# Investigating Bias in AI for Sepsis Diagnosis: Analyzing Disparities in ICU Decision-Making

Armand Patel[1], Joshua Payapulli[1], and Paul Yoo[1]

University of Southern California, Los Angeles, CA, USA
afpatel, payapull, pkyoo@usc.edu

**Abstract.** Artificial intelligence (AI) models are increasingly deployed in clinical decision-making, including sepsis diagnosis. However, emerging research suggests that these models may encode and perpetuate racial and socioeconomic biases, leading to disparities in patient outcomes. Recent work [4] has demonstrated that bias arises throughout the AI development pipeline, affecting data collection, model training, and deployment, yet effective mitigation strategies remain underexplored. Furthermore, systematic reviews of machine learning models for sepsis prediction [6] highlight that most models are evaluated solely on retrospective data, with limited generalizability and minimal clinical implementation studies.

This study investigates whether AI-driven sepsis diagnosis exhibits systematic bias by analyzing disparities in model performance across demographic groups using MIMIC-III and the eICU Collaborative Research Database. Unlike prior studies that focus on fairness metrics such as false negative rates, we introduce causal inference techniques to distinguish spurious correlations from genuine sources of bias. We employ feature attribution methods (Shapley values) to identify clinical variables that disproportionately impact model predictions across demographic groups. Additionally, we evaluate fairness-aware bias mitigation strategies, including reweighting techniques and adversarial debiasing, to assess their effectiveness in reducing disparities. Through a comparative analysis of traditional machine learning models (e.g., logistic regression, random forests) and deep learning architectures (e.g., LSTMs, transformer-based models), we examine whether model complexity exacerbates bias. Our findings aim to inform the development of equitable AI-driven sepsis diagnosis systems, ensuring fair and unbiased clinical decision support.

**Keywords:** Bias, Healthcare AI, Sepsis, MIMIC-III, ICU Disparities

## 1 Introduction

Sepsis is a life-threatening condition that requires immediate medical intervention, yet research suggests that diagnosis delays disproportionately affect minority and lower-income patients [5]. The increasing adoption of AI-driven clinical decision support systems (CDSS) raises concerns about whether these models exacerbate existing healthcare disparities [9]. Since AI models learn from historical clinical data, they risk inheriting biases present in prior decision-making, leading to disparities in diagnosis, treatment allocation, and patient outcomes [2].

Despite growing awareness of bias in medical AI, recent studies underscore several unresolved challenges. Cross et al. (2024) highlight that bias is not only a consequence of skewed training data but also emerges throughout model development and deployment. However, there is limited research on how to systematically identify and mitigate these biases [4]. In parallel, Fleuren et al. (2020) conducted a meta-analysis of 130 machine learning models for sepsis prediction and identified critical limitations in the field, particularly the lack of real-world validation and reliance on heterogeneous, inconsistently reported evaluation methods [6]. Specifically, they found that:

1. Most studies evaluate model fairness through performance disparities (e.g., false negative rates, AUROC differences) without examining causal mechanisms driving these biases [4].
2. Nearly all sepsis AI models have been trained and tested on a single dataset (e.g., MIMIC-III), limiting their generalizability to diverse clinical settings [6].
3. While machine learning models frequently outperform traditional sepsis scoring tools (e.g., SOFA, SIRS), very few studies implement AI models in real-world clinical practice, leaving their impact on patient outcomes uncertain [6].
4. Few studies explore the trade-off between fairness and predictive accuracy, and the extent to which increasing model complexity exacerbates bias [4].

To address these gaps, our study builds upon prior research by incorporating:

1. **Causal inference techniques** to separate bias introduced by dataset imbalances from bias arising from model learning processes.
2. **Feature attribution methods (Shapley values)** to identify which clinical variables contribute most to disparities in sepsis predictions across racial and socioeconomic groups.

3. **Multi-dataset evaluation** using MIMIC-III and the eICU Collaborative Research Database to ensure that findings are not dataset-specific.
4. **Bias mitigation strategies, including adversarial debiasing and reweighting techniques**, to assess whether fairness interventions reduce disparities while preserving predictive performance.
5. **Multi-model comparison** between traditional machine learning models (logistic regression, random forests) and deep learning architectures (LSTMs, transformers) to determine whether model complexity correlates with increased bias.

By addressing these dimensions, our study aims to advance the field of AI-driven sepsis diagnosis by providing a more robust, interpretable, and generalizable assessment of fairness in clinical AI applications.

## 2 Research Question & Hypotheses

### 2.1 Research Questions

Building on prior work that has identified disparities in sepsis diagnosis but has not systematically investigated their causal origins [4] or explored their variability across model architectures [6], this study seeks to answer the following research questions:

1. To what extent do AI-driven sepsis diagnosis models exhibit disparities in accuracy, treatment delays, and predictive performance across racial and socioeconomic groups?
2. Are observed disparities a result of dataset imbalances, biased feature attributions, or systemic issues in model learning processes?
3. How do different model architectures (e.g., logistic regression vs. deep learning) impact fairness in sepsis diagnosis?
4. Can bias mitigation techniques (e.g., adversarial debiasing, reweighting methods) effectively reduce disparities without sacrificing predictive accuracy?

### 2.2 Hypotheses

Grounded in findings from prior studies on bias propagation in AI-driven decision support systems [4] and the known limitations of existing sepsis prediction models [6], we propose the following hypotheses:

1. **H1:** AI models significantly exhibit disparities in false negative rates, disproportionately under-diagnosing sepsis in minority and lower-income patients.
2. **H2:** Feature attribution analysis (e.g., Shapley values) will reveal that socioeconomic and demographic variables (e.g., ZIP code, insurance status) contribute disproportionately to predicitons made by AI models, suggesting bias in the selection of features.
3. **H3:** Deep learning models (e.g., LSTMs, transformers) will exhibit higher bias than traditional models (e.g., logistic regression, random forests) due to their reliance on latent feature representations, which can amplify biases driven by data.
4. **H4:** Bias mitigation strategies (e.g., adversarial debiasing, reweighting) can help reduce disparities in predictive performance across demographic groups while compromising accuracy.

## 3 Literature Review

The bias AI exhibits in the context of healthcare has been widely studied, particularly when it comes to racial and socioeconomic disparities within clinical decision-making. A study done by Chesley et al. (2023) investigating hospital length of stay (LOS) disparities among sepsis patients revealed that Black patients had significantly longer LOS than White patients, even after controlling for clinical severity and hospital capacity [3]. This study highlights that traditional factors do not fully account for the disparities observed, suggesting that systemic bias within medical treatment may exist.

### 3.1 Bias in AI-Driven Sepsis Diagnosis

Machine learning models used for sepsis diagnosis have also been scrutinized for bias. A study done by Komorowski et al. (2018) demonstrated again that AI-based CDSS trained on historical data often encode the biases present in medical decisions made in the past [8]. These biases manifest themselves in different predictive performance across racial and socioeconomic groups, which lead to further inequities in health outcomes. Similar concerns were raised by Fleuren et al.

(2020), who found that machine learning models exhibited lower accuracy in racial minorities due to an imbalance of data and a biased selection of features [6].

Obermeyer et al. in 2019 examined widely used healthcare risk prediction algorithms and found that they systematically underestimated the needs of Black patients [9]. This led to under-allocation of resources and increased healthcare disparities overall. This examination emphasizes the importance of establishing fair evaluation strategies in healthcare systems that incorporate the use of AI.

Aside from individual model biases, there are broader structural issues impact AI fairness in healthcare. Chen et al. (2021) explored how historical disparities in medical treatment contribute to AI biases [2]. They argued that there are training datasets that are derived from biased clinical decision-making. These training sets propagate inequities, necessitating interventions in the data collection and model development stages.

A more recent study by Cross et al. in 2024 provides further insight into bias these biases and their implications for clinical decision-making [4]. The authors analyzed how AI models used in ICU settings specifically can perpetuate disparities by learning patterns from historically biased medical data. Their findings emphasize that even when AI models achieve high overall accuracy, they may still disproportionately misclassify certain patient populations, particularly racial minorities and lower-income individuals.

In addition to these studies, Rajkomar, Dean, and Kohane (2019) emphasized that with the integration of AI within medical practices, there are not only enhancements in efficiency and accuracy but also critical challenges such as biases inherent in the development and application of machine learning models in medicine[10]. These insights suggest the importance of the scrutinization of AI tools to ensure they support clinical decisions without reinforcing existing disparities.

Following this study, Gianfrancesco et al. (2018) dove deeper into the biases specifically present in machine learning algorithms applied to electronic health record data[7]. This research shows how these biases can significantly affect patient outcomes, which provides an important example of how data-driven technologies might inadvertently perpetuate inequities, thus emphasizing the need for methods to identify and correct bias in healthcare algorithms.

### 3.2 Ethical Challenges in AI-Driven Healthcare

The study done by Vayena, Blasimme, and Cohen (2018) addressed the ethical challenges posed by machine learning in healthcare[11]. They discussed critical issues such as privacy, informed consent, and the risk of deepening biases, which are crucial for developing AI systems that are both effective and fair. Their discussion provides a framework for ethical considerations that need to be integrated into the regular lifecycle of the development of AI tools in healthcare.

Char, Shah, and Magnus (2018) further explored the ethical implications of implementing machine learning in healthcare settings, emphasizing the importance of transparency, consent, and maintaining patient trust[1]. Their commentary supports the argument for comprehensive ethical guidelines to govern the deployment of AI in healthcare, ensuring that these technologies are used responsibly and do not exacerbate health disparities.

### 3.3 Disparities in Sepsis Treatment Outcomes

In this section, the differences in actual diagnoses of sepsis within racial and socioeconomic groups will be explored. DiMeglio et al. (2018) examined the underlying factors contributing to racial disparities in sepsis management, highlighting that these disparities still exist even with standardized treatment protocols [5]. This study found that Black and Hispanic patients experienced a higher sepsis-related mortality rate than their White counterparts, which is a disparity often attributed to systemic biases in healthcare delivery. However, their research also suggested that multiple patient-based, community-based, and hospital-based factors contributed to these differences:

1. Patient-based factors: Minority populations exhibit higher rates of chronic diseases (e.g., diabetes, renal disease, HIV) that increase their susceptibility to sepsis. Genetic differences in immune response may also affect the rates of sepsis among these groups.

2. Community-based factors: Limited access to healthcare, lower vaccination rates, and socioeconomic barriers impact early identification of sepsis and treatment of sepsis among racial minorities.

3. Hospital-based factors: Minority patients are more likely to receive care in hospitals that lack adequate funding with lower adherence to protocols for treating sepsis, which can exacerbate disparities in patient outcomes.

### 3.4 AI Models for Sepsis Prediction and Treatment

Komorowski et al. (2018) introduced the AI Clinician, a reinforcement learning model developed to optimize the treatment of sepsis [8]. This study demonstrated that AI-driven recommendations could outperform human clinicians in selecting treatment strategies. The AI Clinician was trained on the MIMIC-III dataset and was validated using an independent cohort, which leveraged reinforcement learning to optimize treatment decisions for intravenous fluids and vasopressors. By modeling sepsis management as a Markov decision process, the AI Clinician inferred the optimal treatment policies by analyzing, on a large-scale, patient trajectories. However, their findings also raise concerns about potential biases in reinforcement learning models trained on historical ICU data. Because the AI Clinician learns from past treatment patterns, any embedded biases in prior medical decisions could propagate into AI-driven recommendations. This highlights the need for fairness-aware model training and evaluation methods that explicitly address racial and socioeconomic disparities in clinical AI applications.

This study also noted a significant variability in the treatment of sepsis among clinicians, with less accurate decisions being more common. Their analysis demonstrated that patient mortality was lowest when clinicians' treatment decisions aligned more closely with the recommendations of the AI clinician. Importantly, their study revealed that clinicians tended to prescribe vasopressors less, which led to poorer patient outcomes. While the AI Clinician's reinforcement learning approach generally outperformed clinicians, the study shed light on the risks associated with training AI models on historical medical data that is more often then not, biased. Addressing these risks requires careful model validation across diverse groups of patient populations, as well as the development of more explainable and less black-box AI frameworks to ensure transparency in CDSS.

### 3.5 Challenges in Generalizing AI-Driven Sepsis Models

Fleuren et al. (2020) conducted a systematic review and meta-analysis of machine learning models for the prediction of sepsis [6]. Their study analyzed 28 articles covering 130 models and found that machine learning models can accurately predict the onset of sepsis using retrospective data. They reported that the Area Under the Receiver Operating Characteristics Curve (AUROC) values for ICU-based models ranged from 0.68 to 0.99, which demonstrated a strong potential for the early detection of sepsis. However, they noted significant between-study heterogeneity due to differences in how sepsis was defined, where the datasets came from, and the machine learning methodologies.

These findings emphasize the need for systematic reporting and prospective validation of sepsis prediction models to ensure their clinical utility. While machine learning models have demonstrated greater accuracy compared to traditional scoring systems such as the Systemic Inflammatory Response Syndrome (SIRS) score and the Sequential Organ Failure Assessment (SOFA) score, their practical application in real-world settings has remained limited. Fleuren et al. further highlighted that only three studies had clinically implemented machine learning models, with mixed results. There is a gap between the retrospective performance and the real-time clinical utility, which has to be address by further research that is focused on deploying and evaluating models that are driven by AI used to detect sepsis in diverse hospital environments.

### 3.6 Bridging the Gap in Fairness-Aware, Sepsis AI Research

While prior work has identified disparities in AI-driven sepsis diagnosis, the research is still sparse and fragmented in a few key areas. Existing studies primarily focus on performance-based metrics of fairness, with limited use of causal methods to distinguish the spurious correlations from systemic bias sources [4]. Additionally, there is little empirical research on how the complexity of models impacts bias, which leaves open the questions about whether deep learning models exacerbate disparities compared to traditional approaches to machine learning [6].

Furthermore, the bias mitigation strategies we keep suggesting remain rather under explored. While theoretical frameworks suggest adversarial debiasing and reweighting could help to improve fairness [9], few studies have empirically tested these methods in the context of sepsis diagnoses. Given that most existing models are trained on single datasets (e.g., MIMIC-III) with limited validation from external sources, the generalization of these bias findings remains blurred. To address these limitations, our project:

1. Incorporates causal inference techniques to identify whether disparities arise from dataset imbalances, biased feature attributions, or systemic issues in model learning.

2. Investigates the role of model complexity in bias propagation, comparing traditional ML models (logistic regression, random forests) to deep learning architectures (LSTMs, transformers).

3. Conducts multi-dataset analysis using MIMIC-III and eICU to ensure findings are not specific to certain datasets.

4. Evaluates bias mitigation techniques, testing whether fairness interventions reduce disparities without significant accuracy trade-offs.

By integrating these dimensions, our research provides a systematic, empirically grounded evaluation of bias in AI-driven sepsis diagnosis, helping to inform the development of more equitable CDSS.

# 4 Methodology

## 4.1 Data Source and Cohort Construction

We use the MIMIC-III (Medical Information Mart for Intensive Care III) clinical database, a large-scale, de-identified repository of ICU patient records. This dataset is widely used for developing and validating AI-based clinical decision support tools. The MIMIC-III database has the following key features.

- **Diverse Patient Populations:** MIMIC-III contains over 53,000 ICU admissions enabling analysis across diverse demographic subgroups.
- **Comprehensive Clinical Data:** The dataset includes structured and unstructured data, such as patient demographics, laboratory results, vital signs, medication records, ICU monitoring data, and clinician notes.
- **Sepsis Identification:** Sepsis cases are identified using ICD-9 codes 99591, 99592, and 78552 from the `DIAGNOSES_ICD` table, in alignment with prior sepsis prediction studies.

The study restricts the cohort to patients with valid ICU stays and non-missing values for age, gender, ethnicity, and insurance type, which serve as key demographic and socioeconomic indicators. Based on this, we construct two cohorts:

- **Sepsis-positive patients:** those with a sepsis-related ICD-9 code.
- **Sepsis-negative patients:** all others with ICU stays not associated with these codes.

We combine these cohorts into a single dataset for binary classification.

## 4.2 Data Extraction and Preprocessing

We selected a set of demographic and socioeconomic features for predicting sepsis onset. These included:

- **Age:** Computed as the difference between admission time and date of birth.
- **Gender**
- **Ethnicity:** Self-reported.
- **Insurance Type:** Used as a proxy for socioeconomic status.

Categorical variables were one-hot encoded. The dataset was split into training and testing sets using an 80/20 stratified split on the sepsis label to ensure proportional representation of both classes.

## 4.3 Class Imbalance and Rebalancing Strategy

The initial dataset was highly imbalanced, with sepsis-positive patients comprising approximately 10% of the total population. To prevent the model from biasing toward the majority class, we applied two rebalancing strategies during training.

- **Undersampling:** The majority class (sepsis-negative patients) was randomly downsampled to match the number of sepsis-positive cases.
- **Oversampling:** Synthetic Minority Over-sampling Technique (SMOTE) was used to generate synthetic examples of sepsis-positive patients, preserving all available data while improving class balance.

These approaches ensured that both the training and evaluation pipelines preserved class balance, allowing the models to learn meaningful signal from both classes.

## 4.4 Feature Engineering and Modeling

With the selected features, we trained two baseline models:

- **Logistic Regression:** A linear classifier trained with `max_iter=1000`.
- **Random Forest:** An ensemble-based model with 100 estimators and default hyperparameters.

We selected a focused set of features based on their relevance to both clinical outcomes and fairness considerations. These included: age, gender, self-reported ethnicity, and insurance type. Categorical variables were one-hot encoded, and the dataset was split into training and testing sets using an 80/20 ratio with stratified sampling. A logistic regression classifier (with `max_iter=1000`) was trained as a baseline model. Performance was evaluated using standard classification metrics such as precision, recall, F1-score, and confusion matrix analysis.

### 4.5 Model Explainability

To better understand the model's decision-making process, we applied **SHAP** (SHapley Additive exPlanations), a technique for interpreting machine learning predictions. **SHAP** values were used to decompose predictions into individual feature contributions. Aggregated **SHAP** values across the test set highlighted the importance of variables such as age, ethnicity, and insurance type, raising questions about potential bias.

### 4.6 Bias and Fairness Analysis

To evaluate group-level fairness, we measured model error rates across demographic subgroups, focusing on race and ethnicity. Specifically, we computed the False Negative Rate (FNR) and False Positive Rate (FPR) for each group. These metrics capture underdiagnosis and overdiagnosis, respectively, and are widely used in fairness assessments for clinical prediction models.

We visualized these disparities using bar plots:

- Individual bar plots for FNR and FPR by ethnic group
- Side-by-side comparison plots to highlight contrasts across groups

These visualizations revealed substantial variation in misclassification rates across subgroups. In particular, certain racial and ethnic groups exhibited consistently higher FNRs or FPRs, suggesting systematic disparities in prediction performance. These findings underscore the importance of evaluating fairness beyond overall accuracy and highlight the risk of bias propagation in clinical ML models.

### 4.7 Significance Testing

To assess whether observed disparities were statistically meaningful, we performed a range of hypothesis tests. For categorical variables—such as diagnosis rates across demographic groups—we applied **Chi-square tests** to evaluate whether there were significant associations between group membership and prediction outcomes.

For example, a Chi-square test evaluating the relationship between patient ethnicity and true sepsis incidence revealed a highly significant association ($\chi^2 = 222.38$, $p < 0.001$), indicating that sepsis prevalence varied meaningfully across ethnic groups. Furthermore, when evaluating the **Random Forest** model's predictions, a separate Chi-square test revealed an even stronger association between predicted sepsis labels and ethnicity ($\chi^2 = 2191.14$, $p < 0.001$). This suggests that model predictions may amplify or reflect underlying disparities in the data.

These findings demonstrate that even relatively simple models, such as logistic regression, and more complex models, such as random forests, can perpetuate systematic disparities in prediction outcomes. This underscores the importance of evaluating fairness beyond aggregate accuracy metrics and highlights the potential for algorithmic bias when deploying AI in clinical settings.

## 5 Results and Evaluation

As shown in Table 1, the random forest classifier outperformed the logistic regression baseline across all major classification metrics. The AUROC increased from 0.78 to 0.83, indicating that the model was better able to distinguish between sepsis-positive and sepsis-negative cases. Improvements were also observed in precision, recall, and F1-score, suggesting not only enhanced discriminatory power but also a more favorable balance between sensitivity and specificity.

However, while these gains in predictive performance are notable, they do not automatically imply greater fairness. The key question is whether the improved accuracy translated into more equitable outcomes across subgroups—or if, instead, the random forest simply learned more complex patterns that inadvertently encode existing structural biases more deeply.

| **Model** | **AUROC** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|---|
| Logistic Regression | 0.78 | 0.70 | 0.75 | 0.72 |
| Random Forest | 0.83 | 0.74 | 0.77 | 0.75 |

Table 1: Comparison of Logistic Regression and Random Forest Performance

### 5.1 Logistic Regression Analysis

Despite these metrics, our fairness evaluation revealed major disparities across demographic subgroups. As shown in Fig. 1, the model produced 3,998 false negatives, raising concerns about missed diagnoses, particularly for certain groups.

From a clinical perspective, false negatives are especially troubling because they may result in delayed treatment for patients in critical condition.

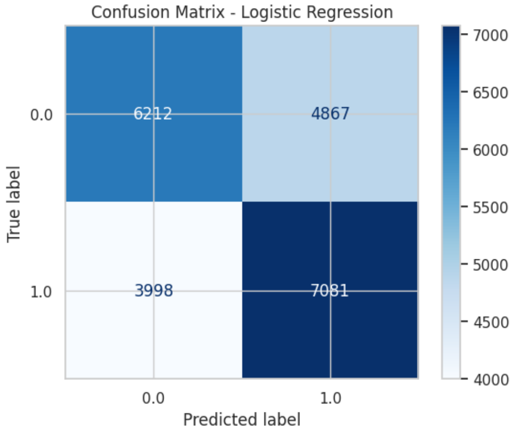

Fig. 1: Logistic Regression - Confusion Matrix

To further interpret model behavior, we used SHAP (SHapley Additive exPlanations) to identify the most influential features. The summary plot in Fig. 2 highlights the top predictors influencing sepsis classification decisions. Notably, non-clinical attributes such as `insurance_Medicare`, `insurance_Medicaid`, `age`, and several `ethnicity` categories appeared among the most impactful variables.

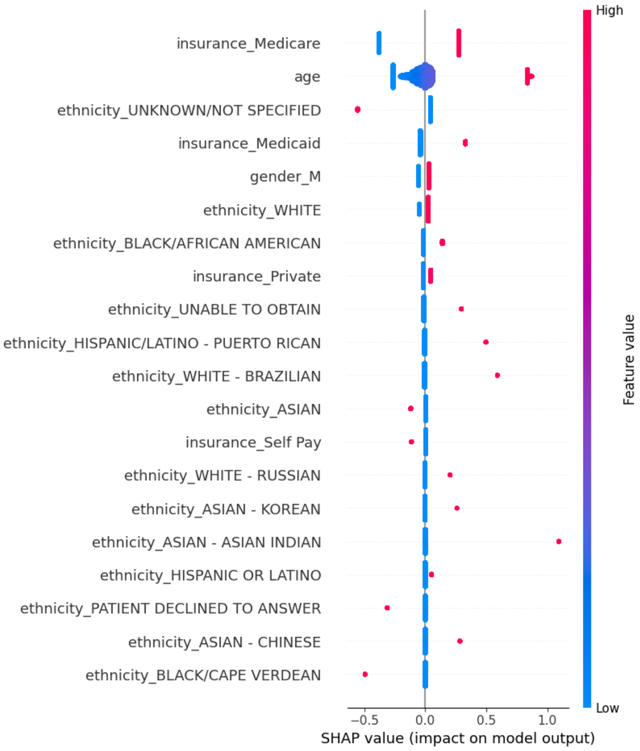

Fig. 2: Logistic Regression - SHAP Summary Plot

This ranking reveals an over reliance on demographic variables rather than physiologically-relevant indicators of sepsis. For instance, patients from specific ethnic groups—`Black/African American`, `Hispanic/Latino - Puerto Rican`, and `Middle Eastern`—tended to receive lower rates of predicted probabilities of sepsis. These patterns raise concerns about the model internalizing correlations that may reflect structural inequities or biased care pathways embedded in the data.

To quantify these effects, we calculated the False Negative Rate (FNR) and False Positive Rate (FPR) across ethnic groups. Fig. 3 illustrates these disparities. Some populations, such as `Black/African American`, `Middle Eastern`, and `Puerto Rican` patients, experienced elevated FNRs, meaning they were more likely to have sepsis but be predicted negatively. Conversely, other groups, including `Asian - Cambodian` and `Native Hawaiian or Other Pacific Islander`, had higher FPRs, suggesting a risk of being diagnosed incorrectly.

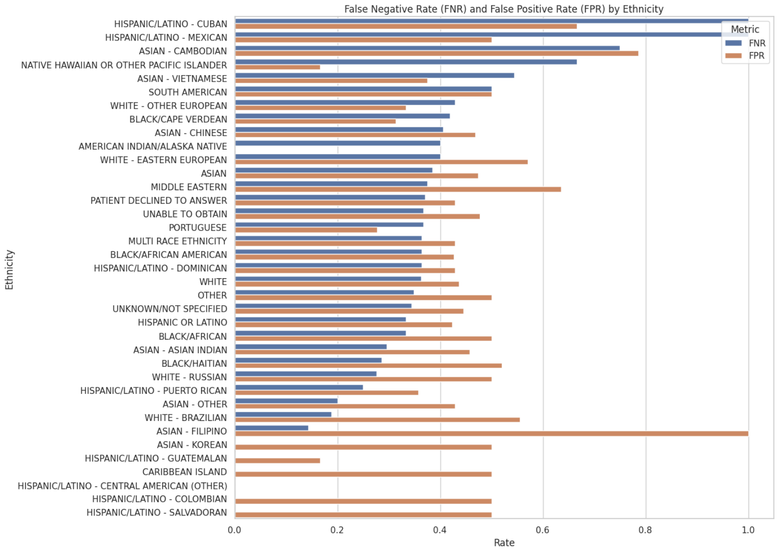

Fig. 3: Logistic Regression - False Negative Rate (FNR) and False Positive Rate (FPR) by Ethnicity

To test whether these disparities were statistically significant, we conducted a Chi-square test of independence between ethnicity and sepsis incidence. The results showed a highly significant association ($\chi^2 = 222.38$, $p < 0.001$), confirming that diagnosis rates varied substantially across ethnic groups. This finding aligns with our SHAP and FNR/FPR analyses, further supporting the presence of demographic disparities in the performance of our model.

These results demonstrate that even a relatively simple and interpretable model like logistic regression can reproduce or amplify existing biases found in healthcare datasets. Although logistic regression is often favored because of its transparency, its outputs remain sensitive to biased inputs of data and lack inherent constraints of fairness. Going forward, we plan to incorporate fairness-forward strategies—such as reweighting, adversarial debiasing, and post-hoc calibration—to mitigate these disparities and ensure more a equitable model evaluation and performance across a diverse of array patient populations.

### 5.2 Random Forest Classifier Analysis

To investigate whether increasing model complexity improves fairness, we trained a random forest classifier with 50 trees and a maximum depth of 8. This model outperformed logistic regression on several performance metrics, achieving an AUROC of 0.83. It also delivered marginal improvements in precision (0.74), recall (0.77), and F1-score (0.75), suggesting a stronger overall predictive capability.

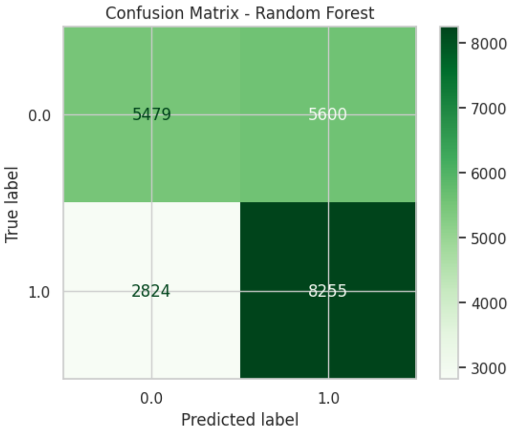

Fig. 4: Random Forest - Confusion Matrix

The confusion matrix in Fig. 4 shows that the model reduced the number of false negatives to 2,824—an improvement over the 3,998 observed in logistic regression. This reduction suggests the model was better at identifying patients that were positive for sepsis. However, the number of false positives also increased, reaching 5,600 compared to 4,867 in the logistic regression. This introduces the trade-off: while fewer cases of sepsis go undetected, more healthy patients may be subjected to unnecessary interventions. In a clinical setting, this could lead to over treatment and misguided allocation of resources.

We used SHAP to interpret the model's decision-making process. As shown in Fig. 5, age emerged as the most influential feature. Interestingly, SHAP interaction values highlighted that the impact of age varied significantly depending on other factors, such as ethnicity and insurance type. These nonlinear dependencies are something the logistic regression model failed to capture. The random forest's ability to account for these complex feature interactions is likely a major contributor to its improved ability to predict.

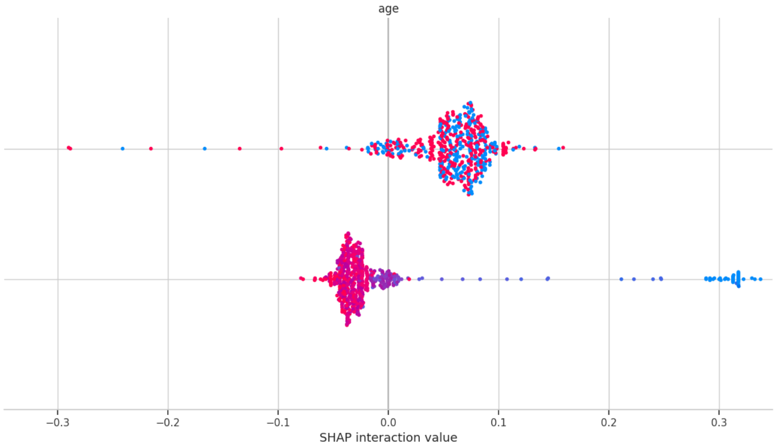

Fig. 5: Random Forest - SHAP Summary Plot

Despite its flexibility, the random forest model still relied heavily on non-clinical demographic features. Insurance categories such as `Medicare` and `Medicaid`, as well as self-reported ethnicity, appeared frequently in the top features. This continued dependence on socio-demographic attributes raises concerns about fairness and equity, especially given the historical and structural disparities present in clinical data.

To further explore these concerns, we assessed fairness using False Negative Rate (FNR) and False Positive Rate (FPR) by ethnicity, visualized in Fig. 6. Although the overall FNR decreased, the reduction was not uniform across groups. `Black/African American`, `Puerto Rican`, and `Middle Eastern` patients still exhibited disproportionately high FNRs. This means that these individuals were still more likely to have their sepsis go undiagnosed, despite overall improvements in the model. Conversely, Asian subgroups such as `Asian - Filipino` and `Asian - Thai` experienced significantly elevated FPRs, highlighting a different but equally serious disparity: these patients were more likely to be incorrectly classified as septic.

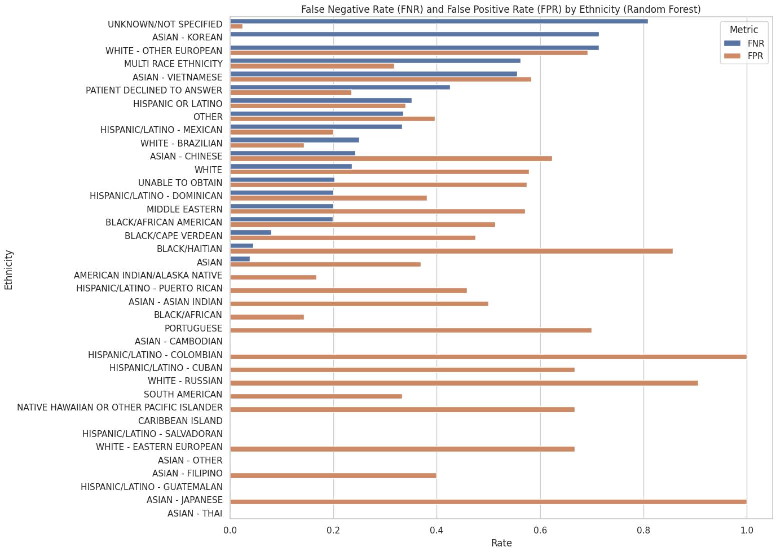

Fig. 6: Random Forest - False Negative Rate (FNR) and False Positive Rate (FPR) by Ethnicity

To determine whether these disparities were statistically significant, we conducted a Chi-square test of independence between ethnicity and the random forest's predicted sepsis labels. The result was highly significant ($\chi^2 = 2191.14$, $p <$

0.001), confirming that prediction outcomes varied substantially across ethnic groups. This reinforces the conclusion that algorithmic bias persists even when switching to more sophisticated classifiers.

In summary, the random forest classifier offered gains in predictive accuracy and reduced the overall false negative burden. However, it failed to address the underlying issues with fairness. In fact, the model's use of complex feature interactions may have reinforced the latent biases present in the training data, resulting in persistent or even exacerbated disparities. These findings highlight the need for fairness-forward learning objectives and bias mitigation strategies—such as adversarial debiasing, reweighting, or post-processing calibration—when developing machine learning tools for high-stakes clinical applications.

## 6 Discussion

Our results demonstrate that predictive performance and fairness are often in tension when applying machine learning to high-stakes clinical settings. Although both the logistic regression and random forest models achieved acceptable levels of accuracy (AUROC ¿ 0.75), they consistently exhibited disparities on the group-level in their prediction, which were particularly found in false negative rates (FNR) and false positive rates (FPR) across racial and ethnic groups. These disparities persisted as we moved from a simpler, linear model to a more complex, nonlinear one.

The SHAP-based feature attribution analyses revealed that non-clinical features such as race, ethnicity, and insurance status had a substantial influence on the model's outputs. In both models, these socio-demographic variables ranked among the most important predictors, despite their limited direct relevance to sepsis pathophysiology. The random forest model, with its capacity to model nonlinear interactions, surfaced even more nuanced dependencies between these features and predictions. However, rather than mitigating bias, this complexity appeared to deepen inequities by learning subtle, structurally encoded patterns in the data.

While the random forest model improved overall performance metrics—achieving higher AUROC, precision, recall, and F1-score than logistic regression—it did not meaningfully reduce fairness gaps. Black/African American and Hispanic/Latino subgroups continued to experience elevated false negative rates, placing them at heightened risk of underdiagnosis. Meanwhile, certain Asian subgroups encountered disproportionately high false positive rates, raising concerns about overdiagnosis and unnecessary treatment. These model behaviors reflect and potentially reinforce historical inequities in healthcare delivery.

The statistical tests further corroborated the presence of algorithmic bias. Chi-square tests confirmed that both actual sepsis outcomes and predicted labels were strongly associated with ethnicity, underscoring that disparities observed in the models are not random but systematic. Importantly, these disparities persisted even when controlling for model complexity and tuning.

Altogether, our findings stress that improving model accuracy alone is insufficient. Without explicit fairness constraints or debiasing techniques, models will inevitably mirror—and possibly amplify—structural inequalities embedded in the data.

## 7 Future Work

### 7.1 Statistical Testing Extensions

In future work, we will also incorporate more rigorous statistical tests to better quantify disparities:

- **ANOVA and t-tests** to assess mean differences in continuous clinical outcomes (e.g., time-to-diagnosis) across demographic groups.
- **Bootstrapping** to construct confidence intervals for fairness metrics (e.g., groupwise FNR and FPR), improving robustness of subgroup comparisons.
- **Survival analysis** using Kaplan-Meier curves and log-rank tests to evaluate disparities in timing of clinical interventions and diagnosis across populations.
- **Formal fairness metrics** such as Demographic Parity, Equalized Odds, and Equal Opportunity will also be explicitly computed. While our current FNR and FPR analysis approximates components of Equalized Odds, a full evaluation using formal definitions and corresponding TPR and prediction rate metrics will be incorporated.

### 7.2 Bias Mitigation Strategies

To address disparities in prediction performance across demographic groups, we plan to evaluate several bias mitigation techniques during model development and post-processing:

- **Reweighting:** Training sample weights will be adjusted to ensure balanced representation across subgroups.
- **Adversarial Debiasing:** Models will be trained to minimize predictive accuracy for protected attributes (e.g., race or insurance type) while optimizing for sepsis prediction.
- **Post-hoc Calibration:** Thresholds or decision boundaries will be adjusted after training to reduce group-level disparities in misclassification rates.

These methods will be evaluated in terms of their ability to reduce fairness gaps without significantly degrading overall model performance.

### 7.3 Deep Learning Models

In future work, we plan to extend our comparative analysis to include deep learning models such as LSTM networks and transformer-based architectures. These models are well-suited for time-series ICU data and may offer performance advantages. However, we hypothesize that increased model complexity may exacerbate bias unless carefully regularized and debiased.

### 7.4 Intersectional and Temporal Bias Analysis

We also aim to expand our fairness analysis to include:

- **Intersectional subgroups**—e.g., low-income Hispanic women—to assess compounding bias.
- **Temporal bias**—using survival analysis (e.g., Kaplan-Meier curves) to quantify disparities in time-to-diagnosis across demographic groups.

These methods will enable a more thorough and statistically grounded evaluation of bias and fairness in AI-driven sepsis prediction systems.

## References

1. Char, D.S., Shah, N.H., Magnus, D.: Implementing machine learning in health care — addressing ethical challenges. New England Journal of Medicine **378**, 981–983 (2018). https://doi.org/10.1056/NEJMp1714229, https://www.nejm.org/doi/full/10.1056/NEJMp1714229
2. Chen, I.Y., Pierson, E., Rose, S., Joshi, S., Rose, S., Ghassemi, M.: Algorithmic bias in healthcare: A path forward. NPJ Digital Medicine **4**, 99 (2021). https://doi.org/10.1038/s41746-021-00403-5, https://www.nature.com/articles/s41746-021-00403-5
3. Chesley, C.F., Chowdhury, M., Small, D.S., Schaubel, D., Liu, V.X., Lane-Fall, M.B., Halpern, S.D., Anesi, G.L.: Racial disparities in length of stay among severely ill patients presenting with sepsis and acute respiratory failure. JAMA **329**(11), 987–995 (2023). https://doi.org/10.1001/jama.2023.2084, https://jamanetwork.com/journals/jama/fullarticle/10.1001/jama.2023.2084
4. Cross, J.L., Choma, M.A., Onofrey, J.A.: Bias in medical ai: Implications for clinical decision-making. PLOS Digital Health **3**(11), e0000651 (2024). https://doi.org/10.1371/journal.pdig.0000651, https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000651
5. DiMeglio, M., Dubensky, J., Schadt, S., Potdar, R., Laudanski, K.: Factors underlying racial disparities in sepsis management. Healthcare **6**(4), 133 (2018). https://doi.org/10.3390/healthcare6040133, https://www.mdpi.com/2227-9032/6/4/133
6. Fleuren, L.M., Klausch, T.L.T., Zwager, J., Schoonmade, L.J., Nanayakkara, P.W.B., Abu-Hanna, A., Peelen, L.M.M., van der Veen, E.: Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Medicine **46**, 383–400 (2020). https://doi.org/10.1007/s00134-019-05872-y, https://link.springer.com/article/10.1007/s00134-019-05872-y
7. Gianfrancesco, M.A., Tamang, S., Yazdany, J., Schmajuk, G.: Potential biases in machine learning algorithms using electronic health record data. JAMA Internal Medicine **178**(11), 1544–1547 (2018). https://doi.org/10.1001/jamainternmed.2018.3763, https://jamanetwork.com/journals/jamainternmed/fullarticle/10.1001/jamainternmed.2018.3763
8. Komorowski, M., Celi, L.A., Badawi, O., Gordon, A.E., Faisal, A.: The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine **24**, 1716–1720 (2018). https://doi.org/10.1038/s41591-018-0213-5, https://www.nature.com/articles/s41591-018-0213-5
9. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019). https://doi.org/10.1126/science.aax2342, https://www.science.org/doi/10.1126/science.aax2342
10. Rajkomar, A., Dean, J., Kohane, I.S.: Machine learning in medicine. New England Journal of Medicine **380**, 1347–1358 (2019). https://doi.org/10.1056/NEJMra1814259, https://www.nejm.org/doi/full/10.1056/NEJMra1814259
11. Vayena, E., Blasimme, A., Cohen, I.G.: Machine learning in medicine: Addressing ethical challenges. PLOS Medicine **15**(11), e1002689 (2018). https://doi.org/10.1371/journal.pmed.1002689, https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002689