9):BSR20220995, 2022.

[64] Xiaobin Liu, Yu Zhao, Yingyi Qin, Dan Wang, Xi Yin, Renqi Yao, Qimin Ma, Yusong Wang, Shihui Zhu, Shaolin Ma, et al. A machine learning predictive model of in-hospital mortality in patients with sepsis complicated by anemia: a retrospective study based on the mimic-iii database. 2021.

[65] Xuefang Liu, Yanlin Feng, Xinyu Zhu, Ying Shi, Manting Lin, Xiaoyan Song, Jiancheng Tu, and Enwu Yuan. Serum anion gap at admission predicts all-cause mortality in critically ill patients with cerebral infarction: evidence from the mimic-iii database. *Biomarkers*, 25(8):725–732, 2020.

[66] Yan-Qiong Liu, Jin Hua Shen, BO Min, Quan Zhou, Xiang-Jie Duan, Ya Fen Guo, Xue Qing Zhang, et al. Relationship between the red cell distribution width-to-platelet ratio and in-hospital mortality among critically ill patients with acute myocardial infarction: a retrospective analysis of the mimic-iv database. *BMJ open*, 12(9):e062384, 2022.

[67] Yang Liu, Kun Gao, Hongbin Deng, Tong Ling, Jiajia Lin, Xianqiang Yu, Xiangwei Bo, Jing Zhou, Lin Gao, Peng Wang, et al. A time-incorporated sofa score-based machine learning model for predicting mortality in critically ill patients: a multicenter, real-world study. *International Journal of Medical Informatics*, 163:104776, 2022.

[68] Mengdi Luo, Yang Chen, Yuan Cheng, Na Li, and He Qing. Association between hematocrit and the 30-day mortality of patients with sepsis: A retrospective analysis based on the large-scale clinical database mimic-iv. *PloS one*, 17(3):e0265758, 2022.

[69] Andrés Alejandro Ramos Magna, Héctor Allende-Cid, Carla Taramasco, Carlos Becerra, and Rosa L Figueroa. Application of machine learning and word embeddings in the classification of cancer diagnosis using patient anamnesis. *Ieee Access*, 8:106198–106213, 2020.

[70] Saumil Maheshwari, Aman Agarwal, Anupam Shukla, and Ritu Tiwari. A comprehensive evaluation for the prediction of mortality in intensive care units with lstm networks: patients with cardiovascular disease. *Biomedical Engineering/Biomedizinische Technik*, 65(4):435–446, 2020.

[71] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[72] Chuizheng Meng, Loc Trinh, Nan Xu, and Yan Liu. Mimic-if: Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *arXiv preprint arXiv:2102.06761*, 2021.

[73] Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1):7166, 2022.

[74] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[75] Karel GM Moons, Douglas G Altman, Johannes B Reitsma, John PA Ioannidis, Petra Macaskill, Ewout W Steyerberg, Andrew J Vickers, David F Ransohoff, and Gary S Collins. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration. *Annals of internal medicine*, 162(1):W1–W73, 2015.

[76] Zhale Nowroozilarki, Arash Pakbin, James Royalty, Donald KK Lee, and Bobak J Mortazavi. Real-time mortality prediction using mimic-iv icu data via boosted nonparametric hazards. In *2021 IEEE EMBS international conference on biomedical and health informatics (BHI)*, pages 1–4. IEEE, 2021.

[77] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *International journal of surgery*, 88:105906, 2021.

[78] Ke Pang, Liang Li, Wen Ouyang, Xing Liu, and Yongzhong Tang. Establishment of icu mortality risk prediction models with machine learning algorithm using mimic-iv database. *Diagnostics*, 12(5):1068, 2022.

[79] Jiang-Chen Peng, Fang Nie, Yu-Jie Li, Qiao-Yi Xu, Shun-Peng Xing, Wen Li, and Yuan Gao. Favorable outcomes of anticoagulation with unfractioned heparin in sepsis-induced coagulopathy: a retrospective analysis of mimic-iii database. *Frontiers in Medicine*, 8:773339, 2022.

[80] Tom J Pollard, Alistair E W Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[81] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.

[82] Zuoan Qin, Nuohan Liao, Xuelin Lu, Xiangjie Duan, Quan Zhou, and Liangqing Ge. Relationship between the hemoglobin-to-red cell distribution width ratio and all-cause mortality in ischemic stroke patients with atrial fibrillation: an analysis from the mimic-iv database. *Neuropsychiatric Disease and Treatment*, 18:341, 2022.

[83] Manish Raghavan. What should we do when our ideas of fairness conflict? *Communications of the ACM*, 67(1):88–97, 2023.

[84] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

[85] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[86] Eliane Röösli, Brian Rice, and Tina Hernandez-Boussard. Bias at warp speed: how ai may contribute to the disparities gap in the time of covid-19. *Journal of the American Medical Informatics Association*, 28 (1):190–192, 2021.

[87] Eliane Röösli, Selen Bozkurt, and Tina Hernandez-Boussard. Peeking into a black box, the fairness and generalizability of a mimic-iii benchmarking model. *Scientific Data*, 9(1):24, 2022.

[88] Gerta Salillari and Nadav Rappoport. Comparison of classification with reject option approaches on mimic-iv dataset. In *International Conference on Artificial Intelligence in Medicine*, pages 210–219. Springer, 2022.

[89] Veit Sandfort, Alistair EW Johnson, Lauren M Kunz, Jose D Vargas, and Douglas R Rosing. Prolonged elevated heart rate and 90-day survival in acutely ill patients: data from the mimic-iii database. *Journal of intensive care medicine*, 34(8):622–629, 2019.

[90] James Sanii and Wai Yip Chan. Explainable machine learning models for pneumonia mortality risk prediction using mimic-iii data. In *2022 9th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pages 68–73. IEEE, 2022.

[91] Kailash Karthik Saravanakumar. The impossibility theorem of machine fairness – a causal perspective, 2021.

[92] Mohammed Sayed, David Riano, and Jesús Villar. Predicting duration of mechanical ventilation in acute respiratory distress syndrome using supervised machine learning. *Journal of Clinical Medicine*, 10(17): 3824, 2021.

[93] Yasser Selim, Élise Di Lena, Nawaf Abu-Omar, Zarrukh Baig, Kevin Verhoeff, Julie La, Kieran Purich, Samantha Albacete, Rahim Valji, Ali Safar, et al., 2022.

[94] H Shi, S-Y Sun, Y-S He, and Q Peng. Association between early vasopressor administration and in-hospital mortality in critically ill patients with acute pancreatitis: A cohort study from the mimic-iv database. *European Review for Medical & Pharmacological Sciences*, 27(2), 2023.

[95] Tingting Shu, Jian Huang, Jiewen Deng, Huaqiao Chen, Yang Zhang, Minjie Duan, Yanqing Wang, Xiaofei Hu, and Xiaozhu Liu. Development and assessment of scoring model for icu stay and mortality prediction after emergency admissions in ischemic heart disease: a retrospective study of mimic-iv databases. *Internal and Emergency Medicine*, 18(2):487–497, 2023.

[96] Yingjie Su, Cuirong Guo, Shifang Zhou, Changluo Li, and Ning Ding. Early predicting 30-day mortality in sepsis in mimic-iii by an artificial neural networks model. *European Journal of Medical Research*, 27 (1):294, 2022.

[97] Chun Sun, Deqing Chen, Xin Jin, Guangtao Xu, Chenye Tang, Xiao Guo, Zhiling Tang, Yixin Bao, Fei Wang, and Ruilin Shen. Association between acute kidney injury and prognoses of cardiac surgery patients: Analysis of the mimic-iii database. *Frontiers in Surgery*, 9, 2022.

[98] Wen Sun, Yang Yan, Shidong Hu, Boyan Liu, Shuying Wang, Wenli Yu, and Songyan Li. The effects of midazolam or propofol plus fentanyl on icu mortality: a retrospective study based on the mimic-iv database. *Annals of Translational Medicine*, 10(4), 2022.

[99] Hai Tang, Zhuochen Jin, Jiajun Deng, Yunlang She, Yifan Zhong, Weiyan Sun, Yijiu Ren, Nan Cao, and Chang Chen. Development and validation of a deep learning model to predict the survival of patients in icu. *Journal of the American Medical Informatics Association*, 29(9):1567–1576, 2022.

[100] Rui Tang, Wen Tang, and Daoxin Wang. Predictive value of machine learning for in-hospital mortality for trauma-induced acute respiratory distress syndrome patients: an analysis using the data from mimic iii. *Zhonghua wei Zhong Bing ji jiu yi xue*, 34(3):260–264, 2022.

[101] Jared Thacker. *A Machine Learning Pipeline for Readmission Prediction with MIMIC-III*. PhD thesis, Auburn University, 2023.

[102] Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6): 1–17, 2022.

[103] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[104] Evan J Tsiklidis, Talid Sinno, and Scott L Diamond. Predicting risk for trauma patients using static and dynamic information from the mimic iii database. *Plos one*, 17(1):e0262523, 2022.

[105] Jeffrey L Tully, William Zhong, Sierra Simpson, Brian P Curran, Alvaro A Macias, Ruth S Waterman, and Rodney A Gabriel. Machine learning prediction models to reduce length of stay at ambulatory surgery centers through case resequencing. *Journal of Medical Systems*, 47(1):71, 2023.

[106] Chunxia Wang, Jianli Zheng, Jinxia Wang, Lin Zou, and Yucai Zhang. Cox-lasso analysis for hospital mortality in patients with sepsis received continuous renal replacement therapy: a mimic-iii database study. *Frontiers in Medicine*, 8:778536, 2022.

[107] Dongyan Wang, Xiaoyan Guo, Wenwen Xia, Zhijuan Ru, Yihai Shi, and Zhengyu Hu. Effect of admission serum calcium levels and length of stay in patients with acute pancreatitis: Data from the mimic-iii database. *Emergency Medicine International*, 2022, 2022.

[108] Dongyan Wang, Jie Lu, Pan Zhang, Zhengyu Hu, and Yihai Shi. Relationship between blood glucose levels and length of hospital stay in patients with acute pancreatitis: An analysis of mimic-iii database. *Clinical and Translational Science*, 16(2):246–257, 2023.

[109] Junjie Wang, Lingqu Zhou, Yinyin Zhang, Haifeng Zhang, Yong Xie, Zhiteng Chen, Boshui Huang, Kuan Zeng, Juan Lei, Jingting Mai, et al. Minimum heart rate and mortality in critically ill myocardial infarction patients: an analysis of the mimic-iii database. *Annals of translational medicine*, 9(6), 2021.

[110] Shaosheng Wu, Xiaoting Shi, Quan Zhou, Xiangjie Duan, Xiongfei Zhang, and Huajing Guo. The association between systemic immune-inflammation index and all-cause mortality in acute ischemic stroke patients: analysis from the mimic-iv database. *Emergency medicine international*, 2022:1–10, 2022.

[111] Zuoxun Xia, Peng Xu, Ye Xiong, Yunbo Lai, Zhaohui Huang, et al. Survival prediction in patients with hypertensive chronic kidney disease in intensive care unit: A retrospective analysis based on the mimic-iii database. *Journal of Immunology Research*, 2022, 2022.

[112] Wanqiu Xie, Yue Li, Xianglin Meng, and Mingyan Zhao. Machine learning prediction models and nomogram to predict the risk of in-hospital death for severe dka: A clinical study based on mimic-iv, eicu databases, and a college hospital icu. *International Journal of Medical Informatics*, 174:105049, 2023.

[113] Jinghong Xu, Li Tong, Jiyou Yao, Zilu Guo, Ka Yin Lui, XiaoGuang Hu, Lu Cao, Yanping Zhu, Fa Huang, Xiangdong Guan, et al. Association of sex with clinical outcome in critically ill sepsis patients: a retrospective analysis of the large clinical database mimic-iii. *Shock*, 52(2):146–151, 2019.

[114] Jun Xu, Hongliu Cai, and Xia Zheng. Timing of vasopressin initiation and mortality in patients with septic shock: analysis of the mimic-iii and mimic-iv databases. *BMC Infectious Diseases*, 23(1):199, 2023.

[115] Cheng-Chang Yang, Oluwaseun Adebayo Bamodu, Lung Chan, Jia-Hung Chen, Chien-Tai Hong, Yi-Ting Huang, and Chen-Chih Chung. Risk factor identification and prediction models for prolonged length of stay in hospital after acute ischemic stroke using artificial neural networks. *Frontiers in Neurology*, 14: 1085178, 2023.

[116] Siyue Yang, Paul Varghese, Ellen Stephenson, Karen Tu, and Jessica Gronsbell. Machine learning approaches for electronic health records phenotyping: a methodical review. *Journal of the American Medical Informatics Association*, 30(2):367–381, 2023.

[117] Wei Yang, Hong Zou, Meng Wang, Qin Zhang, Shadan Li, and Hongyin Liang. Mortality prediction among icu inpatients based on mimic-iii database results from the conditional medical generative adversarial network. *Heliyon*, 9(2), 2023.

[118] Weiyan Ye, Xiaoli Chen, Yongbo Huang, Yuchong Li, Yonghao Xu, Zhenting Liang, Danlin Wu, Xiaoqing Liu, and Yimin Li. The association between neutrophil-to-lymphocyte count ratio and mortality in septic patients: a retrospective analysis of the mimic-iii database. *Journal of Thoracic Disease*, 12(5):1843, 2020.

[119] Zixiang Ye, Shuoyan An, Yanxiang Gao, Enmin Xie, Xuecheng Zhao, Ziyu Guo, Yike Li, Nan Shen, Jingyi Ren, and Jingang Zheng. The prediction of in-hospital mortality in chronic kidney disease patients with coronary artery disease using machine learning models. *European Journal of Medical Research*, 28 (1):1–13, 2023.

[120] Yue Yu, Jun Wang, Qing Wang, Junnan Wang, Jie Min, Suyu Wang, Pei Wang, Renhong Huang, Jian Xiao, Yufeng Zhang, et al. Admission oxygen saturation and all-cause in-hospital mortality in acute myocardial infarction patients: data from the mimic-iii database. *Annals of translational medicine*, 8(21), 2020.

[121] Suru Yue, Shasha Li, Xueying Huang, Jie Liu, Xuefei Hou, Yumei Zhao, Dongdong Niu, Yufeng Wang, Wenkai Tan, and Jiayuan Wu. Machine learning for the prediction of acute kidney injury in patients with sepsis. *Journal of translational medicine*, 20(1):1–12, 2022.

[122] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[123] Zhixuan Zeng, Shuo Yao, Jianfei Zheng, and Xun Gong. Development and validation of a novel blending machine learning model for hospital mortality prediction in icu patients with sepsis. *BioData mining*, 14: 1–15, 2021.

[124] HuanRui Zhang, Wen Tian, and YuJiao Sun. The value of anion gap for predicting the short-term all-cause mortality of critically ill patients with cardiac diseases, based on mimic-iii database. *Heart & Lung*, 55: 59–67, 2022.

[125] Jingqing Zhang, Luis Daniel Bolanos Trujillo, Ashwani Tanwar, Julia Ive, Vibhor Gupta, and Yike Guo. Clinical utility of automatic phenotype annotation in unstructured clinical notes: intensive care unit use. *BMJ Health & Care Informatics*, 29(1):e100519, 2022.

[126] Rongting Zhang, Shanshan Shi, Weihua Chen, Yani Wang, Xueqin Lin, Yukun Zhao, Lihua Liao, Qian Guo, Xiaoying Zhang, Weiguo Li, et al. Independent effects of the triglyceride-glucose index on all-cause mortality in critically ill patients with coronary heart disease: analysis of the mimic-iii database. *Cardiovascular Diabetology*, 22(1):10, 2023.

[127] Tang Zhang, Yao-Zong Guan, and Hao Liu. Association of acidemia with short-term mortality of acute myocardial infarction: a retrospective study base on mimic-iii database. *Clinical and Applied Thrombosis/Hemostasis*, 26:1076029620950837, 2020.

[128] Lina Zhao, Jing Yang, Cong Zhou, Yunying Wang, and Tao Liu. A novel prognostic model for predicting the mortality risk of patients with sepsis-related acute respiratory failure: a cohort study using the mimic-iv database. *Current Medical Research and Opinion*, 38(4):629–636, 2022.

[129] Yu Zhao, Rusen Zhang, Yi Zhong, Jingjing Wang, Zuquan Weng, Heng Luo, and Cunrong Chen. Statistical analysis and machine learning prediction of disease outcomes for covid-19 and pneumonia patients. *Frontiers in cellular and infection microbiology*, 12:838749, 2022.

[130] Shiyu Zhou, Zhenhua Zeng, Hongxia Wei, Tong Sha, and Shengli An. Early combination of albumin with crystalloids administration might be beneficial for the survival of septic patients: a retrospective analysis from mimic-iv database. *Annals of intensive care*, 11:1–10, 2021.

[131] Yao Zhu, Zonglin He, Ya Jin, Sui Zhu, Weipeng Xu, Bingxiao Li, Chuan Nie, Guosheng Liu, Jun Lyu, and Shasha Han. Serum anion gap level predicts all-cause mortality in septic patients: A retrospective study based on the mimic iii database. *Journal of Intensive Care Medicine*, 38(4):349–357, 2023.

[132] James Zou and Londa Schiebinger. Design ai so that it's fair. *Nature*, 559(7714):324–326, 2018.

[133] Zhi-ye Zou, Jia-jia Huang, Ying-yi Luan, Zhen-jia Yang, Zhi-peng Zhou, Jing-jing Zhang, Yong-ming Yao, and Ming Wu. Early prophylactic anticoagulation with heparin alleviates mortality in critically ill patients with sepsis: a retrospective analysis from the mimic-iv database. *Burns & Trauma*, 10:tkac029, 2022.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 5.
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 5.2
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 5.3.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5.2.
3. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 5.3.
   (b) Did you mention the license of the assets? [Yes] Listed in detail in our "Datasheet for database'.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] 'Model card' and 'Fairness report' for HML risk prediction models.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Obtained access from Physionet owners to work on this study.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We worked with already publicly available data.

# A Appendix

## A.1 Systematic survey

PubMed is searched using *('Medical information mart for intensive care AND MIMIC AND MIMIC-IV') AND ('machine learning' OR 'artificial intelligence' OR 'deep learning' OR 'neural network' OR 'prediction model')* search terms yielding 819 works whereas, Google Scholar yielded 1000 records with the search term *'machine learning'|'prediction model'|'artificial intelligence'|'deep learning' AND 'MIMIC IV'|'the medical information mart for intensive care'|'MIMIC III'*.

We abstracted information on:

Table 3: Inclusion and Exclusion Criteria of MIMIC trained Healthcare ML Studies.

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

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

### A.2 HML Risk prediction tasks

1. **Mortality Prediction** [1–4, 8, 9, 14, 17, 18, 23, 24, 26, 31–34, 36, 38, 39, 41–44, 47, 50, 53–56, 58, 59, 61, 63–67, 70, 76, 78, 79, 81, 82, 88–90, 94–96, 98–100, 104, 106, 109–114, 117–120, 123, 124, 126–131, 133] - Predict the likelihood of a patient dying.
   (a) In-hospital and ICU [22, 117] - Estimating the risk of a patient dying during their hospital stay and in ICU.
   (b) Short-term [68, 124] - Assessing the risk of death shortly after ICU admission, typically within 2-3 days.
   (c) *Long-term* [1, 57] - Evaluating the likelihood of death over a longer period, typically from 30 days up to a year after hospital discharge.
2. **Length of Stay (LOS)** [15, 19, 29, 30, 61, 93, 95, 95, 105, 107, 108, 115]
   (a) Predicting the duration of hospital stays for admissions, focusing on stays longer than 3 and 7 days. A custom number of days is also a topic of research interest.
3. **Readmission** [13, 101]
   (a) Identifying patients at risk of returning to the hospital within 30, 60, 90, 120 days, or custom time frames after discharge.
4. **Phenotype Labeling and ICD-9/10 Code Grouping**
   (a) *Phenotype Labeling* [116, 125] - Classifying patients into specific groups based on clinical data for disease prediction and treatment customization.
   (b) *ICD Code Grouping* [37, 51] - Categorizing diseases based on diagnosis codes to streamline classification.
5. **Specific Health Conditions**
   (a) *Heart failure* [2, 53] - Predicting the occurrence, progression, and prognosis of heart failure in patients, using historical and real-time health data.
   (b) *Chronic Kidney Disease (CKD)* [97, 121] - Assessing the risk and progression of CKD to inform treatment plans and manage patient health outcomes.
   (c) *Chronic obstructive pulmonary disease (COPD)* [62] - Forecasting COPD exacerbations and identifying patients at higher risk for hospitalization or severe outcomes.
   (d) *Coronary artery disease (CAD)* [117, 119] - Utilizing clinical data to predict the development or worsening of CAD for early intervention.
   (e) *Sepsis* [5, 121] - Early detection and prediction of sepsis in hospitalized patients to improve treatment response and survival rates.
   (f) *Cancer* [69] - Leveraging patient data to predict cancer risks, progression, and treatment responses.
   (g) *Ventilation failure* [92] - Predicting the risk of ventilation failure in critically ill patients to guide intervention strategies.

### A.3 Sensitive attribute statistics of MIMIC IV ICU Mortality data

**Sensitive attribute statistics of MIMIC IV ICU Mortality data**: We followed the work of Meng et al. [72] and grouped Ethnic demography into 5 categories based on the geographic origin. Data contains 68% of the White population followed by the Other (13.8%) subgroup. Age is categorized into 6 buckets 1 following the work of Röösli et al. [87]. 24% are of 30 to 49 and 50 to 69 age bins followed by 70 to 80 (23.6%). Children below 17 are not part of this CRD. Gender-wise, only Males and Females are recorded, while 56% of the patients are Males. Medicaid, Medicare, and Other were the 3 types of Insurance listed, with English being the only language recorded, whereas others are left as '?'.

All the TS features and the above-mentioned static variables are listed below. There are inconsistencies like in-hospital expiry information and hospital admit and discharge timestamps, detailed in the datasheet. The admission table contains multiple reports of the same patient's death, leading to inconsistencies. However, the owners of MIMIC have mentioned the potential for bias within the CRD.

Table 4: Time series and Static features of MIMIC IV used to train all the analyzed models for ICU mortality task

| **Variable** | **Type** |
|---|---|
| minute | Time series |
| ALP | Time series |
| ALT | Time series |
| AST | Time series |
| Albumin | Time series |
| Albumin 25% | Time series |
| Albumin 5% | Time series |
| Amiodarone | Time series |
| Anion Gap | Time series |
| BUN | Time series |
| Base Excess | Time series |
| Basophils | Time series |
| Bicarbonate | Time series |
| Bilirubin (Direct) | Time series |
| Bilirubin (Indirect) | Time series |
| Bilirubin (Total) | Time series |
| CRR | Time series |
| Calcium Free | Time series |
| Calcium Gluconate | Time series |
| Calcium Total | Time series |
| Cefazolin | Time series |
| Chest Tube | Time series |
| Chloride | Time series |
| Colloid | Time series |
| Creatinine Blood | Time series |
| Creatinine Urine | Time series |
| D5W | Time series |
| DBP | Time series |
| Dextrose Other | Time series |
| Dopamine | Time series |
| EBL | Time series |
| Emesis | Time series |
| Eoisinophils | Time series |
| Epinephrine | Time series |
| Famotidine | Time series |
| Fentanyl | Time series |
| FiO2 | Time series |
| Fiber | Time series |

| Variable | Type |
|---|---|
| Free Water | Time series |
| Fresh Frozen Plasma | Time series |
| Furosemide | Time series |
| GCS_eye | Time series |
| GCS_motor | Time series |
| GCS_verbal | Time series |
| GT Flush | Time series |
| Gastric | Time series |
| Gastric Meds | Time series |
| Glucose (Blood) | Time series |
| Glucose (Serum) | Time series |
| Glucose (Whole Blood) | Time series |
| HR | Time series |
| Half Normal Saline | Time series |
| Hct | Time series |
| Height | Time series |
| Heparin | Time series |
| Hgb | Time series |
| Hydralazine | Time series |
| Hydromorphone | Time series |
| INR | Time series |
| Insulin Humalog | Time series |
| Insulin NPH | Time series |
| Insulin Regular | Time series |
| Insulin largine | Time series |
| Jackson-Pratt | Time series |
| KCl | Time series |
| KCl (Bolus) | Time series |
| LDH | Time series |
| Lactate | Time series |
| Lactated Ringers | Time series |
| Lorazepam | Time series |
| Lymphocytes | Time series |
| Lymphocytes (Absolute) | Time series |
| MBP | Time series |
| MCH | Time series |
| MCHC | Time series |
| MCV | Time series |
| Magnesium | Time series |
| Magnesium Sulfate (Bolus) | Time series |
| Magnesium Sulphate | Time series |
| Metoprolol | Time series |
| Midazolam | Time series |
| Milrinone | Time series |
| Monocytes | Time series |
| Morphine Sulfate | Time series |
| Neosynephrine | Time series |
| Neutrophils | Time series |
| Nitroglycerine | Time series |
| Nitroprusside | Time series |
| Norepinephrine | Time series |
| Normal Saline | Time series |
| O2 Saturation | Time series |

| Variable | Type |
|---|---|
| OR/PACU | Time series |
| Crystalloid | |
| PCO2 | Time series |
| PO intake | Time series |
| PO2 | Time series |
| PT | Time series |
| PTT | Time series |
| Packed RBC | Time series |
| Pantoprazole | Time series |
| Phosphate | Time series |
| Piggyback | Time series |
| Piperacillin | Time series |
| Platelet Count | Time series |
| Potassium | Time series |
| Pre-admission Intake | Time series |
| Pre-admission | Time series |
| Output | |
| Propofol | Time series |
| RBC | Time series |
| RDW | Time series |
| RR | Time series |
| Residual | Time series |
| SBP | Time series |
| SG Urine | Time series |
| Sodium | Time series |
| Solution | Time series |
| Sterile Water | Time series |
| Stool | Time series |
| TPN | Time series |
| Temperature | Time series |
| Total CO2 | Time series |
| Urine | Time series |
| Vacomycin | Time series |
| Vasopressin | Time series |
| WBC | Time series |
| Weight | Time series |
| pH Blood | Time series |
| pH Urine | Time series |
| first_careunit | Static |
| gender | Static |
| anchor_age | Static |
| insurance | Static |
| language | Static |
| marital_status | Static |
| ethnicity | Static |

### A.4 Datasheet for Database

### A.5 Fairness metrics Results for HML models

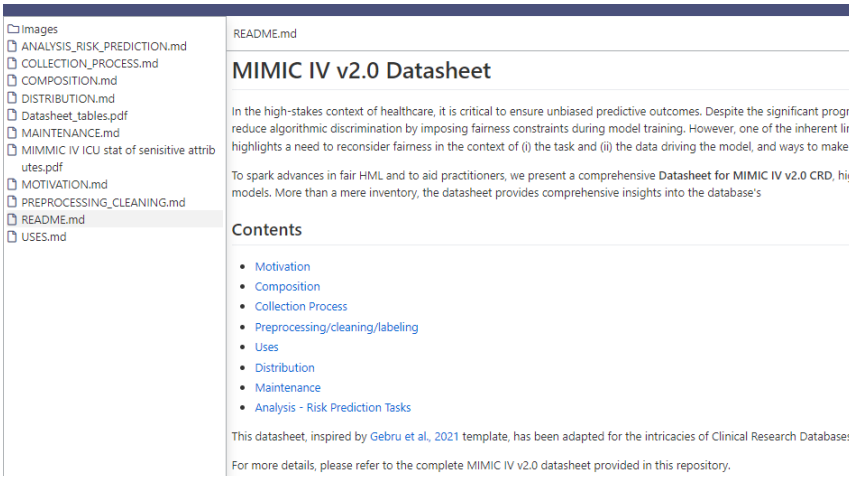



Figure 5: Snapshot of our comprehensive datasheet webpage, tailored for the MIMIC IV v2.0 database.

Table 5: Evaluation of Fairness Metrics (Demographic Parity (DP), Equalized Odds for True Positive Rate (TPR)/Equalized Opportunity (EOp), and Equalized Odds for False Positive Rate (FPR)) Across Model Types: This table presents a comparative analysis of fairness metrics for different models, stratified by ethnic groups.

| **Type** | **Models** | **Ethnicity** | **DP** $\leftrightarrow$ | **EO(TPR)/ EOp** $\uparrow \leftrightarrow$ | **EO(FPR)** $\downarrow \leftrightarrow$ |
|---|---|---|---|---|---|
| Static | Logistic Regression | White | 0.230 | 0.624 | 0.202 |
| | | Other | 0.309 | 0.828 | 0.249 |
| | | Black | 0.265 | 0.825 | 0.225 |
| | | Hispanic/Latino | 0.251 | 0.583 | 0.230 |
| | | Asian | 0.285 | 0.785 | 0.240 |
| | **XG Boost** | White | 0.196 | 0.858 | 0.145 |
| | | **Other** | **0.267** | **0.936** | **0.171** |
| | | Black | 0.192 | 0.866 | 0.145 |
| | | Hispanic/Latino | 0.139 | 0.923 | 0.112 |
| | | Asian | 0.225 | 0.903 | 0.149 |
| Time Series | LSTM | White | 0.460 | 0.709 | 0.215 |
| | | Other | 0.534 | 0.759 | 0.275 |
| | | Black | 0.421 | 0.716 | 0.236 |
| | | Hispanic/Latino | 0.426 | 0.868 | 0.159 |
| | | Asian | 0.545 | 0.739 | 0.333 |
| | STraTS | White | 0.289 | 0.799 | 0.230 |
| | | Other | 0.352 | 0.876 | 0.241 |
| | | Black | 0.304 | 0.871 | 0.252 |
| | | Hispanic/Latino | 0.271 | 0.812 | 0.214 |
| | | Asian | 0.299 | 0.816 | 0.209 |
| | **DuETT (SOTA)** | White | 0.473 | 0.410 | 0.485 |
| | | Other | 0.525 | 0.470 | 0.541 |
| | | Black | 0.553 | 0.520 | 0.560 |
| | | **Hispanic/Latino** | **0.600** | **0.580** | **0.603** |
| | | Asian | 0.525 | 0.511 | 0.527 |