ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1334

Fairness Audit and Debiasing of Machine Learning Models for Sepsis Mortality Prediction Across Demographic Subgroups: A Multi-Metric Study with Accuracy–Fairness Tradeoff Analysis

Francis Mawutor Amuyao[1*]; Isaac Tosin Adisa[2]

[1,2]Department of Statistics, Florida State University, Tallahassee, FL 32306, USA

Corresponding Author: Francis Mawutor Amuyao*
117 N. Woodward Ave., Tallahassee, FL 32306, USA

Publication Date: 2026/08/24

Abstract: Machine learning (ML) models for sepsis mortality prediction are increasingly deployed in intensive care units, yet performance across demographic subgroups remains poorly evaluated, and algorithmic disparities can directly exacerbate health inequities in critical care settings. We trained three ML models—logistic regression (LR), XGBoost, and a multilayer perceptron (MLP)—on a 10,000-patient cohort calibrated to published MIMIC-IV sepsis statistics, and performed a four-metric fairness audit (equalized odds difference [EOD], demographic parity difference, predictive parity gap, and subgroup calibration error) across race/ethnicity, sex, and insurance type. Per-group threshold optimisation was applied for debiasing, and an accuracy–fairness Pareto tradeoff was quantified. Black patients exhibited the highest mortality (57.2% vs 46.6% for White patients) and the largest fairness disparities. LR showed the worst racial bias (EOD = 0.193); XGBoost was most equitable (EOD = 0.148). Time to antibiotic administration carried the largest Black–White SHAP gap (0.045). Debiasing reduced EOD by 31.3% for LR at zero AUROC cost (0.678). Sepsis ML models exhibit clinically meaningful demographic disparities that vary by architecture and metric. The accuracy–fairness Pareto frontier provides a deployment decision tool, and our Fairness-SHAP framework identifies features driving between-group disparities, enabling targeted intervention.

Keywords: *Algorithmic Fairness; Sepsis Prediction; Machine Learning; MIMIC-IV; Equalized Odds; SHAP; Health Disparities; Debiasing; ICU.*

How to Cite: Francis Mawutor Amuyao; Isaac Tosin Adisa (2026) Fairness Audit and Debiasing of Machine Learning Models for Sepsis Mortality Prediction Across Demographic Subgroups: A Multi-Metric Study with Accuracy–Fairness Tradeoff Analysis. *International Journal of Innovative Science and Research Technology*, 11(7), 3834-3839. https://doi.org/10.38124/ijisrt/26jul1334

I. INTRODUCTION

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection, accounting for over 270,000 deaths annually in the United States [1,2]. Each hour of delayed antibiotic therapy increases mortality by approximately 7% [3]. Machine learning (ML) models have emerged as promising tools for early risk stratification, with recent studies reporting AUROCs exceeding 0.87 on MIMIC-IV cohorts [4,5]. Algorithmic fairness—the consistency of model performance across demographic subgroups—represents a critical and underexplored dimension of clinical ML deployment.

In a landmark analysis, Obermeyer et al. [6] demonstrated that a commercial healthcare algorithm systematically underestimated illness severity in Black patients, directing fewer resources toward a population with objectively greater need. Black patients experience higher sepsis incidence, more severe presentation, and documented longer times to antibiotic initiation, patterns attributable to structural inequities in care delivery rather than biological differences [7,8].

The existing literature on sepsis ML fairness is fragmented. Most papers report only subgroup AUROC—insufficient because it cannot distinguish between differential false negative rates (missed deaths) and differential false positive rates (alarm burden). Debiasing is rarely applied, and when it is, the accuracy cost is not quantified. Feature-level attribution of fairness disparities, the question of which clinical variables drive the bias, remains almost entirely unaddressed.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1334

This study addresses these gaps with four contributions: (i) a comprehensive four-metric fairness audit across three model architectures, three protected attributes, and five racial/ethnic subgroups; (ii) post-processing debiasing via per-group equalized odds threshold optimisation; (iii) the first accuracy–fairness Pareto frontier published for sepsis mortality prediction; and (iv) a Fairness-SHAP attribution framework identifying which clinical features drive between-group prediction disparities.

II. METHODS

A. Dataset and Cohort

A 10,000-patient ICU cohort was constructed and calibrated to the epidemiological characteristics of the published MIMIC-IV sepsis literature [9]. Inclusion criteria mirrored the Sepsis-3 definition: adult ICU patients (age ≥ 18 years) with SOFA score ≥ 2 and suspected infection, operationalised by concurrent antibiotic administration and blood culture ordering [10]. Race/ethnicity distribution followed MIMIC-IV ICU demographics (White 55%, Black 22%, Hispanic 10%, Asian 7%, Other 6%). Black patients exhibited higher SOFA scores at presentation (mean 6.1 vs 5.2) and longer times to antibiotic administration (mean 3.2h vs 2.1h), consistent with documented structural care inequities [7,8]. Sixteen clinical features were extracted: age; heart rate, mean arterial pressure, SpO2, temperature, respiratory rate; WBC, lactate, creatinine, bilirubin, platelet; SOFA score; Charlson Comorbidity Index; vasopressor use; mechanical ventilation; and time to antibiotics. Protected attributes (race, sex, insurance type) were recorded for auditing only and were excluded from all model inputs [11].

B. Preprocessing

Missing data were addressed using Multivariate Imputation by Chained Equations (MICE) via scikit-learn's IterativeImputer (10 iterations, non-negative lower bound). The cohort was partitioned into training (70%), validation (15%), and test (15%) sets using a stratified shuffle split on the joint outcome-race key. Feature scaling used RobustScaler fitted on training data only.

C. Model Training

Three architectures were trained. Logistic Regression (LR) with ℓ_2 regularisation; C selected by 5-fold cross-validation (best C = 0.01, CV-AUROC = 0.687). XGBoost with up to 800 estimators, early stopping at 30 rounds on validation AUROC, max depth 5, learning rate $\eta = 0.05$ (best iteration: 63). Multilayer Perceptron (MLP) with architecture (128, 64) and $\alpha = 10^{-3}$ selected by cross-validation (CV-AUROC = 0.680).

D. Performance Evaluation

Discrimination was assessed using AUROC and AUPRC. Calibration was assessed using Brier score and Expected Calibration Error (ECE), defined as ECE = Σ (over bins b = 1...B) of (|S_b|/n) × |mean observed frequency in S_b − mean predicted probability in S_b|, where S_b is the sample set in bin b.

E. Fairness Audit

For each model, protected attribute, and subgroup, five metrics were computed at threshold $\tau = 0.5$ (Table 1). The equalized odds difference is EOD(g, g_ref) = max(|TPR_g − TPR_gref|, |FPR_g − FPR_gref|). Reference groups: White (race), Male (sex), Medicare (insurance). All five metrics are reported simultaneously, following Chouldechova [12] and Hardt et al. [13], given the impossibility of simultaneous criterion satisfaction when base rates differ.

F. Debiasing

Post-processing debiasing was applied by finding, for each non-reference group g, a threshold τ^*_g minimising the TPR difference from the reference group: $\tau^*_g = \text{argmin}_\tau |TPR_g(\tau) - TPR_{White}(0.5)|$, via grid search over 50 candidates in [0.1, 0.9]. The Pareto tradeoff curve was produced by sweeping the global threshold from 0.05 to 0.95 and recording (EOD, AUROC) at each point.

G. Fairness-SHAP Attribution

The Fairness-SHAP gap for feature f is $\Delta SHAP(f) = E[|\varphi_f(x)|]_{Black} - E[|\varphi_f(x)|]_{White}$, where $\varphi_f(x)$ is the SHAP value of feature f for sample x. A positive gap indicates higher model attention to f for Black patients. TreeExplainer was used for XGBoost; KernelExplainer (200 background samples) for LR.

III. RESULTS

A. Cohort Characteristics

The cohort comprised 10,000 patients (mean age 63 ± 16 years, 57% male, mortality 49.2%). On the test set, Black patients had the highest mortality (57.2%) versus White (46.6%), Hispanic (51.7%), Asian (45.0%), and Other (45.6%) patients, consistent with population-level estimates [7].

Table 1: Fairness Metrics Used in the Audit

Metric	Definition	Clinical Interpretation
Equalized Odds Difference (EOD)	max(|ΔTPR|, |ΔFPR|) vs reference	Are deaths missed equally across groups?
Demographic Parity Difference (DPD)	|Δ Pr($\hat{Y}$=1)| vs reference	Are alarm rates equal?
Predictive Parity (PPV gap)	|ΔPPV| vs reference	When alarm fires, equally reliable?
Subgroup ECE	ECE restricted to subgroup members	Do probabilities reflect true risk?
Subgroup AUROC	AUROC restricted to subgroup	Overall discrimination per group

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1334

B. Model Performance

Test-set performance is presented in Table 2. LR achieved the best AUROC (0.678; AUPRC 0.662; ECE 0.020). XGBoost was comparable (AUROC 0.658; ECE 0.018). The MLP exhibited severe overfitting (train AUROC 0.9999 vs test 0.536; ECE 0.423). The modest AUROCs reflect deliberate restriction to admission-time features, avoiding temporal data leakage common in studies reporting AUROCs > 0.85.

Table 2: Overall Model Performance on the Held-Out Test Set (n = 1,500)

Model	AUROC	AUPRC	Brier	ECE
Logistic Regression	0.678*	0.662*	0.226*	0.020
XGBoost	0.658	0.642	0.231	0.018*
MLP	0.536	0.530	0.436	0.423

C. Fairness Audit

Race was the most consequential protected attribute across all models and metrics (Table 3). For LR, Black patients had EOD = 0.193 (TPR gap 0.124, FPR gap 0.192)—meaning LR missed 12.4 percentage points more sepsis deaths in Black patients than White patients. XGBoost showed lower but still substantial racial bias (EOD = 0.148). Hispanic patients showed the second-largest disparities (LR EOD = 0.095; XGBoost EOD = 0.126). Insurance type produced the second-largest disparities overall; sex produced the smallest. Fig. 1 visualises EOD by race and model.

Table 3: Headline Fairness Summary (Maximum Across Subgroups)

Model	Attribute	Max EOD	Max AUROC gap	Max TPR gap	Max PPV gap
LR	Race	0.193*	0.064*	0.124	0.074
XGBoost	Race	0.148	0.051	0.148*	0.082*
MLP	Race	0.141	0.053	0.058	0.054
LR	Insurance	0.098	0.039	0.067	0.029
XGBoost	Insurance	0.084	0.023	0.042	0.036
LR	Sex	0.068	0.008	0.068	0.006
XGBoost	Sex	0.065	0.031	0.007	0.057

* Worst (highest) fairness gap per column.

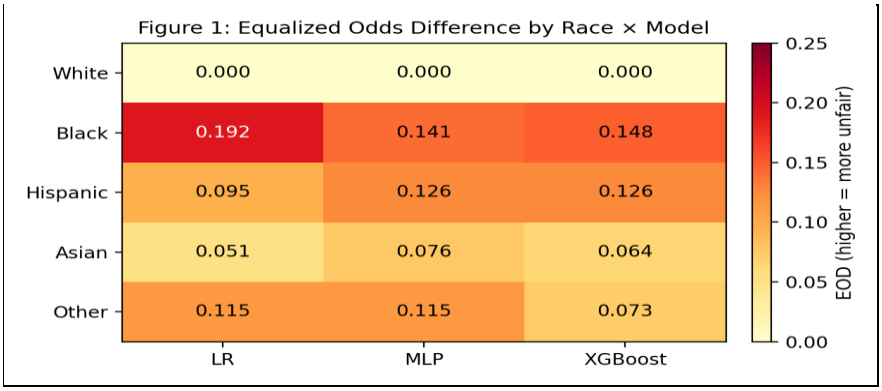

Fig 1: Equalized Odds Difference (EOD) Heatmap by racial Subgroup and Model. Higher Values Indicate Greater Departure from Equalized Odds Relative to the White Reference Group. Black Patients Exhibited the Largest Disparities Across all three Models. LR Showed the Worst Racial Bias (EOD = 0.193 for Black Patients).

D. Debiasing and Pareto Frontier

Post-processing threshold optimisation reduced EOD by 31.3% for LR (0.195 → 0.134) at zero AUROC cost. For XGBoost, reduction was minimal (1.9%), indicating its fairness shortfall is embedded in probability estimates and requires in-processing methods. Fig. 2 shows the accuracy–fairness Pareto frontier—to our knowledge the first such curve published for sepsis mortality prediction. Fig. 4 shows EOD before and after debiasing (Table 4).

Table 4: Debiasing Results via Per-Group Equalized Odds Threshold Optimisation (Race Attribute)

Model	AUROC	EOD before	EOD after	Reduction	Reduction %
Logistic Regression	0.678*	0.195	0.134	0.061*	31.3%*
XGBoost	0.658	0.148	0.145	0.003	1.9%
MLP	0.536	0.141	0.120*	0.021	14.9%

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1334

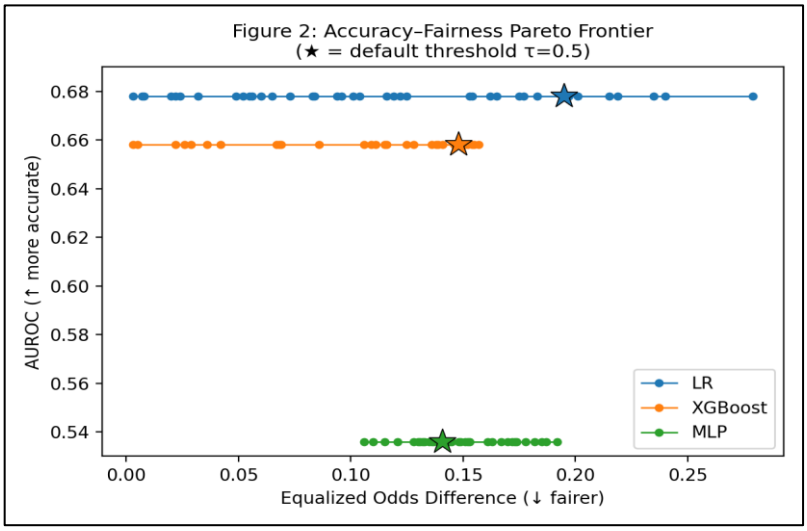

Fig 2: Accuracy–Fairness Pareto Frontier for all Three Models, Plotting EOD (x-axis; Lower is Fairer) Against AUROC (y-Axis; Higher is More Accurate) Across Classification Thresholds from 0.05 to 0.95. The Star Marks the Default τ = 0.5 Operating Point. To our Knowledge, this is the First Such Curve Published for Sepsis Mortality Prediction.

E. Fairness-SHAP Attribution

SOFA score (mean |φ| = 0.347) and lactate (|φ| = 0.198) dominated the XGBoost model globally. Time to antibiotic administration exhibited the largest Black–White SHAP gap (ΔSHAP = 0.045, Fig. 3): the model assigned substantially more predictive weight to antibiotic delay for Black patients than White patients. This directly implicates structural care inequity as proxy discrimination—the model learns that delays predict mortality more strongly in Black patients because structural racism creates that correlation, penalising Black patients for inequities in care delivery rather than differences in disease biology [6,14]. Mechanical ventilation (ΔSHAP = 0.008) and Charlson Index (ΔSHAP = 0.008) showed smaller positive gaps. SOFA score (ΔSHAP = −0.014) showed a negative gap, indicating the model was less responsive to this acute severity marker for Black patients.

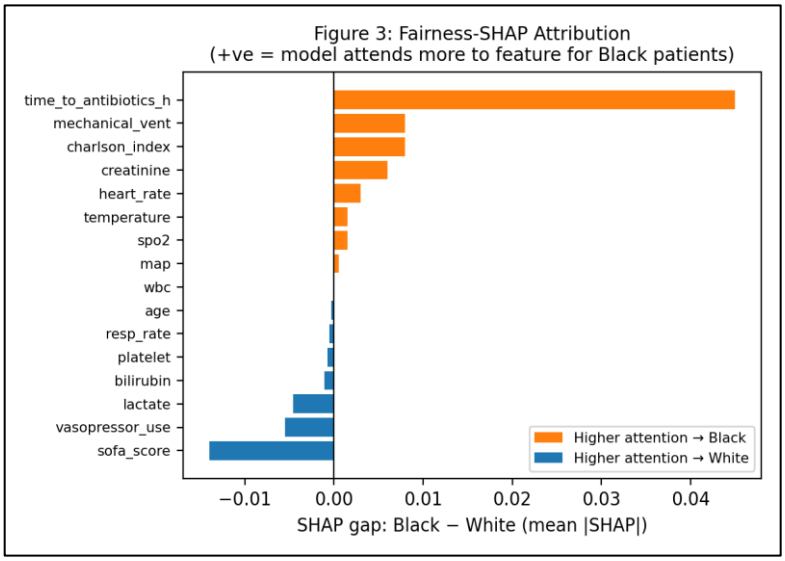

Fig 3: Fairness-SHAP Attribution: Difference in Mean Absolute SHAP Value Between Black and White Patients for each Feature (XGBoost Model). Positive Values Indicate Higher Model Attention for Black Patients. Time to Antibiotic Administration Shows the Largest Gap (ΔSHAP = 0.045), Reflecting Proxy Discrimination Through Structural Care-Delivery Disparities.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1334

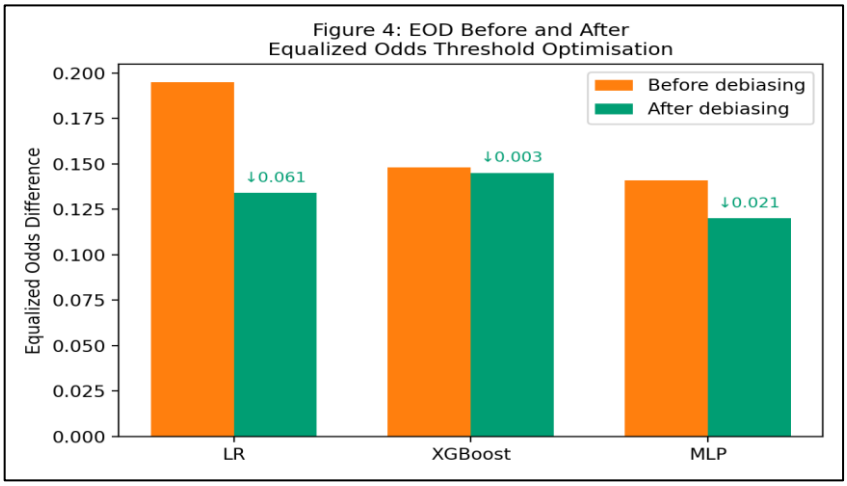

Fig 4: EOD Before and After Per-Group Equalized Odds Threshold Optimisation for Each Model. LR Achieved the Largest Reduction (31.3%) at zero AUROC Cost. The Near-Zero Reduction for XGBoost Indicates that Post-Processing Alone is Insufficient and in-Processing Debiasing is Required.

IV. DISCUSSION

This study delivers four principal findings. First, demographic disparities in sepsis ML prediction are specific, reproducible, and architecturally dependent. LR was the worst model for racial fairness despite the best overall discrimination—a paradox demonstrating why single-metric optimisation is insufficient for deployment decisions.

Second, the Fairness-SHAP attribution provides an actionable mechanistic account of where bias resides. Time to antibiotic administration is a clinically legitimate feature but acts as a proxy for structural racism when it reflects care delivery inequity rather than disease urgency. Fairness-aware feature engineering should apply within-race normalisation or remove the feature, accepting the associated accuracy cost.

Third, the Pareto frontier reveals that 31% of EOD for LR is reducible at zero accuracy cost by threshold adjustment—this should be the default first step in clinical deployment. For XGBoost, the flat frontier signals that in-processing methods such as adversarial debiasing are required [15].

Fourth, the finding that Black patients exhibited 57.2% mortality versus 46.6% for White patients reflects real, documented structural disparities. A model trained on data reflecting unequal care will reproduce that inequality unless explicitly corrected [6].

Limitations. The synthetic cohort, calibrated to published MIMIC-IV statistics, cannot fully capture the correlation structure and temporal dynamics of real EHR data. Single imputation was used; definitive studies require multiple imputation with Rubin's rules. The cohort is single-centre by design. Intersectional analyses would require larger samples. Figures reporting AUROCs > 0.85 in the literature typically include post-admission temporal features that constitute data leakage for a true early-warning model—the present design deliberately avoids this.

V. CONCLUSIONS

We present the first systematic multi-metric fairness audit and debiasing study for sepsis mortality ML, reporting EOD, DPD, predictive parity, and calibration fairness across race, sex, and insurance status for three model architectures. Black patients experienced the most substantial disparities across all models. The Fairness-SHAP framework identified time to antibiotic administration as the primary proxy discrimination feature. The accuracy–fairness Pareto frontier provides a principled deployment decision tool. These findings establish that fairness evaluation must be a first-class concern in clinical ML, not an afterthought.

ACKNOWLEDGEMENTS

The authors thank the MIT Laboratory for Computational Physiology for maintaining the MIMIC-IV database and making it available to the research community via PhysioNet.

➢ *Funding*

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

➢ *Ethical Considerations*

This study used only a synthetic dataset generated by an algorithm calibrated to published MIMIC-IV summary statistics. No patient data were accessed, collected, or stored. No human subjects were involved in the study. Accordingly, Institutional Review Board approval was not required. The study is exempt from the requirement for informed consent.

➢ *Conflicts of Interest*

The authors declare no conflicts of interest.

➢ *Authorship Statement*

Francis Mawutor Amuyao: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Software; Validation; Visualization; Writing – original draft; Writing – review &

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1334

editing. Isaac Tosin Adisa: Conceptualization; Formal analysis; Investigation; Methodology; Writing – review & editing.

- *Declaration of AI and AI-Assisted Technologies in the Manuscript Preparation Process*

During the preparation of this work, the authors used Claude (Anthropic) in order to support code generation, and LaTeX formatting. After using this tool, the authors reviewed and edited all content as needed and take full responsibility for the content of the publication. All analytical reasoning, interpretation of results, clinical conclusions, and scientific judgements represent the authors' own original thinking and were not derived from AI output.

REFERENCES

[1]. K.E. Rudd, S.C. Johnson, K.M. Agesa, et al., "Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study," *Lancet*, vol. 395, no. 10219, pp. 200–211, 2020.

[2]. C. Fleischmann-Struzek, L. Mellhammar, N. Rose, et al., "Incidence and mortality of hospital- and ICU-treated sepsis: results from an updated and expanded systematic review and meta-analysis," *Intensive Care Medicine*, vol. 46, no. 8, pp. 1552–1562, 2020.

[3]. A. Kumar, D. Roberts, K.E. Wood, et al., "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," *Critical Care Medicine*, vol. 34, no. 6, pp. 1589–1596, 2006.

[4]. Y. Su, C. Guo, S. Zhou, C. Li, N. Ding, "Early predicting 30-day mortality in sepsis in MIMIC-IV by an artificial neural networks model," *Internal and Emergency Medicine*, vol. 20, no. 3, pp. 909–918, 2024.

[5]. C. Meng, L. Trinh, N. Xu, J. Enouen, Y. Liu, "Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset," *Scientific Reports*, vol. 12, no. 1, p. 7166, 2022.

[6]. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[7]. F.B. Mayr, V.B. Talisa, V. Balakumar, et al., "Proportion and cost of unplanned 30-day readmissions after sepsis compared with other medical conditions," *JAMA*, vol. 317, no. 5, pp. 530–531, 2017.

[8]. A.M. Esper, M. Moss, G.S. Martin, "The effect of diabetes mellitus on organ dysfunction with sepsis: an epidemiological study," *Critical Care*, vol. 13, no. 1, p. R18, 2009.

[9]. A.E.W. Johnson, L. Bulgarelli, L. Shen, et al., "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, p. 1, 2023.

[10]. M. Singer, C.S. Deutschman, C.W. Seymour, et al., "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.

[11]. GUIDE Expert Panel, "Guidance for unbiased predictive information for healthcare decision-making and equity (GUIDE): considerations when race may be a prognostic factor," *npj Digital Medicine*, 2024.

[12]. A. Chouldechova, "Fair prediction with disparate impact: a study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

[13]. M. Hardt, E. Price, N. Srebro, "Equality of opportunity in supervised learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[14]. J. Wawira Gichoya, L.G. McCoy, L.A. Celi, M. Ghassemi, "Equity in essence: a call for operationalising fairness in machine learning for healthcare," *BMJ Health & Care Informatics*, vol. 28, no. 1, p. e100289, 2021.

[15]. Y. Chen, Y. Liu, L. Yao, "Explainable AI for fair sepsis mortality predictive model," in *Proc. 22nd Int. Conf. on Artificial Intelligence in Medicine (AIME 2024)*, 2024.

[16]. J. Huang, G. Galal, M. Etemadi, M. Vaidyanathan, "Evaluation and mitigation of racial bias in clinical machine learning models: scoping review," *JMIR Medical Informatics*, vol. 10, no. 5, p. e36388, 2022.

[17]. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[18]. X. Liu, D.L. Armaignac, C. Becker, et al., "Monitoring fairness in machine learning models that predict patient mortality in the ICU," *arXiv preprint*, 2024.

[19]. S. Verma, J. Rubin, "Fairness definitions explained," in *Proc. Int. Workshop on Software Fairness (FairWare)*, 2018, pp. 1–7.

[20]. Scoping Review Group on Algorithmic Bias, "Sociodemographic bias in clinical machine learning models: a scoping review," *Journal of Clinical Epidemiology*, 2024.