Supplementary Information

Implementation Variability Makes Sepsis-3 an Inconsistent Measuring Model for Early AI Prediction on Clinical Data

Jan Mackensen[1], Michael Staniek[1], Yutong He[1], Verena Schneider-Lindner[2], and Stefan Riezler[1,3]

[1]Department of Computational Linguistics, Heidelberg University, Im Neuenheimer Feld 325, 69120 Heidelberg, Germany
[2]Department of Anesthesiology, Surgical Intensive Care Medicine and Pain Medicine, Medical Faculty Mannheim, Heidelberg University, Theodor-Kutzer-Ufer 1-3, 68167 Mannheim, Germany
[3]Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, INF 205, 69120 Heidelberg, Germany

S1 eICU

S1.1 ICD-Based Sepsis Labels

The sepsis labels for the eICU dataset were generated using the dataset's diagnosis table, which contains diagnoses and timestamps entered by medical staff for each patient, with diagnoses specified via ICD-9 codes. Labels were derived from these codes as follows: for patients with an ICD code explicitly indicating sepsis (785.52, 995.91, or 995.92), the timestamp of that diagnosis was used directly as the sepsis label. For patients without a sepsis-specific ICD code, the Angus criterion [1] was applied instead. This criterion classifies a patient as septic if they have ICD codes indicating both an infection and organ failure [1]. For the eICU dataset, temporal proximity between the two codes was additionally required: the organ failure code had to be assigned within 48 hours before or 24 hours after the infection-related code. Where this criterion was met, the earlier of the two timestamps was taken as the sepsis onset.

S1.2 Sepsis-3 Compared to ICD-Based Annotations

A comparison of the ICD-based sepsis labels generated for the eICU dataset with the annotations produced by the implementation variants of the Sepsis-3 definition confirms many of the findings from Experiment 2, while also revealing accuracy problems with labels generated from ICD codes. An analysis of variance (LMEM) applied to the eICU data showed that the two components accounting for the most variance in the metrics were the OD function and sepsis onset. These were also two of the three highest-variance components in Experiment 2. The third component, the SI definition, is absent from the eICU analysis, since only the "Multiple Antibiotics" variant of the SI definitions can be applied to eICU. As in Experiment 2, the SOFA version, data imputation method, and SI window contributed little to no variance, depending on the metric considered. A notable difference from Experiment 2, however, is that sepsis onset accounts for a substantially larger share of the variance in the specificity metric for the eICU data (see Table S1).

Variance Component	Sensitivity σ^2 (%)	Specificity σ^2 (%)	MCC σ^2 (%)	F1 σ^2 (%)
Sepsis-Onset	**21.5012**	30.4298	**37.2031**	**34.0450**
OD-Function	10.6839	**33.4513**	21.6736	17.4022
SOFA-Version	1.1988	3.8639	6.5947	5.2812
Data-Imputation	1.8866	7.6105	5.7636	3.6713
SI-Window	1.1039	2.3684	2.8341	2.7392
Residual	63.6257	22.2761	25.9308	36.8611

Table S1: Variance in sensitivity, specificity, MCC, and F1 across different Sepsis-3 implementation variants was analyzed using the eICU dataset and assessed with a linear mixed-effects model.

S2 Interaction Effects of the Sepsis-3 Components

In addition to the LMEM used in Experiment 2, which modeled the individual components of the Sepsis-3 definition as independent effects, we fitted a second LMEM that additionally accounts for pairwise interaction effects between components. The results are shown in Table S2. As in the independent-effects model, SI-Definition, Sepsis-Onset, and OD-Function remain the largest variance components across all metrics. However, the independent effect of OD-Function is markedly smaller for Sensitivity, MCC, and F1 than in the Experiment 2 model, since for these three metrics the interaction effect between Sepsis-Onset and OD-Function is substantial. All other interaction effects are negligible across metrics, and the independent effects of SOFA-Version, Data-Imputation, and SI-Window remain small, consistent with Experiment 2.

Overall, this analysis confirms the findings of Experiment 2: SI-Definition, Sepsis-Onset, and SI-Window contribute the most variance, while the remaining components contribute little to annotation variability. Additionally, the interaction between Sepsis-Onset and OD-Function substantially reduces the independent effect of OD-Function. This can be explained as follows: except for the *Baseline* variant, all OD-Function variants detect a SOFA score increase of ≥ 2 within the SI-Window; they differ not in whether this change is detected, but in *when* it is detected, which time point is defined as t_{SOFA}. Crucially, this shift in t_{SOFA} only affects the outcome if the chosen Sepsis-Onset variant also uses t_{SOFA} to determine t_{SEPSIS}. This dependency produces the strong interaction effect, which accounts for much of the independent OD-Function effect observed in Experiment 2.

S3 Results for Experiments 2 for public-domain Sepsis-3 implementations.

This section summarizes the results of Experiments 2 and 3 for the well-known public-domain Sepsis-3 implementations (MIMIC-Code [2], PhysioNet [3], and ricu [4]). Additionally, the best- and worst-performing implementation variants from Experiment 2 are presented to contextualize the results of the public-domain implementations. The results of the various implementations for Experiment 2 are shown in Table S3. While ricu is the best-performing of the three public implementations in terms of F1, it is outperformed by the best possible implementation variant by 6%. Beyond F1, ricu is also outperformed by this variant on all other metrics except specificity. MIMIC-Code, by contrast, performs close to the worst implementation variant in terms of F1, achieving an F1 score of approximately 9%, compared to 6% for the worst variant. PhysioNet's performance falls between the other two implementations and closely matches the intercept of our LMEM across all metrics. Notably, PhysioNet achieves the highest specificity among the five definitions.

Variance Component	Sensitivity σ^2 (%)	Specificity σ^2 (%)	MCC σ^2 (%)	F1 σ^2 (%)
Independent Effects				
SI-Definition	**57.2999**	**61.2768**	1.2366	**35.9644**
Sepsis-Onset	12.0679	0.4513	**33.0522**	24.2109
OD-Function	4.5761	29.6737	14.1985	2.8176
SOFA-Version	0.0000	0.0000	0.2047	0.0000
Data-Imputation	0.0000	0.1279	0.0000	0.0000
SI-Window	0.4152	3.4646	0.9605	0.0000
Interaction Effects				
Sepsis-Onset × OD-Function	**14.4559**	0.2239	**33.6188**	**22.0505**
SI-Definition × OD-Function	1.3834	**1.3378**	0.8132	4.9157
Sepsis-Onset × SI-Window	2.4499	0.3939	3.4889	2.5479
OD-Function × SI-Window	2.3718	1.2361	1.4197	1.9262
SI-Definition × Sepsis-Onset	1.3547	0.4046	3.1781	0.9779
OD-Function × Data-Imputation	0.2222	0.1660	1.1369	0.5063
Sepsis-Onset × SOFA-Version	0.1273	0.0035	0.6620	0.3950
OD-Function × SOFA-Version	0.1242	0.0745	0.2394	0.2673
SOFA-Version × Data-Imputation	0.4495	0.4778	0.0111	0.2403
SI-Definition × Data-Imputation	0.0519	0.0252	0.3617	0.1963
SI-Definition × SOFA-Version	0.0964	0.0098	0.2840	0.1925
Sepsis-Onset × Data-Imputation	0.0311	0.0002	0.1686	0.1189
SI-Definition × SI-Window	0.2668	0.1188	0.3671	0.0807
SI-Window × SOFA-Version	0.0021	0.0675	0.2483	0.0476
SI-Window × Data-Imputation	0.0023	0.0174	0.0558	0.0116
Residual	2.2516	0.4486	4.2940	2.5322

Table S2: Variance in matching SepsisExp ground truth labels in % including the effects of pairwise interactions.

Definition	**Configuration**					**Metrics**			
	SOFA-Version	SI-Definition	SI-Window	OD-Function	t_{SEPSIS}	Sensitivity	Specificity	MCC	F1
Worst Definition	SOFA-1	Multiple Antibiotics	[48, 24]	Baseline	$\min(t_{\text{SOFA}}, t_{\text{SI}})$	0.05	0.80	-0.17	0.06
MIMIC-Code[2]	SOFA-1	Antibiotics and Body Fluids	[48, 24]	Baseline	$\min(t_{\text{SOFA}}, t_{\text{SI}})$	0.14	0.66	-0.16	0.09
PhysioNet[3]	SOFA-1	Antibiotics ≥ 4 Days and Blood Culture	[24, 12]	Current vs. Minimum (24h)	$\min(t_{\text{SOFA}}, t_{\text{SI}})$	0.16	0.83	-0.02	0.19
ricu[4]	SOFA-1	Antibiotics and Body Fluids	[48, 24]	Current vs. Minimum (SI-Window)	t_{SOFA}	0.25	0.72	-0.03	0.22
Best Definition	SOFA-1	Antibiotics ≥ 2 doses and Body Fluids	[48, 24]	Current vs. Minimum (24h)	$\max(t_{\text{SOFA}}, t_{\text{SI}})$	0.37	0.68	0.04	0.28

Table S3: Overview of the most frequently used implementation configurations, plus the worst and best configurations according to F1 in our evaluation (SepsisExp dataset). In all implementations, LOCF was chosen as the Data-Imputation method.

References

[1] Angus, DC, Linde-Zwirble, WT, Lidicker, J, Clermont, G, Carcillo, J, and Pinsky, MR. "Epidemiology of severe sepsis in the united states: analysis of incidence, outcome, and associated costs of care". In: *Critical care medicine* 29.7 (2001), pp. 1303–1310. DOI: https://doi.org/10.1097/00003246-200107000-00002. URL: https://www.ovid.com/00003246-200107000-00002.

[2] Johnson, Alistair E W, Stone, David J, Celi, Leo A, and Pollard, Tom J. "The mimic code repository: enabling reproducibility in critical care research". In: *Journal of the american medical informatics association* 25.1 (2018), pp. 32–39. DOI: https://doi.org/10.1093/jamia/ocx084.

[3] Reyna, Matthew A. et al. "Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019". In: *Critical care medicine* 48.2 (2019), pp. 210–217. DOI: https://doi.org/10.1097/CCM.0000000000004145.

[4] Bennett, Nicolas, Plecko, Drago, and Ukor, Ida-Fong. *Ricu: intensive care unit data with r*. https://github.com/eth-mds/ricu, 2023.