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 select-

ing 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 Dataset

This study utilizes the MIMIC-III (Medical Information Mart for Intensive Care III) and eICU Collaborative Research Database, two large-scale, publicly available datasets containing de-identified health records of ICU patients. These datasets have been widely used for clinical research and the development of AI-driven decision support systems. Our study leverages these resources to analyze potential disparities in the AI-based diagnosis of sepsis across different demographic subgroups.

1. **Key Features of MIMIC-III and eICU**
   (a) **Diverse Patient Populations:** MIMIC-III contains over 53,000 ICU admissions, while eICU provides data from over 200 hospitals across the United States, enabling analysis across institutions and demographic subgroups.
   (b) **Comprehensive Clinical Data:** Both datasets include structured and unstructured data, such as patient demographics, laboratory results, vital signs, medication records, ICU monitoring data, and clinician notes.
   (c) **Sepsis Identification:** Sepsis cases are identified using ICD-9 codes from the `DIAGNOSES_ICD` table, in alignment with prior sepsis prediction studies.
2. **Data Extraction and Preprocessing**
   (a) **Patient Cohort Selection:** We extract patients diagnosed with sepsis based on ICD-9 codes and include only ICU stays lasting longer than 24 hours to focus on patients receiving sustained critical care.
   (b) **Feature Engineering:** Construct relevant clinical features, including vital signs, lab results, sequential organ failure assessment (SOFA) scores, and timestamps for treatment initiation.
   (c) **Handling Missing Data:** Employ multiple imputation techniques for missing numerical data (e.g., mean, mode, and median imputation) and introduce separate categories for missing categorical variables.
   (d) **Socioeconomic Data Approximation:** Approximate socioeconomic status (SES) using ZIP codes linked to external census data on median income and healthcare access.
   (e) **Demographic Representation:** Extract race, gender, and insurance type attributes to analyze disparities in model performance across key demographic groups.

4.2 Analysis Plan

1. **Sepsis Prediction Modeling**
   (a) Train multiple AI models to predict sepsis onset:
   – **Baseline Models:** Logistic Regression, Random Forest
   – **Deep Learning:** Long Short-Term Memory (LSTM) networks and transformer-based models for time-series ICU data
   (b) Optimize models using cross-validation (80/20 train-test split).
   (c) Evaluate model performance using AUROC, precision-recall curves, and false negative rates.
2. **Bias and Fairness Analysis**
   (a) **Disparity Metrics:** Assess disparities in sepsis diagnosis using false negative rates (FNR), false positive rates (FPR), and predictive parity across demographic groups.
   (b) **Feature Attribution Analysis:** Utilize Shapley values to identify which clinical features disproportionately influence sepsis predictions for different racial and socioeconomic groups.
   (c) **Intersectional Bias Analysis:** Investigate whether bias is compounded for subgroups at the intersection of race, gender, and SES (e.g., Black, low-income, female patients).
   (d) **Temporal Bias Analysis:** Examine disparities in time-to-diagnosis and treatment initiation using survival analysis techniques.
3. **Bias Mitigation Strategies**
   (a) **Reweighting Methods:** Adjust training sample weights to ensure equitable subgroup representation.
   (b) **Fair Representation Learning:** Train debiased feature representations using adversarial debiasing techniques.
   (c) **Post-hoc Bias Correction:** Apply calibration strategies to mitigate disparate misclassification rates across demographic groups.

4.3 Statistical Tests for Bias

1. **Fairness Metrics:**
   (a) **Demographic Parity:** Measure differences in diagnosis rates across groups.
   (b) **Equalized Odds:** Compare false negative and false positive rates across demographic groups.
   (c) **Equal Opportunity:** Ensure comparable true positive rates across racial and socioeconomic subgroups.
2. **Survival Analysis for Time-to-Diagnosis:**
   (a) **Kaplan-Meier Curves:** Compare survival distributions and time-to-diagnosis across demographic subgroups.
3. **Significance Testing:**
   (a) **Chi-square tests:** Assess disparities in categorical clinical outcomes.
   (b) **ANOVA and t-tests:** Compare means of key clinical variables across demographic subgroups.
   (c) **Bootstrapping:** Compute confidence intervals for fairness metrics.

5 Evaluation

5.1 Evaluation

1. **Evaluation Metrics**
   (a) **Model Performance:** Evaluate AUROC, sensitivity, specificity, and F1-score for sepsis prediction models.
   (b) **Bias Metrics:** Measure disparities in false negative rates, time-to-treatment, and equalized odds across racial and socioeconomic groups.
   (c) **Fairness-Accuracy Trade-offs:** Assess the impact of bias mitigation techniques on overall model performance.

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