# Artificial Intelligence for Sepsis Prediction: Bias & User Interfaces

A Literature Review

Kathleen C. Fort

University College Dublin, Arcadia University

Table of Contents

Artificial Intelligence for Sepsis Prediction: Bias & User Interfaces

A Literature Review

## 1 Introduction

In recent years, machine learning (ML) and artificial intelligence (AI) have become integral to various aspects of life (Yang et al., 2023). As ML progresses, there is growing interest in its potential to assist clinicians in medical diagnoses and treatments (Eloranta et al., 2022). One notable area of focus is the prediction of sepsis onset.

Sepsis is a condition with one of the highest mortality and morbidity rates in the United States and other high-income countries (Singer et al., 2016). The average 30-day mortality rate for sepsis is 24.4%, which increases to 34.7% for septic shock (Bauer et al., 2020). Early identification and antibiotic treatment have been shown to improve patient outcomes significantly, with delays of just six hours increasing the mortality rate by 7.6% (Yan et al., 2022). Currently, sepsis risk is primarily assessed using diagnostic tools such as the Sepsis-related Organ Failure Assessment (SOFA) Score and the Acute Physiology and Chronic Health Evaluation (APACHE) (Islam et al., 2019). Adams et al. (2022) highlight machine learning's potential to reduce the global burden of sepsis by analyzing vital signs from a patient, assessing their sepsis risk, and helping clinicians anticipate the need for antibiotics.

A sepsis-predicting ML tool raises concerns about bias affecting the model's performance in clinical settings. Developers and clinicians must identify and mitigate biases to prevent the deployment of ineffective systems. Sources of bias include non-representative data sets, erroneous imputation techniques, and existing prejudices within the healthcare system (Fleuren et al., 2020).

This paper will assess past uses of ML tools for predicting sepsis and examine their associated risks of bias. Additionally, it will investigate the potential applications of AI in sepsis prediction, the challenges associated with implementing ML in a clinical setting, and areas where further research is required. Furthermore, this review will explore the gap between ML outputs and user comprehension, emphasizing considerations for creating an effective AI user interface (UI) for clinicians. Specifically, it will address Explainable AI (XAI), strategies to avoid screen fatigue, and the overall usability of the UI.

## 2 Definitions

Predicting sepsis is challenging in part due to the lack of unique identifying biomarkers. Additionally, the definition of sepsis varies between hospitals. According to an international task force, sepsis is broadly defined as a life-threatening organ dysfunction caused by the body's abnormal response to infection. This task force has identified the SOFA score as the best indicator for predicting in-hospital mortality from sepsis (Evans, 2018). Adams et al. (2022) found that a patient's risk of sepsis can be most efficiently and accurately assessed using the quick SOFA (qSOFA) score, which evaluates blood pressure, respiratory rate, and Glasgow Coma Scale score.

In cases where sepsis progresses, a patient may develop septic shock, defined as a subset of sepsis where circulatory, cellular, and metabolic abnormalities are associated with a higher risk of mortality compared to sepsis alone (Evans, 2018). Despite these standardized definitions, their integration into clinical practice remains inconsistent. Moreover, gaps remain in the definitions of sepsis, such as the lack of a precise definition of "infection" (Evans, 2018).

By extension, comparing different algorithms is challenging due to variations in the timing of sepsis detection. Sepsis prediction can occur before sepsis develops or at the onset of sepsis. Furthermore, models can either predict sepsis continuously (right-leaning algorithms) or at a specific point in time (left-leaning algorithms) (Fleuren et al., 2020). Additionally, depending on the model, sepsis identification can detect sepsis, severe sepsis, or septic shock. In the ICU, models are more likely to identify sepsis or severe sepsis, while in hospitals, the focus is often on septic shock.

## 3 Results

### *3.1 Performance and Bias*

Among algorithms for sepsis prediction, XGBoost and Random Forest models have consistently demonstrated high performance. These models have shown significant advantages in predicting sepsis in ICU patients, exhibiting higher accuracy (ACC) and concordance index (c-index) values compared to other models. However, a systematic review of models for sepsis prediction by Adams et al. (2022) found that most exhibited a high risk of bias, while the remaining models had an unknown risk of bias. These risks were attributed to factors such as small sample sizes, missing data, and the misinterpretation of complex data.

One source of bias in these algorithms is incorporation bias, which often occurs when investigational predictors are also determinative factors in defining the outcome (Prasad et al., 2023). This could involve using diagnostic data to predict the likelihood of a patient developing sepsis. This presents bias because such tests are typically ordered only if a clinician suspects the patient is deteriorating. A similar case can occur when using the administration of antibiotics as an indicator of sepsis risk. Training a model on this data risks categorizing patients as “low risk” simply because no diagnostic tests or antibiotics were ordered for them, causing both the model and the clinician to overlook those most at risk of delayed antibiotic treatment due to vague symptoms and no obvious vital sign abnormalities. A study reviewing 107 predictive algorithms found that none accounted for “informative observations,” where the presence of a diagnostic observation was non-random and driven by clinician concern. This oversight not only increases the risk of misdiagnosis but also exacerbates the likelihood of a “diagnostic deadlock,” where a clinician's suspicions are reinforced by the model, leading to further delays in diagnosis (Prasad et al., 2023).

One strategy to mitigate bias through data integration involves using “bland” data during model training. Prasad et al. investigated the effectiveness of employing “bland” clinical data to counteract bias within ML models. Their study revealed that the bland model exhibited subpar performance metrics, including elevated rates of false positives and poor sensitivity (Prasad et al., 2023).

*3.2 Imputation Methods*

Another avenue through which bias can infiltrate AI algorithms is via data imputation, the process by which missing data is handled within the model. Several studies have explored the use of the K-nearest neighbors (KNN) approach in conjunction with the SMOTE oversampling technique to address issues related to imbalanced datasets (Su et al., 2021; Böck et al., 2022; He et al., 2023). Research conducted by Memon et al. (2023) demonstrated that KNN imputation achieved superior precision in predicting missing values for categorical variables compared to methodologies such as Random Forest, Sequential Hot-Deck, and Multiple Imputation by Chained Equations.

Furthermore, Su et al. endorsed the application of KNN imputation specifically within sepsis prediction models, while acknowledging the potential efficacy of alternative methods such as stochastic regression and tree-based models for future comparative analyses (Su et al., 2021). In another investigation, Baniasadi et al. (2021) employed a two-step imputation strategy involving linear interpolation combined with a sample-weighted AdaBoost model to manage missing data effectively, enhancing model robustness and generalizability relative to other approaches.

### *3.3 Screen Fatigue*

Studies on screen fatigue have highlighted that, while electronic interfaces in clinical settings achieve high ratings in terms of data accuracy, precision, and processing efficiency, they also provoke significant dissatisfaction among users. This dissatisfaction stems from the perceived effort required, frequent interruptions, and frustration associated with navigating these interfaces (Hilty et al., 2022).

A meta-analysis by Dunn Lopez et al. emphasized that electronic usage in clinical settings can lead to fatigue, including emotional exhaustion, weariness, and challenges in maintaining engagement. Emotional repercussions such as anger, irritability, and stress were frequently reported, reflecting the adverse psychological impact of prolonged electronic interface interactions. Furthermore, there exists a noticeable absence of standardized assessments, monitoring mechanisms, or intervention protocols specifically tailored for technology in clinical workplaces beyond initial onboarding or training processes. These factors contribute to heightened complexity within the already stressful and high-pressure healthcare environment.

Eye strain from prolonged screen use in clinical settings contributes significantly to fatigue. Krupinski et al. (2010) found that increased screen time leads to symptoms like blurred vision and difficulty focusing. They also noted a rise in clinical errors after extended screen reading, especially when screens were viewed at close distances.

One of the primary drivers of screen fatigue is the extensive time spent using technology. Tutty et al. (2019) found that for every hour physicians spent on direct patient care, they spent nearly three additional hours interacting with electronic health records. This study underscored the critical role of leadership in engaging physicians during technology implementation and emphasized the importance of workflow design in the UI. It advised implementation teams to conduct thorough testing before and after implementation using scenarios aimed at minimizing clinical burden.

*3.4 Interpretability, Explainability, and Usability*

Ensuring the interpretability of AI models by clinicians, developers, and patients is crucial. As models become more complex to achieve higher performance and accuracy, there is a decline in explainability and transparency. Post hoc explanations are necessary to interpret outputs from ML algorithms rather than relying solely on inherent explainability. These methods include analyzing learned features and assessing feature importance and interactions. Two valuable methods for explaining AI in clinical contexts are Local Interpretable Model-Agnostic Explanations (LIME) and the use of Shapley values (SHAP) (van der Velden et al., 2022).

As AI usage expands in medical settings, clinical validation emerges as a fundamental requirement. Clinical validation ensures that AI systems can effectively perform in real-world scenarios, focusing on prediction accuracy and minimizing error rates (Amann et al., 2020). The explainability of AI tools is crucial for resolving disagreements between clinicians and AI by allowing clinicians to evaluate the model’s recommendations. Moreover, explainability allows users to identify model flaws and provide feedback to improve training and performance (Amann et al., 2020). Explainable AI (XAI) techniques, such as assessing feature importance, can be utilized in these scenarios (Di Martino et al., 2023).

## 4 Gaps

For widespread implementation of sepsis prediction algorithms in clinical settings, a standard definition of sepsis is essential to facilitate accurate comparisons of different prediction models and septic survival rates, as well as for identifying potential biases within these models (Bauer et al., 2020). Moreover, there exists a notable absence of a universally accepted checklist for evaluating the quality of diagnostic machine-learning research within medical contexts. Establishing such standards for assessing AI can mitigate the risk of biases in models (Dunn Lopez et al., 2018). Additionally, there is frequently insufficient transparency regarding sepsis prediction models, with many developers of high-performing algorithms withhold essential information such as the training dataset and model specifics. Enhancing transparency is crucial for the widespread adoption of sepsis prediction models in hospital settings (Yan et al., 2022).

Another gap in the implementation of sepsis prediction algorithms is the lack of data on sepsis patients in lower-income countries. This limits the use of sepsis prediction data and therefore sepsis prediction models to population levels (Fleuren et al., 2020).

Furthermore, there is a pressing need to develop models that rely on noninvasive or minimally invasive indicators. Many existing models evaluated in research depend on invasive procedures to achieve accurate predictions (Adams et al., 2022). For example, recent comprehensive studies suggest that arterial blood gas (ABG) results obtained through invasive methods may offer more accurate indicators of sepsis than previously understood (Fleuren et al., 2020). To address this challenge, our research group aims to mitigate bias by improving the accuracy of Pulse-Oximeter results as an alternative to ABG testing. This approach seeks to enhance the reliability and accessibility of sepsis prediction without the need for invasive procedures.

## 5 Conclusion

Machine learning has great potential to aid in early identification and prompt treatment of sepsis, leading to improved patient outcomes. However, its integration into clinical settings requires careful management of biases. Thus, it is crucial to ensure that training data is comprehensive and representative of diverse demographics, including age, sex, and ethnicity.

Furthermore, introducing machine learning UIs in clinical practice demands thoughtful consideration of usability factors to prevent screen fatigue, burnout among healthcare providers, and user frustration. Addressing these challenges through robust data practices and user-centered design principles can significantly enhance sepsis detection and management, ultimately improving patient care and reducing mortality rates.

References

Adams, R., Henry, K. E., Sridharan, A., Soleimani, H., Zhan, A., Rawat, N., Johnson, L., Hager, D. N., Cosgrove, S. E., Markowski, A., Klein, E. Y., Chen, E. S., Saheed, M. O., Henley, M., Miranda, S., Houston, K., Linton, R. C., Ahluwalia, A. R., Wu, A. W., & Saria, S. (2022). Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nature medicine*, *28*(7), 1455–1460. https://doi.org/10.1038/s41591-022-01894-0

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Precise4Q consortium (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, *20*(1), 310. https://doi.org/10.1186/s12911-020-01332-6

Baniasadi, A., Rezaeirad, S., Zare, H., & Ghassemi, M. M. (2021). Two-Step Imputation and AdaBoost-Based Classification for Early Prediction of Sepsis on Imbalanced Clinical Data. *Critical care medicine*, *49*(1), e91–e97. https://doi.org/10.1097/CCM.0000000000004705

Bauer, M., Gerlach, H., Vogelmann, T., Preissing, F., Stiefel, J., & Adam, D. (2020). Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019- results from a systematic review and meta-analysis. *Critical care (London, England)*, *24*(1), 239. https://doi.org/10.1186/s13054-020-02950-2

Böck, M., Malle, J., Pasterk, D., Kukina, H., Hasani, R., & Heitzinger, C. (2022). Superhuman performance on sepsis MIMIC-III data by distributional reinforcement learning. *PloS one*, *17*(11), e0275358. https://doi.org/10.1371/journal.pone.0275358

Di Martino, F., & Delmastro, F. (2023). Explainable AI for clinical and remote health applications: a survey on tabular and time series data. *Artificial intelligence review*, *56*(6), 5261–5315. https://doi.org/10.1007/s10462-022-10304-3

Dunn Lopez, K., & Fahey, L. (2018). Advocating for Greater Usability in Clinical Technologies: The Role of the Practicing Nurse. Critical Care Nursing Clinics of North America, 30(2), 247–257. https://www.sciencedirect.com/science/article/abs/pii/S0899588518300091?via%3Dihub

Eloranta, S., & Boman, M. (2022). Predictive models for clinical decision making: Deep dives in practical machine learning. Journal of Internal Medicine, 292(2), 278–295. https://doi.org/https://doi.org/10.1111/joim.13483

Evans, T. (2018). Diagnosis and management of sepsis. Clinical Medicine, 18(2), 146–149. https://www.sciencedirect.com/science/article/pii/S1470211824017354

Fleuren, L.M., Klausch, T.L.T., Zwager, C.L. *et al.* Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med* 46, 383–400 (2020). https://doi.org/10.1007/s00134-019-05872-y

He, Y., Liu, Y., Liu, Y., He, H., Liu, W., Huang, D., Gu, Z., & Ju, M. (2023). A machine-learning approach for prediction of hospital mortality in cancer-related sepsis. Clinical EHealth, 6, 17–23. https://www.sciencedirect.com/science/article/pii/S2588914123000126

Hilty, D. M., Armstrong, C. M., Smout, S. A., Crawford, A. C., Maheu, M. M., Drude, K. P., Chan, S., Yellowlees, P. M., & Krupinski, E. A. (2022). Findings and guidelines on provider technology, fatigue, and well-being: Scoping review. Journal of Medical Internet Research, 24(5), 1–15. https://www.jmir.org/2022/5/e34451

Islam, M. M., Nasrin, T., Walther, B. A., Wu, C. C., Yang, H. C., & Li, Y. C. (2019). Prediction of sepsis patients using machine learning approach: A meta-analysis. *Computer methods and programs in biomedicine*, *170*, 1–9. https://www.sciencedirect.com/science/article/pii/S016926071831602X?via%3Dihub

Krupinski, E. A., Berbaum, K. S., Caldwell, R. T., Shartz, K. M., & Kim, J. (2010). Long Radiology Workdays Reduce Detection and Accommodation Accuracy. Journal of the American College of Radiology, 7(9), 698–704. https://www.sciencedirect.com/science/article/pii/S1546144010001341

Memon, S. M., Wamala, R., & Kabano, I. H. (2023). A comparison of imputation methods for categorical data. Informatics in Medicine Unlocked, 42. https://www.sciencedirect.com/science/article/pii/S2352914823002289

Prasad, V., Aydemir, B., Kehoe, I.E., Kotturesh, C., O'Connell, A., et al. (2023). Diagnostic suspicion bias and machine learning: Breaking the awareness deadlock for sepsis detection. PLOS Digital Health. https://doi.org/https://doi.org/10.1371/journal.pdig.0000365

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J. D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J. L., & Angus, D. C. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, *315*(8), 801–810. https://doi.org/10.1001/jama.2016.0287

Su, L., Xu, Z., Chang, F., Ma, Y., Liu, S., Jiang, H., Wang, H., Li, D., Chen, H., Zhou, X., Hong, N., Zhu, W., & Long, Y. (2021). Early Prediction of Mortality, Severity, and Length of

Stay in the Intensive Care Unit of Sepsis Patients Based on Sepsis 3.0 by Machine Learning Models. *Frontiers in medicine*, *8*, 664966. https://doi.org/10.3389/fmed.2021.664966

Tutty, M. A., Carlasare, L. E., Lloyd, S., & Sinsky, C. A. (2019). The complex case of EHRs: examining the factors impacting the EHR user experience. *Journal of the American Medical Informatics Association : JAMIA*, *26*(7), 673–677. https://doi.org/10.1093/jamia/ocz021

van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis, 79. https://www.sciencedirect.com/science/article/pii/S1361841522001177

Yan, M. Y., Gustad, L. T., & Nytrø, Ø. (2022). Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review. *Journal of the American Medical Informatics Association : JAMIA*, *29*(3), 559–575. https://doi.org/10.1093/jamia/ocab236

Yang, Z., Cui, X., & Song, Z. (2023). Predicting sepsis onset in ICU using machine learning models: a systematic review and meta-analysis. *BMC infectious diseases*, *23*(1), 635. https://doi.org/10.1186/s12879-023-08614-0