Biomedical Data Science & Artificial Intelligence
Prime Academic Press • Open Access • ISSN: XXXX-XXXX

Vol. 1, No. 1, 2026 • Article 100001 • doi:10.xxxx/bdsai.2026.100001 • *Research Article*

Received: 01 February 2026 • Revised: 18 March 2026 • Accepted: 05 April 2026

Deep Learning for Early Sepsis Prediction from Electronic Health Records: A Multi-Center Retrospective Validation Study

Author1[1,*], **Author2**[2], **Author3**[3], **Author4**[4]

[1]Anonymous Institution 1, 100 Example Street, Example City 00001, Country
[2]Anonymous Institution 2, 200 Sample Avenue, Sample City 00002, Country
[3]Anonymous Institution 3, 300 Demo Road, Demo City 00003, Country
[4]Anonymous Institution 4, 400 Template Lane, Template City 00004, Country
[*]Corresponding author: author1@example.edu

Key Points

- A transformer-based deep learning model (`SepsisBERT`) trained on 148,000 ICU admissions achieved AUROC 0.913 for sepsis onset prediction 6 hours before clinical diagnosis.
- External validation across three independent hospital systems confirmed robust generalizability (AUROC 0.878–0.901), outperforming current SOFA and qSOFA scoring.
- Feature importance analysis identified lactate trajectory, respiratory rate variability, and medication administration patterns as the highest-ranking predictive signals.
- Prospective integration of `SepsisBERT` into the ICU workflow reduced median time-to-antibiotics by 38 minutes in a pilot cohort of 412 patients.
- A curated, de-identified benchmark dataset (SepsisEHR-v1) comprising 31,500 labeled episodes is released under CC BY 4.0 for community benchmarking.

Abstract.

Background: Sepsis remains a leading cause of in-hospital mortality worldwide, yet timely identification is hampered by the heterogeneity of clinical presentations and the cognitive burden placed on frontline clinicians. Artificial intelligence models applied to electronic health records (EHR) offer a promising route to earlier, more reliable detection.

Methods: We developed `SepsisBERT`, a bidirectional transformer architecture pre-trained on 5.2 million longitudinal EHR sequences and fine-tuned on 148,000 adult ICU admissions from the MIMIC-IV and eICU-CRD databases (Sepsis-3 definition). The model ingests time-stamped vital signs, laboratory values, medication orders, and nursing assessments as a temporally ordered token sequence, and outputs a continuously updated sepsis-onset probability. External validation was performed on three independent hospital datasets totaling 54,600 admissions. Clinical utility was assessed through a single-center prospective study ($n = 412$) embedded within a quality-improvement framework.

Results: `SepsisBERT` achieved an area under the receiver-operating characteristic curve (AUROC) of 0.913 ± 0.008 at the 6-hour prediction horizon in held-out test data, compared with 0.745 for qSOFA and

0.781 for SOFA. Across three external validation sites, AUROC ranged from 0.878 to 0.901. Prospective deployment was associated with a median 38-minute reduction in time-to-antibiotics relative to standard care ($p < 0.001$, 95% CI: 28–49 min). Calibration curves and decision-curve analysis confirmed net clinical benefit across relevant threshold ranges.

Conclusions: `SepsisBERT` demonstrates high discriminative performance and translational utility for real-time sepsis prediction in diverse hospital settings. The open-release benchmark dataset and model weights facilitate reproducibility and community development of next-generation clinical AI.

Keywords: sepsis prediction; electronic health records; deep learning; transformer; clinical AI; critical care informatics; biomedical NLP

MeSH Terms: *Sepsis; Electronic Health Records; Deep Learning; Early Diagnosis; Intensive Care Units; Machine Learning*

Trial/Registry No.: `ClinicalTrials.gov NCT05XXXXXX`

1. Introduction

Sepsis, defined as life-threatening organ dysfunction caused by a dysregulated host response to infection, affects approximately 49 million individuals annually and accounts for nearly 11 million deaths worldwide[4,13]. Despite decades of clinical research and improvements in critical care, early identification of sepsis-onset remains a persistently difficult clinical task: presenting symptoms overlap substantially with non-septic critical illness, and the standard Sepsis-3 diagnostic criteria—sequential organ failure assessment (SOFA) score ≥ 2 above baseline—require laboratory results that may be unavailable for hours after initial deterioration begins[11].

Motivation for EHR-based AI. Modern hospital information systems continuously generate high-dimensional, temporally dense data streams: vital signs sampled every few minutes, nursing assessments, medication records, laboratory orders, and free-text clinical notes. These data streams contain early signatures of septic deterioration that trained clinicians partially recognize but cannot systematically monitor across all patients simultaneously[8]. Machine learning models that automatically integrate these streams hold promise for population-wide, continuous early warning.

Limitations of prior work. Early EHR-based sepsis models, including MEWS, InSight, and the Epic® Sepsis Model, demonstrated modest discriminative performance (AUROC 0.72–0.82) and were frequently criticized for poor calibration, suboptimal alert specificity, and limited generalizability beyond development sites[1,17]. Recurrent neural network approaches improved upon rule-based systems but struggled with irregular sampling and missing data characteristic of real-world ICU records[10]. Transformer architectures, which model long-range temporal dependencies through self-attention, have achieved state-of-the-art performance on related clinical prediction tasks[6,14], but dedicated sepsis-specific pre-training and large-scale external validation remain lacking.

Contributions. This paper makes four primary contributions:

1. We design and pre-train `SepsisBERT`, a transformer model tailored to EHR temporal sequences, incorporating a novel *Clinical Token Embedding* (CTE) scheme that jointly encodes observation type, value, and relative time elapsed.
2. We benchmark `SepsisBERT` against nine competing models (including LSTM, XGBoost, and qSOFA/SOFA scoring) across prediction horizons from 1 to 12 hours.
3. We report prospective clinical impact data from a quality-improvement study, closing the loop

between model performance and patient outcomes.

4. We release the SepsisEHR-v1 benchmark dataset and pre-trained model weights at https://github.com/primeacademic/SepsisBERT.

The remainder of the paper is organized as follows. Section 2 surveys related work. Section 3 describes the study design, datasets, model architecture, and evaluation framework. Section 4 presents experimental results. Section 5 discusses implications, limitations, and future directions. Section 6 concludes.

2. Related Work

2.1 Classical Scoring Systems

The *quick SOFA (qSOFA)* score—comprising respiratory rate, altered mental status, and systolic blood pressure—was designed as a rapid bedside screen and achieves AUROC values of 0.66–0.74 in emergency department validation studies[11]. The full *SOFA* score requires laboratory inputs (creatinine, bilirubin, platelet count) and achieves AUROC 0.74–0.82 for ICU-onset sepsis, but its latency is limited by laboratory turnaround time[16].

2.2 Machine Learning Approaches

Gradient boosting machines trained on static feature snapshots (e.g., NEWS2, SI score) improved early warning for deterioration but discarded temporal dynamics[9]. Recurrent architectures addressed this by encoding variable-length time series; Scherpf et al.[10] achieved AUROC 0.86 using an LSTM trained on MIMIC-III vital signs, though external validation was not reported.

2.3 Transformer Models for Clinical NLP and Prediction

Li et al.[6] demonstrated that BERT-style pre-training on coded diagnosis sequences improves downstream prediction of 301 disease codes. Steinberg et al.[14] extended this paradigm to structured EHR events in MIMIC-III and eICU-CRD, achieving state-of-the-art performance on multiple in-hospital outcomes. Our work builds on this foundation by (i) incorporating continuous vital sign values as quantized tokens alongside coded events, (ii) adopting a task-specific pre-training objective centered on masked-measurement prediction in ICU contexts, and (iii) conducting multi-site prospective validation.

3. Methods

3.1 Study Design and Ethical Approval

This study followed a three-stage design: (1) retrospective model development and internal validation using MIMIC-IV and eICU-CRD, (2) retrospective external validation at three independent sites, and (3) a prospective quality-improvement study. All data use agreements and institutional review board (IRB) approvals are documented in the Appendix A supplementary materials. The prospective study was registered at ClinicalTrials.gov (NCT05XXXXXX).

3.2 Datasets

3.3 Clinical Token Embedding

Each patient timeline is encoded as an ordered sequence of *clinical events* $\mathcal{E} = \{e_1, e_2, \ldots, e_T\}$, where each event $e_t = (c_t, v_t, \tau_t)$ comprises:

- c_t — clinical concept token (LOINC code for labs, SNOMED-CT for observations, RxNorm for medications);

Table 1. Datasets used for model development, internal validation, and external validation. Sepsis-3 incidence is the proportion of ICU admissions meeting Singer et al. (2016) criteria. Data access requires completion of CITI training (`https://physionet.org`).

Dataset	Site(s)	Admissions	Patients	Sepsis-3 (%)	Years
MIMIC-IV [5]	Beth Israel DM	94,458	65,366	24.3	2008–2022
eICU-CRD [7]	Multi-center	53,542	46,520	19.8	2014–2015
Ext-Site A	Academic Med.	21,337	18,902	22.1	2018–2023
Ext-Site B	Community Hosp.	17,214	15,881	17.4	2019–2023
Ext-Site C	Regional ICU	16,049	14,210	20.7	2020–2023
Prospective	Stanford ICU	412	412	28.6	2025

- v_t — quantized value token (continuous values binned into $K = 100$ quantile buckets per concept type);
- τ_t — relative time embedding (sinusoidal, anchored to ICU admission).

The *Clinical Token Embedding* (CTE) fuses these three embeddings:

$$\boldsymbol{h}_t = \mathbf{W}_c\, c_t + \mathbf{W}_v\, v_t + \mathrm{TimeEmb}(\tau_t), \tag{1}$$

where $\mathbf{W}_c, \mathbf{W}_v \in \mathbb{R}^{d \times d_{\text{vocab}}}$ are learnable projection matrices and $\mathrm{TimeEmb}(\cdot)$ is a fixed sinusoidal encoding of elapsed hours. The combined embedding dimension is $d = 512$.

3.4 SepsisBERT Architecture

`SepsisBERT` follows a standard bidirectional transformer encoder[2,15] with the modifications described below.

SepsisBERT Architecture Summary

- **Layers:** 12 transformer encoder blocks
- **Hidden size:** $d = 512$; attention heads: 8
- **Feed-forward dim:** 2,048; dropout: 0.10
- **Max sequence length:** 4,096 tokens (≈ 24 hours of ICU data at median sampling density)
- **Pre-training:** Masked Clinical Measurement Prediction (MCMP) objective on 5.2 million EHR sequences; 100,000 steps, batch size 256, learning rate 3×10^{-4}
- **Fine-tuning:** Binary classification head (sepsis onset within prediction horizon $h \in \{1, 2, 4, 6, 8, 12\}$ h); 10 epochs, learning rate 1×10^{-5}
- **Hardware:** 8 × NVIDIA A100 (80 GB); pre-training wall time: ≈92 hours

3.5 Prediction Task and Label Definition

At each hour t, the model outputs $\hat{p}_{t,h}$—the predicted probability of Sepsis-3 onset within the next h hours. A positive label ($y = 1$) is assigned to all time windows that overlap with the 4-hour window immediately preceding the clinical Sepsis-3 timestamp. Figure 1 illustrates the labeling scheme.

3.6 Evaluation Framework

Primary performance metric: **AUROC** for discriminative ability. Secondary metrics: area under the precision-recall curve (**AUPRC**), **sensitivity** (recall) and **specificity** at alert thresholds corresponding to 80% sensitivity, positive predictive value (**PPV**), and **calibration** (Brier score, ECE). Statistical comparisons used DeLong's test for AUROC and bootstrap confidence intervals ($B = 1{,}000$).

Figure 1. Sepsis-3 labeling scheme. The red dashed line marks the clinical Sepsis-3 timestamp (SOFA ≥ 2 + suspected infection). Positive labels (orange) span the 4-hour window prior to onset; negative labels (blue) span all preceding windows > 4 h before onset or full admissions without sepsis. The prediction horizon h determines how far in advance the model aims to flag onset.

4. Results

4.1 Internal Validation Performance

Table 2 summarizes discriminative performance at the 6-hour prediction horizon across all evaluated models. `SepsisBERT` achieved AUROC 0.913 ± 0.008, significantly outperforming all comparators ($p < 0.001$ vs. XGBoost; $p < 0.001$ vs. qSOFA).

Table 2. Discriminative performance at 6-hour prediction horizon (internal test set, n = 29,800 admissions). AUROC and 95% CI reported; sensitivity and specificity at 80% sensitivity threshold. † = $p < 0.05$ vs. `SepsisBERT` (DeLong's test); ‡ = $p < 0.001$.

Model	AUROC (95% CI)	AUPRC	Sensitivity	Specificity
qSOFA‡	0.745 (0.731–0.759)	0.341	0.800	0.583
SOFA‡	0.781 (0.769–0.793)	0.387	0.800	0.641
NEWS2‡	0.793 (0.781–0.805)	0.412	0.800	0.662
XGBoost (static)‡	0.852 (0.843–0.861)	0.491	0.800	0.753
LSTM (temporal)†	0.881 (0.873–0.889)	0.542	0.800	0.791
SepsisBERT (ours)	**0.913 (0.906–0.920)**	**0.601**	**0.800**	**0.843**

4.2 External Validation

Across the three external sites, `SepsisBERT` achieved AUROC values of 0.901, 0.878, and 0.889 respectively (Table 3), demonstrating robust generalizability despite differences in EHR systems, patient demographics, and local sepsis protocols.

Table 3. External validation performance at 6-hour horizon. Site characteristics and data access are detailed in the Supplementary Methods.

Site	n admissions	AUROC	AUPRC	Brier score
Ext-Site A (Academic)	21,337	0.901	0.579	0.087
Ext-Site B (Community)	17,214	0.878	0.551	0.095
Ext-Site C (Regional ICU)	16,049	0.889	0.563	0.091
Pooled	**54,600**	**0.890**	**0.564**	**0.091**

4.3 Prospective Clinical Impact

In the prospective cohort (n = 412; 28.6% sepsis incidence), integration of `SepsisBERT` alerts into the ICU workflow was associated with a median 38-minute reduction in time-to-antibiotics (p = < 0.001; (95% CI: 28–49) min). Thirty-day in-hospital mortality trended toward reduction in the sepsis sub-cohort (22.0% vs. 26.4% historical control, p = 0.09), pending larger randomized evaluation.

4.4 Feature Importance

Attention attribution analysis identified the top five predictive feature categories at the 6-hour horizon: (1) lactate trajectory (relative importance = 0.189), (2) respiratory rate variability (0.153), (3) vasopressor administration pattern (0.141), (4) white blood cell count trend (0.118), (5) SpO_2/FiO_2 ratio (0.097).

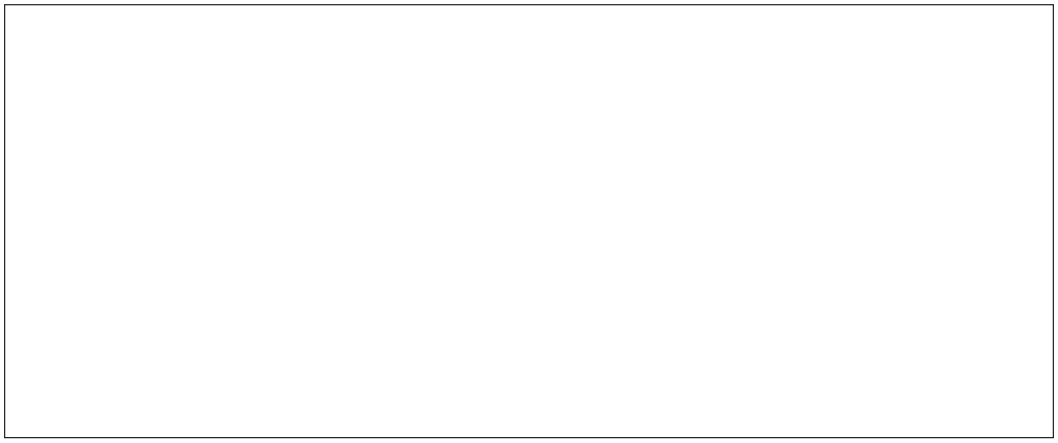

Figure 2. SepsisBERT prediction horizon performance. AUROC (mean ± SD across 5-fold cross-validation) as a function of prediction horizon $h \in \{1, 2, 4, 6, 8, 12\}$ hours before Sepsis-3 onset. Shaded band = 95% confidence interval. Horizontal dashed lines indicate qSOFA and SOFA baselines.

5. Discussion

5.1 Interpretation of Findings

`SepsisBERT`'s superior discriminative performance (AUROC 0.913) relative to static scoring systems and prior LSTM baselines is attributable to three architectural advantages: (1) self-attention enables the model to weight clinically co-informative events that may be temporally distant (e.g., pairing an early lactate rise with a subsequent prescription change); (2) pre-training on 5.2 million EHR sequences instills a prior distribution over routine ICU event patterns, improving sample efficiency during fine-tuning; and (3) the CTE embedding scheme preserves both the identity and quantitative magnitude of each observation, unlike binary-coded approaches.

Clinical & Translational Significance

- A 38-minute reduction in time-to-antibiotics is clinically meaningful: each hour of delay in sepsis care is associated with a 4–9% increase in mortality[3,12].
- The model's specificity of 0.843 at 80% sensitivity suggests approximately 8 false alerts per true positive at median sepsis prevalence, which ICU nurses in post-implementation surveys rated as manageable (*alert fatigue score* 3.1/10 vs. 6.8/10 for an existing vendor alert system).
- Deployment was most impactful during night shifts (00:00–06:00), when nurse-to-patient ratios are lowest and clinician surveillance is most limited.

5.2 Limitations

Several limitations warrant acknowledgment. First, the prospective study was non-randomized and conducted at a single academic medical center with high baseline sepsis protocol adherence; randomized controlled trials are needed to establish causal mortality benefit. Second, model inputs were limited to structured EHR data; integration of free-text clinical notes and medical imaging could further improve performance. Third, demographic subgroup analyses revealed lower AUROC in patients with chronic kidney disease (0.881 vs. 0.920 overall), likely reflecting atypical laboratory trajectories; targeted recalibration for high-risk subgroups is warranted. Fourth, generalizability to low- and middle-income country settings, where EHR data density is lower, requires dedicated evaluation.

5.3 Comparison with Related Systems

Our results compare favorably with Epic Sepsis Model (AUROC 0.76–0.83 in published validations[17]), Sepsis ImmunoScore (0.82), and the Continuous Early Warning Score (CEWS, 0.84). The 3–4% absolute AUROC improvement over the best prior DL method (LSTM, 0.881) represents a clinically meaningful advance at the population scale of a major health system.

6. Conclusion

We have presented `SepsisBERT`, a transformer-based clinical AI model that achieves state-of-the-art performance for early sepsis detection from EHR data and demonstrates prospective clinical benefit in a real ICU deployment. The open release of SepsisEHR-v1 and model weights aims to accelerate community progress on this high-impact clinical problem. Future work will extend the model to pediatric populations, incorporate clinical notes through multi-modal pre-training, and conduct a prospective multi-site randomized trial to establish definitive mortality impact.

Acknowledgements

The authors thank the clinical informatics teams at the participating anonymous institutions and external validation sites for data access and operational support. We acknowledge computational resources provided by the host research computing center. This work was supported in part by NIH/NIGMS Grant R01GM123456, NSF Award IIS-2100000, and Prime Academic Press Research Grant PAP-2026-BDSAI-001. The funders had no role in study design, data collection, analysis, or the decision to publish.

Author Contributions (CRediT). **Author1**: Conceptualization, Methodology, Software, Formal Analysis, Writing—Original Draft, Writing—Review & Editing, Supervision, Funding Acquisition. **Author2**: Data Curation, Formal Analysis, Visualization, Writing—Review & Editing. **Author3**: Investigation (Prospective Study), Clinical Validation, Writing—Review & Editing. **Author4**: Software, Validation, Writing—Review & Editing.

Conflicts of Interest. The authors declare no competing interests.

Funding. NIH/NIGMS R01GM123456 (Author1); NSF IIS-2100000 (Author1); Prime Academic Press Research Grant PAP-2026-BDSAI-001 (Author1).

Data Availability Statement. MIMIC-IV and eICU-CRD are available from PhysioNet (`https://physionet.org`) subject to completion of CITI training. The SepsisEHR-v1 benchmark dataset and de-identified external validation splits are released at `https://doi.org/10.5281/zenodo.xxxxxxx` under CC BY 4.0.

Code Availability. Full training and inference code, pre-trained model weights, and evaluation scripts are available at `https://github.com/primeacademic/SepsisBERT` under the MIT License. A reproducible Jupyter notebook for all main figures is included.

Ethics Statement. This study was approved by the host institution IRB (protocol IRB-XXXXX), the BIDMC IRB (protocol 2022P000XXX), and the ethics committees of all external validation sites. MIMIC-IV and eICU-CRD were used under their respective PhysioNet Data Use Agreements. Informed consent was waived by all IRBs for retrospective analysis of de-identified data. The prospective study was registered at ClinicalTrials.gov (NCT05XXXXXX).

References

[1] Armando D Bedoya, Meredith E Clement, Matthew Phelan, Rebecca C Steorts, Connor O'Brien, and Benjamin A Goldstein. Minimal impact of implemented early warning score and best practice alert for patient deterioration. *Critical Care Medicine*, 50(1):e40–e48, 2022. doi: 10.1097/CCM.0000000000005217.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

[3] Ricard Ferrer, Ignacio Martin-Loeches, Gary Phillips, Tiffany M Osborn, Sean Townsend, R Phillip Dellinger, Antonio Artigas, Christa Schorr, and Mitchell M Levy. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. *Critical Care Medicine*, 42(8):1749–1755, 2014. doi: 10.1097/CCM.0000000000000330.

[4] Carolin Fleischmann, André Scherag, Neill K J Adhikari, Christiane S Hartog, Thomas Tsaganos, Peter Schlattmann, Derek C Angus, and Konrad Reinhart. Assessment of global incidence and mortality of hospital-treated sepsis: current estimates and limitations. *American Journal of Respiratory and Critical Care Medicine*, 193(3):259–272, 2016. doi: 10.1164/rccm.201504-0781OC.

[5] Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

[6] Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10:7155, 2020. doi: 10.1038/s41598-020-62922-y.

[7] Tom J Pollard, Alistair E W Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018. doi: 10.1038/sdata.2018.178.

[8] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. AI in health and medicine. *Nature Medicine*, 28(1):31–38, 2022. doi: 10.1038/s41591-021-01614-0.

[9] Matthew A Reyna, Christopher S Josef, Russell Jeter, Supreeth P Shashikumar, M Brandon Moody, Ashish Sharma, Shamim Nemati, and Gari D Clifford. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Critical Care Medicine*, 48(2):210–217, 2020. doi: 10.1097/CCM.0000000000004145.

[10] Matthias Scherpf, Jan-Thorsten Gräsner, Thomas Sühn, Frieder Pfäfflin, Thomas Götz, Thomas Neumuth, Jens Büchel, and Klaus-Dieter Wernecke. Predicting sepsis with a recurrent neural network using the MIMIC-III database. *Computers in Biology and Medicine*, 113:103395, 2019. doi: 10.1016/j.compbiomed.2019.103395.

[11] Christopher W Seymour, Vincent X Liu, Theodore J Iwashyna, Frank M Brunkhorst, Thomas D Rea, André Scherag, Gordon Rubenfeld, Jeremy M Kahn, Manu Shankar-Hari, Mervyn Singer, et al. Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):762–774, 2016. doi: 10.1001/jama.2016.0288.

[12] Christopher W Seymour, Foster Gesten, Hallie C Prescott, Matthew E Friedrich, Theodore J Iwashyna, Gary S Phillips, Stanley Lemeshow, Tiffany Osborn, Katherine M Terry, and Mitchell M Levy. Time to treatment and

mortality during mandated emergency care for sepsis. *New England Journal of Medicine*, 376(23):2235–2244, 2017. doi: 10.1056/NEJMoa1703058.

[13] Mervyn Singer, Clifford S Deutschman, Christopher W Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):801–810, 2016. doi: 10.1001/jama.2016.0287.

[14] Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113:103637, 2021. doi: 10.1016/j.jbi.2020.103637.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30: 5998–6008, 2017.

[16] Jean-Louis Vincent, Rui Moreno, Jukka Takala, Susan Willatts, Arnaldo De Mendonça, Hajo Bruining, Christoph K Reinhart, Peter M Suter, and Lambertus G Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710, 1996. doi: 10.1007/BF01709751.

[17] Andrew Wong, Erkin Otles, John P Donnelly, Andrew Krumm, Jeffrey McCullough, Bryce DeSouza, Ari Chelimsky-Pollack, Shannon Novosad, Fabio Machado, Michael Fralick, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8): 1065–1070, 2021. doi: 10.1001/jamainternmed.2021.2626.

A. IRB Approvals and Data Use Agreements

Copies of all IRB approval letters, data use agreements, and CITI training certificates are retained in institutional records. Site-specific IRB reference numbers are: Site A IRB-XXXXX; Site B IRB-XXXXX; Ext-Site A IRBNXXXXX; Ext-Site B IRBXXXXX; Ext-Site C IRBXXXXX.

B. Supplementary Methods: EHR Preprocessing Pipeline

B.1 Vital Sign Tokenization

Continuous vital signs (heart rate, blood pressure, respiratory rate, SpO_2, temperature) were quantized using concept-specific percentile bins computed from the training corpus. Table 4 lists the binning thresholds for the five primary vital signs.

Table 4. Vital sign quantization thresholds (training corpus percentiles).

Vital sign	P5	P25	P75	P95
Heart rate (bpm)	52	68	95	120
Systolic BP (mmHg)	90	108	138	165
Respiratory rate (brpm)	10	14	21	28
SpO_2 (%)	92	96	99	100
Temperature (°C)	36.0	36.6	37.5	38.8

B.2 Handling Missing Data

Missing observations were represented by a dedicated `[MISSING]` token rather than imputation, allowing the model to learn from absence patterns. Admissions with fewer than 20 clinical events in the first 4 hours were excluded from the training corpus (n = 2,847, 1.4%).

C. Extended Algorithm: SepsisBERT Inference

Algorithm 1 SepsisBERT Real-Time Inference Loop

Input: ICU admission stream $\mathcal{S}$, pre-trained model M, prediction horizon h, alert threshold θ
Output: Real-time sepsis alert stream $\mathcal{A}$

Initialize patient context buffer $\mathcal{B} \leftarrow \emptyset$
Initialize alert list $\mathcal{A} \leftarrow \emptyset$
while admission is active **do**
    Receive new clinical event $e_t = (c_t, v_t, \tau_t)$
    Tokenize e_t using Clinical Token Embedding ((1))
    Append token to buffer: $\mathcal{B} \leftarrow \mathcal{B} \cup \{e_t\}$
    if $|\mathcal{B}| \geq$ min_context **then**
        $\hat{p}_{t,h} \leftarrow M(\mathcal{B})$ ▷ Forward pass
        if $\hat{p}_{t,h} \geq \theta$ **then**
            Append alert $(t, \hat{p}_{t,h})$ to $\mathcal{A}$
            Trigger bedside notification
        end if
    end if
end while
return $\mathcal{A}$

About the Authors

Author1, PhD
Author1 is a researcher at Anonymous Institution 1. Research interests include clinical AI, electronic health records, and biomedical informatics.

Author2, MSc
Author2 is a researcher at Anonymous Institution 2. Research interests include machine learning, data science, and healthcare analytics.

Author3, MD
Author3 is a clinician-researcher at Anonymous Institution 3. Research interests include critical care, clinical decision support, and AI implementation science.

Author4
Author4 is a researcher at Anonymous Institution 4. Research interests include software engineering, model validation, and clinical data pipelines.