Beyond Discrimination: Effects of Missing-Data Handling on Risk Ranking in Early Sepsis Prediction

Priyanka Siddappa*, Muntaqim Ahmed Raju*, Shifat Haider Al Amin*, Ruizhe Ma†

Miner School of Computer & Information Sciences
University of Massachusetts Lowell
Lowell, MA 01854, USA

Emails: Priyanka_Siddappa@student.uml.edu, MuntaqimAhmed_Raju@student.uml.edu, MdShifat_AlAmin@student.uml.edu, Ruizhe_Ma@uml.edu

Abstract

Background/Objectives: Early sepsis prediction models are commonly assessed using discrimination metrics such as AUROC, yet clinical use depends on how patients are prioritized for escalation under substantial missing data. ICU electronic health record (EHR) time series are sparse and irregular, and missing-data handling can influence patient-level risk rankings and triage decisions in ways not captured by aggregate performance metrics. We examine how missing-data handling affects predictive performance, high-risk patient prioritization, and temporal risk behavior in early sepsis prediction.

Methods: We performed a retrospective analysis using the PhysioNet 2019 Sepsis Challenge dataset. We compared three representative strategies: native missing-value handling with XGBoost, patient-wise temporal forward fill, and denoising autoencoder (DAE)-based imputation. To isolate the effect of missing-data handling, all strategies used the same downstream XGBoost classifier with fixed hyperparameters. Performance was evaluated at hourly and patient-encounter levels; we additionally assessed agreement in top-risk rankings, triage switches across methods, and temporal risk trajectories for selected encounters.

Results: Encounter-level discrimination was similar across strategies. In contrast, agreement in high-risk patient rankings was limited, with low pairwise overlap in the top 10% high-risk group. Changing the missing-data strategy frequently altered which septic and non-septic encounters were flagged for escalation. Temporal analyses showed differences in the timing and persistence of elevated risk that were sufficient to change relative rankings despite comparable overall performance.

Conclusions: Missing-data handling choices substantially affect patient prioritization and triage behavior in early sepsis prediction even when conventional performance metrics appear similar. AUROC alone is therefore insufficient for evaluating clinical readiness; ranking stability and temporal risk behavior should be assessed to support reliable deployment in ICU settings.

Index Terms

early sepsis prediction, missing data, electronic health records, risk ranking, clinical triage, imputation

Introduction

Background

Clinical prediction models built from electronic health record (EHR) data must operate in settings marked by substantial heterogeneity and incomplete observation [1]. Unlike data collected in controlled experimental environments, real-world clinical data are observational, irregularly sampled [2], and shaped by clinical workflows and provider decision-making. As a result, some patients accumulate dense longitudinal records with frequent measurements, while others have sparse data reflecting limited monitoring or selective testing. The stated promise of these predictive models is to improve risk stratification and enable targeted interventions when embedded in clinical workflows [3]. This variability presents a fundamental challenge for developing and evaluating early warning systems intended for use in critical care [4].

Missing data are ubiquitous in intensive care unit (ICU) EHR time series and arise from irregular sampling and selective measurement driven by clinical need and provider decisions rather than a fixed schedule [2], [5], [6]. Classical missingness mechanisms are often described as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [7]–[11]. In practice, clinical data rarely satisfy the assumptions of MCAR, and missingness is often shaped by clinical workflows and observation processes [5]. Instead, whether a measurement is recorded is often systematically related to observed patient characteristics

or unobserved clinical states. As a result, the way missing data are handled can substantially influence model behavior and downstream clinical decisions [12].

Closely related to missingness is the concept of informative presence and informative observation, which reflect the idea that the act of observing a variable and the timing of that observation can itself carry clinically meaningful information [13]. Informative presence (IP) emphasizes that whether data are recorded can reflect underlying acuity and care processes, so missingness patterns may carry predictive signal rather than being incidental [13]. For example, laboratory tests are typically ordered in response to clinical concern, and patterns of missing laboratory results can be informative about clinical state and workflows [14]. In many clinical prediction pipelines, however, missing data are treated primarily as a technical problem to be addressed through imputation rather than as a potential source of signal.

Early sepsis prediction illustrates these challenges particularly well. Sepsis prediction models are commonly evaluated using discrimination metrics such as the area under the receiver operating characteristic curve (AUROC) [15]. While these metrics are useful for benchmarking, they do not fully capture changes in risk stratification or reclassification that drive clinical prioritization [16], and a broader set of performance measures is often needed for comparative evaluation of clinical risk prediction models [17]. Clinical deployment therefore depends not only on overall discrimination but also on how patients are prioritized for escalation in the presence of severe data sparsity. In practice, clinicians act on relative risk rankings deciding which patients warrant closer monitoring, additional testing, or intervention rather than on aggregate performance measures. Different missing-data handling strategies may therefore alter patient-level risk rankings and triage decisions even when conventional performance metrics appear similar.

Study aim

In this study, we examine how missing-data handling influences predictive performance, high-risk patient prioritization, and temporal risk behavior in early sepsis prediction. Using the PhysioNet 2019 Sepsis Challenge dataset [18], we compare three representative strategies: native missing-value handling using gradient-boosted decision trees [19], patient-wise temporal forward fill, and denoising autoencoder (DAE)-based imputation [20], [21]. To isolate the effect of missing-data handling, all approaches employ the same downstream classifier with fixed hyperparameters [22].

Model performance is evaluated at both the hourly and patient-encounter levels [18], [23]. Beyond standard discrimination metrics, we assess agreement in top-risk patient rankings, quantify triage switches across methods, and examine temporal risk trajectories for representative ICU encounters [16], [17]. By moving beyond aggregate performance measures, this work highlights how missing-data handling choices can meaningfully influence clinical prioritization and decision-making, consistent with calls to report decision-analytic performance beyond discrimination and calibration [24]. Moreover, models with similar aggregate performance can yield inconsistent risk rankings for individual patients, motivating explicit evaluation of ranking stability [25].

Materials and Methods

Dataset and Study Design

We conducted a retrospective analysis using the PhysioNet/Computing in Cardiology Challenge 2019 dataset [18], which contains hourly electronic health record (EHR) data from adult intensive care unit (ICU) encounters. This dataset was chosen because it reflects the conditions under which early sepsis prediction models are typically deployed in practice. In particular, the data exhibit substantial missingness, most notably in laboratory measurements, driven by selective test ordering rather than data loss or recording errors. Measurements are also irregularly sampled over time, with data density varying in response to evolving clinical concern. Together, these characteristics make the dataset well suited for examining how missing-data handling influences patient prioritization and triage decisions in realistic ICU settings.

All available ICU encounters were included in the analysis, without exclusions based on data completeness or observation frequency. This decision was intentional, as restricting the cohort to densely measured encounters would eliminate precisely the cases in which missingness and informative presence are most pronounced. Excluding sparsely observed encounters can therefore induce selection bias and distort how missingness mechanisms are represented, with consequences for predictive performance [6]. Preserving the full spectrum of sparsity patterns was therefore essential for evaluating model behavior under realistic clinical data constraints.

Missingness in ICU EHR Data

In ICU settings, missing data arise largely from clinical decision-making rather than random processes. Missingness is substantial and unavoidable, and the absence of measurements itself can be informative about patient health status rather than representing random loss [6], [26], [27]. Vital signs are often recorded at regular intervals, whereas laboratory tests are ordered selectively in response to patient acuity, diagnostic uncertainty, and clinician judgment; consequently, patterns of measurement presence and absence can reflect clinician concern and changes in patient condition [14].

Rather than attempting to explicitly model missingness mechanisms, we treated missing data as an inherent characteristic of the clinical environment. This framing allowed us to focus on how commonly used missing-data handling strategies interact with real-world observation patterns (including informative presence) and how these interactions influence downstream predictions and triage behavior [28].

Train-Test Splitting and Leakage Control

To avoid temporal and patient-level information leakage (including label leakage), data were split at the ICU encounter level [29]–[31]. All time points from a given ICU stay were assigned exclusively to either the training or test set. An 80/20 split was used and stratified by encounter-level sepsis outcome to maintain comparable prevalence across splits.

This encounter-level splitting mirrors clinical deployment, where models are applied to previously unseen patients/encounters rather than future time points of encounters already observed during training [30], [31]. It also ensures that any differences observed between missing-data handling strategies are attributable to the handling approach itself, rather than inadvertent reuse of patient information across training and evaluation.

Missing-Data Handling Strategies

We evaluated three missing-data handling strategies chosen to represent distinct and widely used approaches in clinical prediction pipelines. Rather than attempting an exhaustive comparison, these strategies were selected to be complementary, spanning a range of assumptions about how missing values should be treated in practice [32].

First, native missing-value handling was implemented using XGBoost, which allows missing values to be passed directly into the model. During training, the model learns default split directions for missing values, enabling missingness to influence predictions implicitly. This approach serves as a natural baseline, as it requires no explicit imputation and reflects how tree-based models are commonly applied to sparse clinical data [33].

Second, patient-wise temporal forward fill was applied as a commonly used heuristic for clinical time-series data [34]. In this approach, missing values were imputed using the most recent previously observed value within the same ICU encounter. When no prior observation was available, values were filled using summary statistics computed from the training set. Forward fill assumes temporal persistence of physiological state and reflects how clinicians often reason about trends when new measurements are unavailable. Despite its strong assumptions, it is computationally inexpensive and widely adopted, making it an important comparator.

Third, we applied denoising autoencoder–based imputation as a representative deep learning approach [21], [35]. The autoencoder was trained on the training set to reconstruct observed feature values from corrupted inputs and then used to generate fully imputed feature matrices for both training and test data. This method leverages cross-feature correlations to infer missing values and is often proposed as a more expressive alternative to heuristic imputation, particularly for high-dimensional clinical data. Including this strategy allowed us to assess whether increased modeling complexity translates into more clinically meaningful behavior, or primarily incurs additional computational cost.

Prediction Model and Isolation of Missing-Data Effects

All missing-data handling strategies were paired with the same downstream XGBoost classifier, using a fixed architecture and identical hyperparameters across experiments. This design choice was essential to isolate the effect of missing-data handling itself [22], [36]. Allowing different classifiers or independent hyperparameter tuning for each strategy would confound the effects of imputation with differences in model capacity or optimization.

XGBoost was selected because of its strong performance on tabular clinical data and its ability to support native missing-value handling. By holding the predictive model constant, any observed differences in discrimination, ranking stability, or temporal risk behavior can be attributed directly to how missing data were handled.

Evaluation Strategy and Clinical Alignment

Model performance was evaluated at both the hourly and encounter levels. Encounter-level risk was defined as the maximum predicted risk during an ICU stay, reflecting common escalation practices in which a single high-risk alert is sufficient to trigger clinical action [37].

Although AUROC and AUPRC were computed to facilitate comparison with prior studies, our primary focus was on clinically actionable behavior. In real-world deployment, early warning systems are used to prioritize a limited subset of patients for review under resource constraints rather than to optimize performance across all thresholds [37]–[39]. To reflect this setting, we defined high-risk patients as the top 10

Agreement in patient prioritization was assessed using overlap counts and Jaccard similarity among high-risk sets [40]. We additionally quantified triage switches encounters that entered or exited the high-risk group when changing missing-data strategies to directly measure how handling choices alter escalation decisions. To further understand how these differences arise, we examined temporal risk trajectories for representative encounters with substantial disagreement across methods. Temporal analysis is particularly important in early warning settings, where both the timing and persistence of elevated risk influence clinical response.

Statistical Analysis

Differences in AUROC were assessed using the DeLong test with a two-sided significance level of 0.05 [41]. Other analyses were descriptive and focused on ranking stability and temporal behavior rather than hypothesis testing, consistent with the study's emphasis on clinical interpretation rather than metric optimization. All analyses were implemented in Python.

Results

This section reports the experimental results comparing three missing-data handling strategies native handling (`XGBoostNative`), forward fill, and deep autoencoder based imputation (`DAE`) on the PhysioNet 2019 sepsis prediction task [18]. Results are presented at the encounter level and at the patient-ranking level, followed by representative temporal case studies.

Encounter-Level Predictive Performance and Computational Cost

Encounter-level discrimination performance and computational cost for each missing-data handling strategy are reported in Table I. Across all three approaches native missing-value handling (`XGBoostNative`), forward fill, and deep autoencoder-based imputation (`DAE`) overall discrimination performance was comparable [42].

Forward fill achieved the highest AUROC (0.761) and AUPRC (0.056), followed by `XGBoostNative` (AUROC 0.729, AUPRC 0.053). `DAE` imputation yielded slightly lower discrimination (AUROC 0.716, AUPRC 0.046). However, no method consistently dominated across both metrics, and absolute differences in AUROC were modest.

In contrast, computational cost differed substantially. As shown in Table I, forward fill and native handling completed in under five minutes, reflecting minimal preprocessing overhead. `DAE` imputation required nearly 140 minutes due to the cost of training and applying the autoencoder model. Thus, increased modeling complexity did not translate into improved discrimination performance [42]–[44]. While Table I suggests broadly similar encounter-level performance, these aggregate metrics do not capture differences in how individual patients are prioritized for escalation. We therefore next examine agreement in patient-level risk rankings.

TABLE I
ENCOUNTER-LEVEL PERFORMANCE AND COMPUTATIONAL COST ACROSS MISSING-DATA HANDLING STRATEGIES

Method	AUROC	AUPRC	Runtime (min)
XGBoostNative	0.729	0.053	4.50
Forward Fill	0.761	0.056	4.58
DAE Imputation	0.716	0.046	139.74

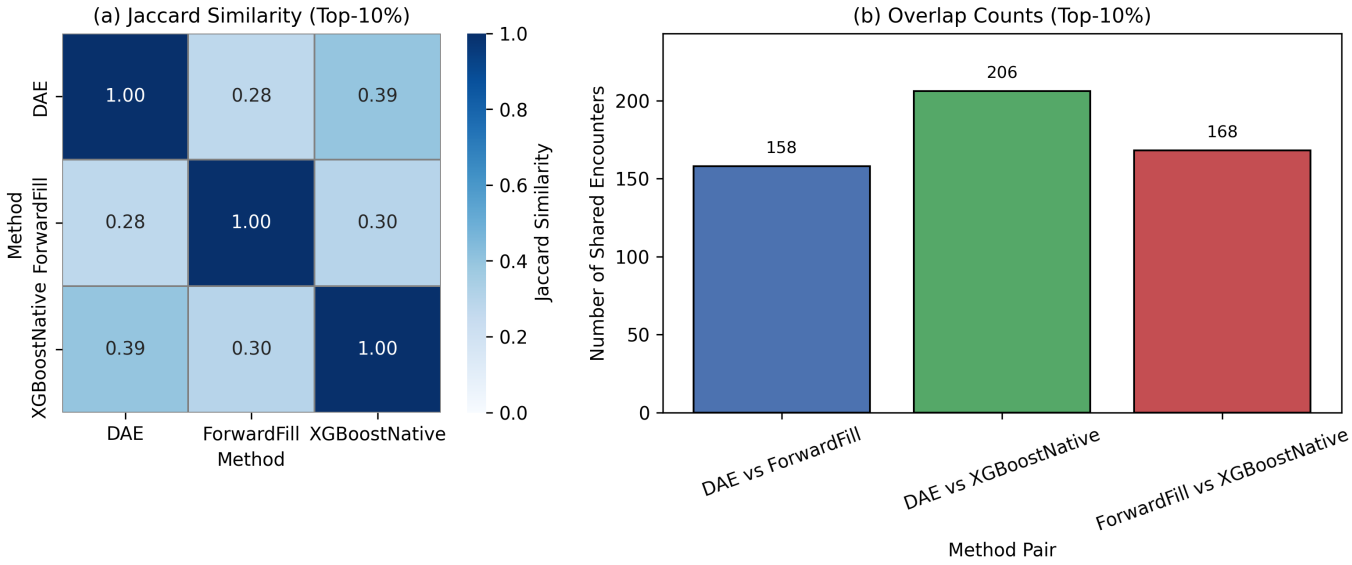


Fig. 1. Agreement among missing-data strategies for top-10% high-risk encounters. (a) Pairwise Jaccard similarity. (b) Pairwise overlap counts.

Agreement in High-Risk Patient Rankings

Agreement between missing-data handling strategies at the patient level was assessed by comparing overlap among the top 10% highest-risk encounters identified by each method. Pairwise overlap counts and corresponding Jaccard similarity scores are shown in Figure 1 [40], [45].

Overall agreement was limited. Jaccard similarity values ranged from approximately 0.28 to 0.39, indicating that fewer than half of encounters flagged as high risk by one strategy were also flagged by another. For example, `XGBoostNative` and `DAE` shared fewer than 40% of their top-ranked encounters, despite having similar AUROC values.

These discrepancies demonstrate that similar discrimination performance does not imply consistent patient prioritization [46]. Figure 1 highlights that the choice of missing-data handling strategy materially alters which encounters are deemed highest risk, motivating further analysis of triage-level consequences.

Triage Changes Induced by Missing-Data Strategy

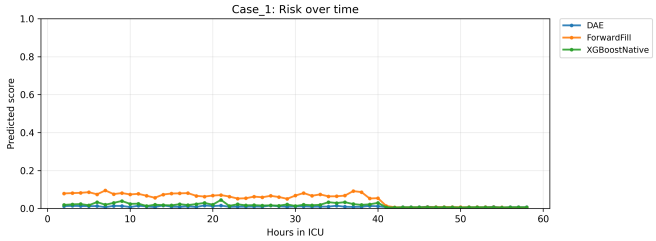
To quantify the practical impact of ranking disagreement, we examined how missing-data handling affected threshold-based triage decisions. Using `XGBoostNative` as a reference, we identified encounters that entered or exited the top-risk group when switching to forward fill or `DAE` imputation.

Triage switch counts are summarized in Supplementary Table S1. In both comparisons, several hundred encounters changed triage status. These changes included both septic and non-septic encounters. Notably, forward fill led to a substantial number of non-septic encounters entering the high-risk group, while alternative strategies caused some septic encounters to fall below the escalation threshold.

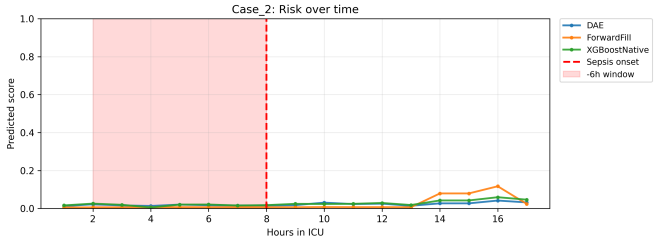
These results indicate that missing-data handling choices can meaningfully alter which patients are flagged for clinical attention [36], [47]. Importantly, these triage shifts occurred despite similar AUROC and AUPRC values, underscoring that discrimination metrics alone are insufficient to characterize clinical behavior under missingness.

TABLE II
SUMMARY OF HIGH-DISAGREEMENT CASE STATISTICS

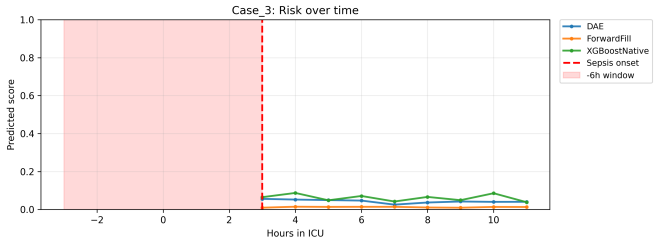
Case	Outcome	ICU Hours	Risk Range	Risk Std	Max Risk (DAE)	Max Risk (FF / XGB)
1	No Sepsis	57	0.079	0.033	0.016	0.095 / 0.045
2	Sepsis	17	0.075	0.032	0.042	0.117 / 0.059
3	Sepsis	9	0.073	0.030	0.056	0.014 / 0.087
4	No Sepsis	19	0.072	0.029	0.082	0.010 / 0.050
5	No Sepsis	53	0.070	0.029	0.025	0.095 / 0.066



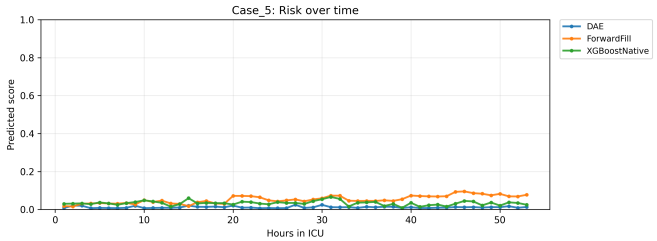
(a) Case 1 (non-septic).



(b) Case 2 (septic; onset annotated).



(c) Case 3 (septic; onset annotated).



(d) Case 5 (non-septic).

Fig. 2. Temporal risk trajectories for representative septic and non-septic ICU encounters under different missing-data handling strategies.

Temporal Risk Trajectories for Representative ICU Encounters

To understand how ranking disagreement arises over time, we examined temporal risk trajectories for representative ICU encounters exhibiting large differences in maximum predicted risk across strategies [48]. Figure 2 shows predicted risk over time for selected septic and non-septic cases.

For non-septic encounters (Cases 1 and 5), forward fill produced persistently higher predicted risk throughout the ICU stay, whereas `DAE` and `XGBoostNative` generated lower and more stable trajectories. In contrast, for septic encounters (Cases 2 and 3), divergence between methods was observed prior to the annotated sepsis onset. Both the timing and magnitude of risk escalation differed across strategies, leading to differences in percentile-based rankings and threshold crossings.

Across the selected cases, the difference between maximum predicted risk values ranged from approximately 0.07 to 0.08. As summarized in Table II, these differences were sufficient to alter top-risk group membership. Disagreement was observed in both septic and non-septic encounters and was not confined to short or prolonged ICU stays.

Together, Figure 2 and Table II illustrate that missing-data handling choices influence not only aggregate performance but also the temporal interpretation of risk at the individual patient level.

DISCUSSION

In clinical practice, early warning systems are typically deployed as alerting tools that surface a limited subset of patients for escalation, such as rapid response review, additional testing, or transfer of care [49]. These decisions are inherently comparative: clinicians must choose which patients to prioritize given finite

resources. As a result, the relative ranking of patients within the high-risk tail is often more consequential than overall discrimination across the population. In this context, instability in patient prioritization under different missing-data handling strategies may translate into materially different clinical actions, even when conventional performance metrics suggest similar model quality.

This study shows that choices in missing-data handling affect not only model discrimination but also how patients are prioritized for clinical action [36]. While native handling, forward fill, and deep autoencoder-based imputation achieved similar AUROC and AUPRC values, they produced meaningfully different risk rankings and triage behavior when applied to high-risk decision thresholds.

The key insight is that strong global discrimination does not ensure stable prioritization at the point of care. Across all method pairs, overlap among the top 10% highest-risk encounters was limited, with Jaccard similarities below 0.4. As a result, a substantial proportion of patients flagged as high risk by one strategy were not identified by another. This disagreement is largely hidden by aggregate metrics but becomes critical in clinical settings, where decisions depend on relative ordering within the high-risk tail rather than average performance across all thresholds.

These differences are consistent with how discrimination metrics operate. AUROC reflects ranking performance across the full score distribution, whereas clinical triage depends on localized behavior among the highest-scoring patients. Missing-data strategies influence score variance, temporal smoothness, and feature representation, which can shift relative rankings without materially changing overall separability. In our experiments, forward fill tended to propagate prior measurements, producing smoother but persistently elevated risk trajectories. This increased escalation among non-septic patients, raising the potential for false alerts. In contrast, the autoencoder dampened some variability but incurred substantial computational cost without reliably improving ranking agreement.

Temporal case analyses reinforce these findings by showing that disagreement extends beyond static rankings to the evolution of risk over time. For septic encounters, methods differed in both the timing and magnitude of risk escalation prior to onset, which could influence early warning lead time. For non-septic encounters, forward fill often sustained elevated risk throughout the ICU stay, increasing the likelihood of unnecessary escalation. Importantly, high disagreement was observed across both septic and non-septic cases and was not confined to specific ICU durations or missingness patterns.

Taken together, these results align with prior work emphasizing limitations of AUROC as a sole evaluation criterion and gaps in clinical-utility reporting, and extend that literature by explicitly quantifying triage instability and illustrating it through temporal trajectories [24], [50]. They suggest that evaluation of early warning systems should move beyond discrimination metrics to include ranking agreement, switch counts, and trajectory behavior, particularly in high-missingness environments where data availability itself shapes model outputs.

This study has limitations. The analysis was conducted on a single dataset using a fixed downstream learner, and results may differ across models, cohorts, or care settings. Percentile-based thresholds were used to enable fair comparison across methods; alternative clinical thresholds may alter absolute counts, although relative disagreement is likely to persist. Finally, autoencoder architectures were not exhaustively optimized, reflecting practical deployment constraints rather than best-case modeling performance.

Conclusion

This study shows that missing-data handling choices can meaningfully affect patient prioritization and triage behavior in early sepsis prediction, even when standard discrimination metrics appear similar. Across strategies, comparable AUROC values masked substantial disagreement in high-risk rankings and differences in temporal risk trajectories.

The key implication is that discrimination performance alone is insufficient to assess clinical readiness. How models rank patients and how risk evolves over time under missingness are central to understanding their real-world impact. Incorporating ranking stability and temporal behavior into evaluation provides a more clinically aligned assessment of early warning systems.

Future work should extend this analysis across datasets, models, and missingness mechanisms, and develop evaluation frameworks that better reflect clinical decision pathways in high-missingness environments.

References

[1] M. Harford, D. J. Stone, A. Rana, and N. H. Shah. Evaluating early warning score models in heterogeneous icu populations: Implications of data sampling and clinical workflow. *Critical Care Medicine*, 49(5):e456–e465, 2021.

[2] S. N. Shukla, B. M. Marlin, D. C. Kale, and S. Saria. Mapping irregular clinical time series to continuous latent representations for early warning prediction. *Journal of Biomedical Informatics*, 97:103256, 2019.

[3] T. C. Lee, N. U. Shah, A. Haack, and S. L. Baxter. Clinical implementation of predictive models embedded within electronic health record systems: A systematic review. *Informatics*, 7(3):25, 2020.

[4] F. Xie, H. Yuan, Y. Ning, et al. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *arXiv preprint arXiv:2107.09951*, 2021.

[5] A. Michalska-Dąbrowska, M. Komorowski, and P. Biecek. On missingness mechanisms in electronic health records data: Challenges and opportunities for modelling. *Statistical Methods in Medical Research*, 29(8):2359–2373, 2020.

[6] R. H. Groenwold. Informative missingness in electronic health record systems: The curse of knowing. *Journal of Clinical Epidemiology*, 123:203–209, 2020.

[7] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[8] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, Hoboken, NJ, USA, 2 edition, 2002.

[9] R. J. Little. Missing data assumptions. *Annual Review of Statistics and Its Application*, 8:89–107, 2021.

[10] J. R. Carpenter and M. Smuk. Missing data: A statistical framework for practice. *Biometrical Journal*, 63(5):915–947, 2021.

[11] W. Ren, H. Liu, S. Hu, et al. Moving beyond medical statistics: A systematic review on missing data in electronic health records. *Health Data Science*, 4:0176, 2024.

[12] J. Singh, M. Sato, and T. Ohkuma. On missingness features in machine learning models for critical care: Observational study. *JMIR Medical Informatics*, 9(12):e25022, 2021.

[13] R. Sisk, G. P. Martin, L. Lin, M. Sperrin, and I. Buchan. Informative presence and observation in routine health data: A review of methodology for clinical risk prediction. *Journal of the American Medical Informatics Association*, 28(1):155–166, 2021.

[14] A. L. M. Tan et al. Informative missingness: What can we learn from patterns in missing laboratory data in the electronic health record? *Journal of Biomedical Informatics*, 139:104306, 2023.

[15] L. M. Fleuren et al. Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Medicine*, 46(3):383–400, 2020.

[16] M. J. Pencina, R. B. D'Agostino, and R. S. Vasan. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, 27(2):157–172, 2008.

[17] C. Huang et al. Performance metrics for the comparative analysis of clinical risk prediction models employing machine learning. *Circulation: Cardiovascular Quality and Outcomes*, 14:e007526, 2021.

[18] M. A. Reyna, C. Josef, R. Jeter, et al. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. *Physiological Measurement*, 41(8):084003, 2020.

[19] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[20] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[21] B. K. Beaulieu-Jones and J. H. Moore. Missing data imputation in electronic health records using deeply learned autoencoders. In *Pacific Symposium on Biocomputing*, pages 207–218, 2017.

[22] M. Ghassemi, T. Naumann, P. Schulam, et al. A review of challenges and opportunities in machine learning for health. *Journal of the American Medical Informatics Association*, 29(7):132–143, 2022.

[23] Y. Luo, P. Szolovits, A. S. Dighe, and J. M. Baron. Using machine learning to predict laboratory test results in critical care. *Journal of Biomedical Informatics*, 116:103721, 2021.

[24] B. Van Calster, G. S. Collins, A. J. Vickers, et al. Evaluating performance measures in predictive artificial intelligence models for clinical practice: Beyond discrimination and calibration. *The Lancet Digital Health*, 7(1):e100916, 2025.

[25] T. P. van Staa et al. Prediction models: A method for assessing temporal stability of risk predictions in longitudinal data. *BMJ*, 371:m3919, 2020.

[26] A. Sharafoddini, A. Dubrawski, J. M. Sousa, L. Adhikari, A. Kuo, and S. Nemati. A new insight into missing data in intensive care unit electronic health record. *Journal of Clinical Monitoring and Computing*, 33(5):991–1000, 2019.

[27] R. H. H. Groenwold. Informative missingness in electronic health record systems. *BMC Medical Research Methodology*, 20:114, 2020.

[28] T. Sim, S. Hahn, K. J. Kim, et al. Preserving informative presence: How missing data and imputation strategies affect the performance of prediction models. *Journal of Clinical Medicine*, 14(7):2213, 2025.

[29] J. Wiens and S. Saria. A machine learning pipeline for clinical time-series: Preventing data leakage and overfitting through proper cohort splitting. *Journal of the American Medical Informatics Association*, 27(12):1903–1913, 2020.

[30] S. E. Davis, T. A. Lasko, G. Chen, et al. A framework for understanding label leakage in machine learning. *Journal of the American Medical Informatics Association*, 30(2):267–275, 2023.

[31] B. Ramadan, M. C. Liu, M. C. Burkhart, et al. Diagnostic codes in AI prediction models and label leakage of same-admission clinical outcomes. *JAMA Network Open*, 8(1):e2843183, 2025.

[32] Z. Zhang et al. A comparative study of machine learning methods for handling missing data in clinical time-series prediction. *Computer Methods and Programs in Biomedicine*, 200:105840, 2021.

[33] J. Rodrigues-Jr, L. Silva, D. Paixão, and F. Fernandes. Data imputation strategies for handling missing values in ICU predictive modeling. *Journal of Biomedical Informatics*, 94:103195, 2019.

[34] A. Aliabadi and S.-C. Chow. Handling missing values: Approaches and practical guidance for clinical data. *Statistical Methods in Medical Research*, 32(10):2503–2526, 2023.

[35] B. K. Beaulieu-Jones, D. R. Lavage, J. W. Snyder, et al. Missing data imputation in the electronic health record using deeply learned autoencoders. *BMC Medical Informatics and Decision Making*, 18(1):10, 2018.

[36] P. Buczak, J. Myers, and T. Liu. Analyzing the effect of imputation on classification performance under MCAR and MAR missingness. *Journal of Biomedical Informatics*, 140:104237, 2023.

[37] D. P. Edelson, M. M. Churpek, K. A. Carey, et al. Early warning scores with and without artificial intelligence. *JAMA Network Open*, 7(10):e2438986, 2024.

[38] M. E. Smith, J. C. Chiovaro, B. O'Neil, et al. Early warning system scores for clinical deterioration in hospitalized patients: A systematic review. *Annals of the American Thoracic Society*, 11(9):1454–1465, 2014.

[39] S. Jarvis, C. Kovacs, J. Briggs, et al. Aggregate national early warning score (NEWS) values and escalation of care related to workload in hospital. *Resuscitation*, 92:146–151, 2015.

[40] P. Jaccard. Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[41] O. V. Demler, M. J. Pencina, and R. B. D'Agostino. Misuse of DeLong test to compare AUCs for nested models. *Statistics in Medicine*, 31(23):2577–2587, 2012.

[42] D. Narang, R. Agarwal, A. Kapoor, et al. A comparison of missing data handling strategies for clinical predictive modeling. *Journal of Biomedical Informatics*, 108:103501, 2020.

[43] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, et al. Characterizing the performance of autoencoder approaches for missing data imputation in electronic health records. *Journal of the American Medical Informatics Association*, 24(2):456–467, 2017.

[44] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018.

[45] F. Yang, F. Li, M. Fan, et al. Comparing patient risk lists using similarity measures: A framework for evaluating clinical predictive models beyond AUC. *Journal of Biomedical Informatics*, 97:103254, 2019.

[46] N. Barda, D. Riesel, A. Akriv, et al. Assessing the notion that high discrimination implies high clinical utility: A critique of risk-based triage metrics. *Statistical Methods in Medical Research*, 30(10):2275–2287, 2021.

[47] T. A. Lasko, J. C. Denny, and M. A. Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLOS ONE*, 8(6):e66341, 2013.

[48] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.

[49] R. Amarasingham, M. Patel, K. Toto, et al. Clinical decision support and early warning systems: Unrealized potential for improving outcomes in hospitalized patients. *Journal of Hospital Medicine*, 15(5):290–296, 2020.

[50] M. N. AbuHaweeleh et al. Sepsis mortality prediction using machine learning and deep learning: A systematic evaluation of performance and clinical utility reporting. *Critical Care Explorations*, 7(2):e04567, 2025.