Variability in Automated Sepsis Case Detection: A Systematic Analysis of Implementation Methods in Clinical Data Repositories

Falk Meyer-Eschenbach[1,2,5,✉], **René Schmiedler**[1], **Johannes v. Stoephasius**[5], **Camille Zhang**[2], **Luis Kronfli**[1], **Nicolas Frey**[1], **Anatol-Fiete Näher**[4], **Jacques Ehret**[2], **Johanna Nothacker**[2], **Christof von Kalle**[2], **Severin Kohler**[6,7], **Elias Grünewald**[1], **Andreas Edel**[3], **Oliver Kumpf**[3], **Jon Barrenetxea**[1], **Felix Balzer**[1]

[1]Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Germany
[2]Clinical Study Center, Berlin Institute of Health, Berlin, Germany
[3]Klinik für Anästhesiologie und Intensivmedizin (CCM|CVK), Charité – Universitätsmedizin Berlin, Germany
[4]Digital Global Public Health, Hasso Plattner Institute, University of Potsdam, Potsdam, Germany
[5]Institute of Informatics, Humboldt University, Berlin, Germany
[6]Institute of Informatics, Freie Universität Berlin, Berlin, Germany
[7]Health Data Science Unit, University Hospital Heidelberg, Heidelberg, Germany

Abstract

Objective: To systematically identify and characterize methodological heterogeneity in sepsis case detection methods using the MIMIC-III database or the eICU-CRD, and to quantify the resulting variability in sepsis detection rates.

Materials and Methods: We conducted a PRISMA-guided systematic review of PubMed and Web of Science (2016–2024), and stratified studies by cohort definition to obtain comparable subsets. We extracted information on sepsis case detection methodology across six domains: parameter coverage, temporal windows, aggregation methods, missing data handling, SOFA calculation, and infection identification. For studies with available source code, we performed detailed code analysis extracting implementation decisions across all six domains.

Results: Of 396 screened publications using the MIMIC-III or the eICU-CRD database, 64 studies met our inclusion criteria. Reported sepsis detection rates ranged from 3.4 % to 65.2 %. Documentation of sepsis case definition methodology was provided in a minority of studies across all domains, including SOFA calculation (53.1 %), infection detection (42.2 %), temporal windows (37.5 %), aggregation methods (26.6 %), and missing data handling (17.2 %). Source code analysis revealed differences in baseline SOFA definitions, temporal alignment, and infection detection timing.

Discussion: The detection of sepsis cases is a key step in critical care research for validating sepsis prediction models. However, implementation of the current sepsis definition varies substantially, thereby undermining research reproducibility.

Conclusion: To improve the reproducibility and robustness of sepsis prediction models, we recommend standardized reporting of sepsis case definition methodology and the publication of version-controlled source code.

Keywords: sepsis; MIMIC-III; eICU-CRD; reproducibility; machine learning; electronic health records; SOFA score

✉ **For correspondence:** falk.meyer-eschenbach@charite.de (FME)

Data availability: All data supporting the findings of this study are available in a Zenodo repository (https://doi.org/10.5281/zenodo.18612351).

Competing interests: The authors declare no competing interests.

Introduction

Sepsis is a major clinical and economic burden globally, accounting for 11 million deaths in 2017 and representing 19 % of global mortality [1, 2]. In intensive care settings, sepsis affects 29.5 % of patients with mortality rates exceeding 35 % [3, 4], which results in substantial healthcare costs [5, 6]. The actual burden is likely higher due to underdiagnosis and inconsistent reporting, highlighting the need for standardized diagnostic criteria to improve timely case detection and epidemiological accuracy [7]. Recent clinical guidelines emphasize the importance of systematic screening for early sepsis detection [8], underscoring the need for reliable automated case detection methods. In this regard, the introduction of a new Sepsis definition in 2016 (Sepsis-3) marked an important step toward standardization of sepsis diagnosis [9]. This definition operationalizes sepsis as suspected infection accompanied by acute organ dysfunction. Organ dysfunction is quantified with the SOFA (Sequential Organ Failure Assessment) score, and an increase of at least 2 SOFA points indicates sepsis when a concomitant infection is apparent. Nevertheless, applying this definition to identical datasets yields substantially different detection rates across studies [10]. Such heterogeneity in case detection directly affects the validity of potential downstream applications like clinical decision support systems. In addition, predictive models trained on differently labeled cohorts may achieve similar performance metrics while predicting different clinical entities. This undermines meaningful comparisons of model validity, generalizability, clinical utility, and reproducibility. Limited transparency in detection details prevents independent validation and translation of research findings into clinical applications [11].

Algorithmic implementations of automated sepsis case detection are complex. As part of the most recent sepsis definition, the use of the SOFA score requires the integration of six organ system assessments (respiratory, cardiovascular, hepatic, coagulation, renal, and neurological). Each is dependent on clinical parameters recorded in electronic health records. These parameters may be recorded under multiple synonymous labels, measured at irregular intervals, and are subject to data quality issues including missing values, outliers, and documentation inconsistencies. Further, automated detection of infections compounds this complexity by requiring either diagnoses (e.g. documented with an International Classification of Disease (ICD) Code) or temporal matching of antibiotic administration with microbiology culture sampling according to criteria proposed by Seymour et al. [12]. This matching process involves multiple calculation decisions regarding temporal windows, antibiotic classification, and specimen type requirements.

The scale of methodological heterogeneity is evident when examining studies applying the current sepsis definition to the Medical Information Mart for Intensive Care III (MIMIC-III [13]) dataset, one of the most widely used publicly available critical care databases. Despite utilizing identical source data and nominally following the same definition, reported detection rates range from 11.3 % to 49.1 % [14]. Recent systematic analysis by Lambden et al. [10] highlighted challenges in sepsis operationalization,

yet a systematic characterization of the specific implementation decisions contributing to this variability is lacking.

Two methodological gaps contribute to these challenges. First, publications frequently omit sepsis case definition details – such as parameter identification strategies, temporal aggregation methods, missing value imputation approaches, SOFA calculation variants and definition of suspicion of infection – making exact reproduction difficult. Second, without consensus on best practices for retrospective sepsis operationalization, researchers make divergent methodological choices across multiple dimensions of data preprocessing and label calculation, each contributing to the observed detection rate variability.

This study addresses these gaps through a systematic, multi-level investigation of sepsis case detection methodologies in MIMIC-III [13] and eICU-CRD [15] research. We combine PRISMA-guided systematic literature review with detailed source code analysis to answer three research questions:

1. What is the range of sepsis detection rates reported across studies applying Sepsis-3 to MIMIC-III/eICU-CRD?
2. Which methodological decisions in data preprocessing and label calculation can be identified from publications and source code?
3. What patterns of methodological heterogeneity emerge across implementations, and to what extent are implementation decisions documented?

To structure this analysis, we decomposed detection decisions into six methodological domains (D1–D6): parameter coverage, temporal windows, aggregation strategies, missing value handling, SOFA calculation variants, and infection detection.

Materials and Methods

We first identified publicly available intensive care databases (MIMIC-III [13], MIMIC-IV [16], eICU-CRD [15], AmsterdamUMCdb [17]). For each database, we conducted literature searches in PubMed and Web of Science to identify related publications. Separately, we compiled a corpus of publications related to the current sepsis definition. We then computed the intersection of each database-specific set with the sepsis corpus to quantify the volume of sepsis detection studies per database. MIMIC-III yielded the largest intersection and was therefore selected for detailed analysis. We also included eICU-CRD as the second dataset for three reasons: (1) sufficient publication volume to enable systematic analysis (110 sepsis publications), (2) healthcare system comparability with MIMIC-III (both US-based), controlling for system-level confounders in documentation practices and clinical workflows, and (3) institutional independence from MIMIC-III, ensuring no patient data overlap. MIMIC-IV was excluded despite having the second-largest corpus (264 publications) because it originates from the same institution as MIMIC-III

with overlapping patient populations (2008–2012), precluding independent validation. AmsterdamUMCdb was excluded due to insufficient publication volume (17 sepsis publications) for robust systematic analysis.

The Medical Information Mart for Intensive Care III (MIMIC-III, version 1.4) contains de-identified electronic health record data from 46,520 patients admitted to intensive care units at Beth Israel Deaconess Medical Center between 2001 and 2012, representing 61,532 ICU stays across 58,976 hospital admissions [13]. MIMIC-III integrates data from two distinct clinical information systems: *CareVue* (primarily 2001–2008) and *MetaVision* (2008–2012), which differ in parameter naming conventions, documentation granularity, and data completeness; notably, microbiology records are underrepresented in *CareVue* data [18]. The eICU Collaborative Research Database (eICU-CRD, version 2.0) contains de-identified electronic health record data for patients admitted to intensive care units across the United States during 2014–2015. The database includes approximately 200,000 ICU stays from over 130,000 patients treated at more than 200 hospitals participating in the Philips eICU telehealth program [15]. Data are organized into over 30 interrelated tables encompassing high-frequency vital sign measurements, laboratory results, medication administrations, care plans, diagnoses, and severity-of-illness scores (including APACHE IV).

We searched PubMed and Web of Science from January 1, 2016 through December 31, 2024, covering the period since the Sepsis-3 definition was published. Search terms for MIMIC-III and Sepsis-3 comprised spelling variations (e.g., “mimic-iii”, “mimic iii”, “mimic-3”, “sepsis-3”, “sepsis 3”), as well as citation-based searches of the original Sepsis-3 definition [9] or MIMIC-III [13] publications. Complete search strategies with counts of resulting publications are provided in Supplement S1. Studies were included if they: (1) were published in English between January 2016 and December 2024, (2) were available as open access full text, (3) used the MIMIC-III or eICU dataset with explicit SOFA-based sepsis detection, and (4) reported a traceable sepsis detection rate with sufficient context to assess the underlying methodology. We excluded review articles, conference abstracts without full papers, studies mentioning but not using MIMIC-III or the current sepsis definition (Sepsis-3), and those without any traceable sepsis cohort derivation. The selection process followed PRISMA guidelines through three phases: initial title/abstract screening, full-text assessment for eligibility, and final inclusion. For MIMIC-III, screening was performed by three independent reviewers, with discrepancies resolved through discussion. For eICU-CRD, the same inclusion and exclusion criteria were applied, publications were reviewed by one reviewer and independently validated by a second reviewer. Application of the current sepsis definition was confirmed through operationalization of SOFA score and suspected infection.

To detect suspected infection from clinical data retrospectively, it was investigated whether implementations adhered to ICD or other methods, like the method proposed by Seymour et al. [12]. This method defines two temporal patterns. If sample collection occurred first, antibiotics must have been administered within 72 hours. If antibiotics were administered first, sample collection must have oc-

curred within 24 hours. The time of suspected infection is determined by whichever event occurred first.

We developed a data extraction protocol covering six methodological domains (D1–D6) that span the complete process from raw clinical data to final sepsis case detection. Parameter coverage (D1) encompasses clinical parameter selection and name variants (e.g., "Bilirubin", "Total Bilirubin", "Bili T."), various mechanical ventilation parameters, antibiotic names/routes for infection detection, and dual information system handling (*CareVue*/*MetaVision*). Temporal resolution (D2) covers SOFA calculation scope, sliding window size/offset, and temporal reference points (ICU admission vs. infection onset). Parameter aggregation (D3) includes strategies for multiple measurements within windows (worst-case vs. central tendency), cumulative parameters (rolling sums for urine output), and vasopressor aggregation. Missing data handling (D4) encompasses imputation methodologies, fallback strategies, formula-based calculations, and patient-specific vs. population-based approaches. SOFA calculation variants (D5) include baseline assumptions, component-specific modifications (weight normalization for vasopressors), and specialized calculations (SpO_2-based Horowitz Index). Infection detection (D6) covers primary approaches (ICD-9 codes, antibiotic-culture matching [12], hybrid methods), diagnostic codes, temporal windows for antibiotic-culture relationships, and antibiotic list comprehensiveness.

In addition, we documented cohort selection criteria including age restrictions, ICU stay duration requirements, exclusion thresholds for patients with excessive data missingness, multiple admission handling policies, and early sepsis exclusion strategies. For each study, we extracted MIMIC-III version, reported cohort sizes and sepsis detection rates (as patients, admissions, or ICU stays), and technical implementation details. Further, in order to facilitate interpretation of detection rate variations, we stratified studies based on cohort comparability: Set A Studies reported adult sepsis detection rates on standard MIMIC-III reference populations, without any other additional filters beyond age restrictions; Set B Studies applied additional filters except comorbidities (e.g., missingness thresholds, outlier removal, minimum observation duration). For studies investigating specific comorbidities or organ dysfunction subgroups (e.g., sepsis-associated AKI), we verified whether an intermediate processing step reported an adult sepsis cohort prior to applying disease-specific filters, enabling extraction of comparable baseline detection rates.

We systematically identified source code through two channels: code repository links provided in publications identified through our systematic search, and supplementary material of included publications. In addition, dependency tracing revealed upstream repositories containing complete Sepsis-3 implementations not captured by the literature search — notably when publications did not reference the current sepsis definition despite implementing it in their preprocessing code, illustrating the very documentation gaps this study characterizes. We selected code repositories meeting three additional criteria: (1) peer-reviewed publication association with studies identified in our systematic literature search, (2) explicit MIMIC-III/eICU-CRD and SOFA-based sepsis detection implementation veri-

fied through source code inspection, and (3) sufficient code completeness for methodology extraction. When re-implementations were identified, we traced code sources back to original implementations. For each identified code repository, we applied the same six-dimensional methodological framework (D1 to D6) to extract implementation details directly from source code. Implementation details were systematically extracted using a structured three-phase approach. First, each repository was profiled to document meta-information including URL and commit hash (for reproducibility), programming languages, file structure, and dependencies. Second, code files were screened to identify those affecting sepsis label calculation; identified relevant files were then analyzed line-by-line to extract implementation decisions across D1–D6 domains, with findings documented in structured tables and parameter mappings stored in JSON format. Third, cross-repository synthesis compared implementations to identify sources of variability.

All analyzed repositories were publicly accessible. Complete documentation is provided in Supplement S2, with a summary in the Results Section. Claude Code [19] (Anthropic Claude Opus 4.5, accessed January 2026) was used to assist with (1) structuring and organizing findings from source code analysis, and (2) generating code for data management and figure preparation. All AI-generated outputs were manually verified and validated by the authors.

Results

Systematic Literature Search

For MIMIC-III, the literature search identified 262 records from Web of Science and 115 from PubMed at the intersection of MIMIC-III and sepsis related publications, supplemented by 8 records from reference screening of seminal articles. After duplicate removal, 286 unique papers remained for screening. For eICU-CRD, a parallel search yielded 110 unique papers after duplicate removal. In total, 396 papers were screened across both databases (M: 286; E: 110), including 32 articles appearing in both searches (Figure 1).

During the screening phase, articles were sequentially excluded based on four criteria: accessibility (records without open access, preprint-only status, or non-English language), dataset used (non-MIMIC-III/eICU-CRD studies and review papers), adherence to current sepsis criteria (studies not implementing SOFA-based sepsis detection or with unrelated subjects), and provision of label calculation details. This yielded 118 articles for eligibility assessment (MIMIC-III: 80; eICU-CRD: 38). In the eligibility phase, articles were further evaluated for traceability of reported sepsis detection rates and comparability of exclusion criteria (excluding studies with hidden conditional filtering or incomparable sepsis labels). The final corpus comprised 64 studies for detailed analysis (MIMIC-III: 44; eICU-CRD: 20).

Among included MIMIC-III studies, primary cohort sources varied: 17 studies used the patients table,

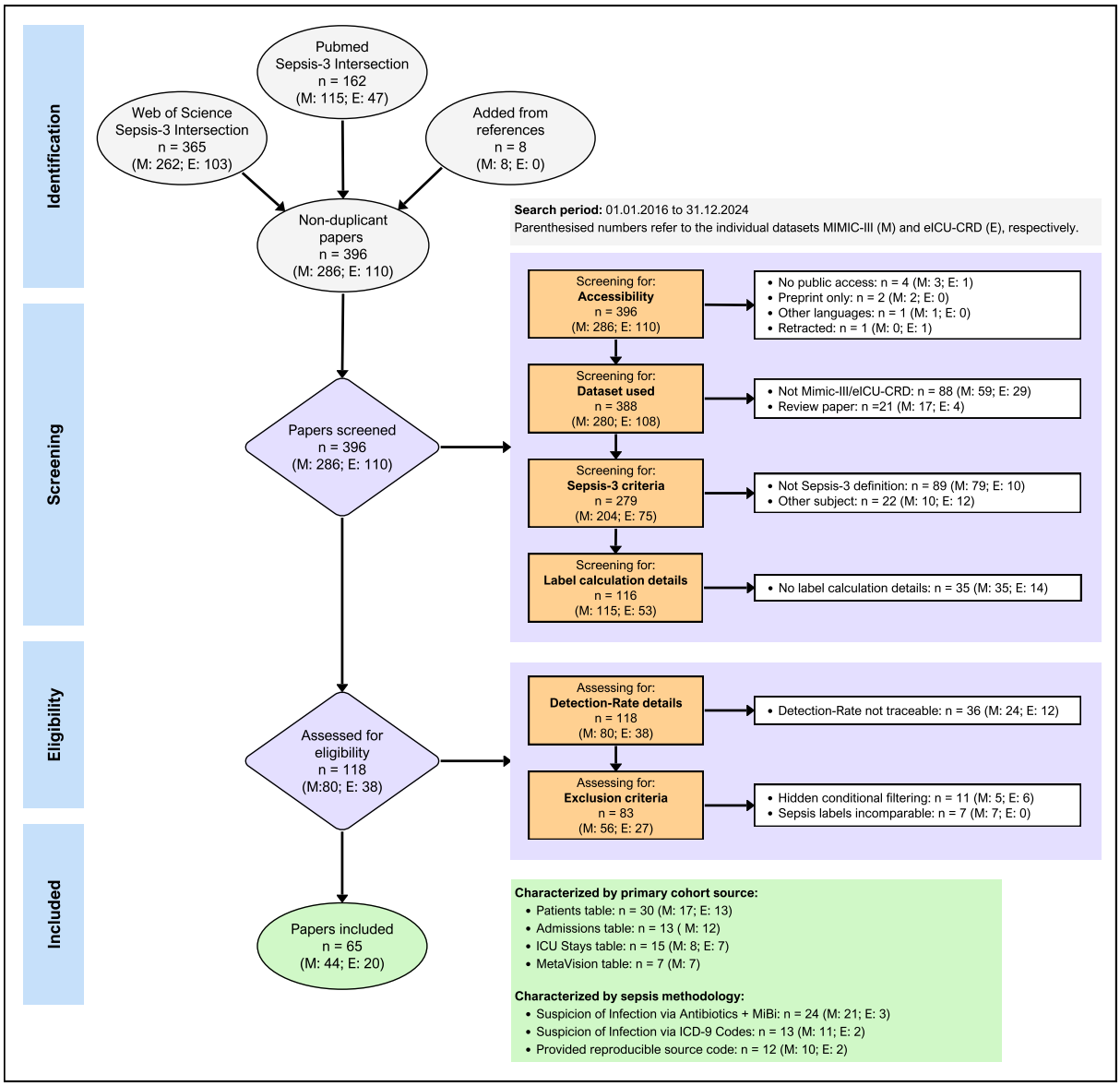

Figure 1. PRISMA flow diagram of the study selection process for MIMIC-III (M) and eICU-CRD (E) datasets. The diagram illustrates the phases of identification, screening, eligibility assessment, and final inclusion. After duplicate removal, 396 papers were screened (M: 286; E: 110). Articles were sequentially excluded based on accessibility, dataset used, adherence to current sepsis criteria, and provision of label calculation details, yielding 118 articles for eligibility assessment (M: 80; E: 38). Following evaluation of detection rate traceability and exclusion criteria comparability, 64 studies were included for detailed analysis (M: 44; E: 20).

---

12 the admissions table, 8 the ICU stays table, and 7 restricted analyses to *MetaVision* data. Regarding infection detection, 21 studies applied the antibiotic-culture temporal matching criteria [12], while 11 relied on ICD-9 codes. Among eICU-CRD studies, 13 used the patients table, 7 the ICU stays table; 3 studies applied antibiotic-culture matching, while 2 relied on ICD-9 codes and 7 did not report their infection detection method.

Sepsis Detection Rate Variability Across Studies

Table 1 presents detection rates, cohort sizes, and key methodological characteristics for all 44 MIMIC-III studies. To facilitate interpretation, we stratified studies into Set A (No. 1–24, n=24), comprising studies analyzing adult sepsis cohorts on standard MIMIC-III reference populations without additional data quality filters beyond age restrictions, and Set B (No. 25–44, n=20), comprising studies applying additional filters such as missingness thresholds, outlier removal, or minimum observation duration requirements.

Table 1. Sepsis detection rates across 44 MIMIC-III/Sepsis-3 studies. Set A (No. 1–24): standard reference populations without additional data quality filters; Set B (No. 25–44): additional filtering applied. Entity indicates cohort source table; # all = total cohort size; # septic = sepsis cases detected. Values in brackets were inferred from methodology or cited references. Abbreviations: ATB+Lab = antibiotics plus microbiology cultures [12]; ICD9 = ICD-9 diagnosis codes; Angus = Angus criteria; NR = not reported. Reference sizes: Patients n=46,520; Admissions n=58,976; ICU Stays n=61,532. Code: ✓ = available ([1]original, [2]derived, [3]MIMIC repository [20]).

No.	Author	Code	Version	Entity	Min. Age	Infection	# all	# septic	Rate
Set A – Standard reference populations									
1	Wang et al. [21]		1.4	Patients	16	ICD9+Lab	[38,641]	23,956	62.00 %
2	Parbhoo et al. [22]	✓	1.4	[Patients]	15	NR	[38,605]	18,200	47.14 %
3	Peng et al. [23]		1.4	Patients	–	SIC score	[46,520]	19,613	42.16 %
4	Chen et al. [24]		1.4	Patients	–	ICD9	46,520	18,592	39.97 %
5	Yao et al. [25]		[1.4]	Patients	16	ICD9	38,641	15,032	38.90 %
6	Chen et al. II [26]		1.4	Patients	15	NR	38,605	14,861	38.50 %
7	Drudi et al. [27]		[1.4]	[Admissions]	16	ATB+Lab	53,423	20,496	38.37 %
8	Huang et al. [28]		1.4	[Admissions]	16	NR	53,423	20,438	38.26 %
9	Chen et al. III [29]		[1.4]	[Admissions]	16	ICD9	53,423	18,592	34.80 %
10	Choi et al. [30]		1.4	[ICU Stays]	–	NR	61,532	20,927	34.01 %
11	Chen et al. IV [31]		1.4	[ICU Stays]	14	ATB+Lab	52,961	13,828	26.11 %
12	Wang et al. [32]		[1.4]	[Patients]	–	ATB+Lab	[46,520]	11,791	25.35 %
13	Shahn et al. [33]		[1.4]	MetaVision	18	ATB+Lab	23,606	5,784	24.50 %
14	Zhang et al. [34]		[1.4]	MetaVision	–	NR	23,630	5,443	23.03 %
15	Hou et al. [35]		1.4	Patients	–	ICD9	46,520	10,704	23.01 %
16	Ren et al. [36]		1.4	[Patients]	–	ICD9	[46,520]	7,851	16.88 %
17	Rosnati et al. [37]	✓	[1.4]	Admissions	15	ATB+Lab	NR	7,936	[14.86] %
18	Zhou et al. [38]		[1.4]	[Admissions]	18	ATB+Lab	[52,862]	7,230	13.68 %
19	Peng et al. [39]		[1.4]	[Admissions]	18	ICD9	[52,862]	5,784	10.94 %
20	Yan et al. [40]		[1.4]	Admissions	16	NR	52,963	4,727	8.93 %
21	Apalak & Kiasaleh [41]		1.4	MetaVision	15	ATB+Lab	19,515	1,725	8.84 %
22	Lan et al. [42]		1.4	Patients	16	ICD9	[38,641]	1,891	4.89 %
23	Wang et al. II [43]		1.4	[ICU Stays]	–	NR	61,532	2,733	4.44 %
24	Sakib et al. [44]		1.4	[Admissions]	–	ATB+Lab	58,976	1,994	3.38 %

Continued on next page

Table 1 continued from previous page

No.	Author	Code	Version	Entity	Min. Age	Infection	# all	# septic	Rate
Set B – Additional filtering applied									
25	Lin et al. [45]		1.3	Patients	18	NR	18,894	12,321	65.21 %
26	Zhong et al. [46]		[1.4]	[Patients]	–	[ICD9]	38,597	20,318	52.64 %
27	Wang et al. III [47]		[1.4]	[ICU Stays]	16	ICD-9 (Angus)	31,951	16,131	50.49 %
28	Johnson et al. [48]	✓	1.4	MetaVision	18	ATB+Lab	11,791	5,784	49.05 %
29	Wang et al. [49]		1.4	MetaVision	16	ATB+Lab	13,783	5,783	41.96 %
30	Liu et al. [50]		1.4	[Patients]	18	ATB+Lab	38,418	15,111	39.33 %
31	Ghalati et al. [51]		1.4	Admissions	–	ATB+Lab	58,976	22,547	38.23 %
32	Rhodes et al. [52]	✓[2]	[1.4]	[ICU Stays]	[18]	[ATB+Lab]	NR	20,954	[34.05] %
33	Rhodes et al. [53]	✓[2]	1.4	[ICU Stays]	[18]	[ATB+Lab]	NR	20,952	[34.05] %
34	Liu et al. [54]	✓	[1.4]	[Patients]	18	ATB+Lab	38,645	10,840	28.05 %
35	Komorowski et al. [55]	✓[1]	[1.4]	[ICU Stays]	18	ATB+Lab	NR	17,083	[27.76] %
36	Chen et al. [56]	✓	1.4	[Patients]	18	NR	NR	10,305	[26.67] %
37	Moor et al. [57]	✓	1.4	Admissions	15	ATB+Lab	36,591	9,541	26.07 %
38	He & Qiu [58]	✓[3]	1.4	[Admissions]	18	ATB+Lab	NR	7,924	[14.67] %
39	Wei et al. [59]		1.4	[Patients]	18	[ICD9]	38,645	5,325	13.78 %
40	Peng et al. [60]		1.4	Admissions	18	NR	43,129	5,784	13.41 %
41	Desautels et al. [18]		1.3	MetaVision	15	ATB+Lab	22,853	2,577	11.28 %
42	Li et al. [61]		1.3	Patients	18	NR	46,476	5,121	11.02 %
43	Solís-García et al. [62]		1.4	MetaVision	15	ATB+Lab	19,073	1,797	9.42 %
44	Barton et al. [63]		1.3	[ICU Stays]	18	ATB+Lab	21,507	1,024	4.76 %

To assess generalizability beyond MIMIC-III, we identified 20 additional publications implementing the current sepsis definition on the eICU-CRD (Table 2), which aggregates data from 208 US hospitals. Across all 44 MIMIC-III studies, reported sepsis detection rates ranged from 3.4 % to 65.2 % (median: 26.4 %, IQR: 13.6 %–38.6 %). Set A studies reported rates from 3.4 % to 62.0 % (median: 24.9 %, IQR: 13.0 %–38.4 %), while Set B studies showed comparable variability from 4.8 % to 65.2 % (median: 27.9 %, IQR: 13.7 %–40.0 %). Within MIMIC-III, the similar median detection rates between Set A and Set B (difference: 3.0 percentage points) suggest that cohort restriction strategies alone do not account for the observed variability.

Table 2. Sepsis detection rates across 20 eICU-CRD/sepsis studies. Entity indicates cohort source table; # all = total cohort size; # septic = sepsis cases detected. Values in brackets were inferred from methodology or cited references. Reference sizes: Patients n=139,367; ICU Stays n=200,859. Abbreviations: ATB+Lab = antibiotics plus microbiology cultures [12]; ICD9 = ICD-9 diagnosis codes; NR = not reported; doc = documented; sus = suspected; con = confirmed; kno = known; inf = infection. Code: ✓ = available ([1]original, [2]derived, [3]MIT-LCP repository). [4]No further infection information is provided. [5]Total cohort size including patients under the min. age.

No.	Author	Code	Version	Entity	Min. Age	Infection	# all	# septic	Rate
1	Fang et al. [64]		[2.0]	[ICU Stays]	[18]	–	[200,859][5]	96,127	47.86 %
2	Zhang et al. [65]		2.0	Patients	18	doc./sus. inf.[4]	[139,367][5]	48,694	34.94 %
3	Zhang et al. [66]		2.0	Patients	18	con./sus. inf.[4]	[139,367][5]	48,694	34.94 %
4	Zhang et al. [67]		2.0	Patients	18	con. inf.[4]	[139,367][5]	48,694	34.94 %
5	Zhang et al. [68]	✓[3]	2.0	Patients	–	–	[139,367]	43,727	31.38 %
6	Yang et al. [69]		2.0	[ICU Stays]	18	–	200,859[5]	55,848	27.80 %
7	Zhang et al. [70]		–	Patients	–	inf.[4]	[139,367]	36,302	26.05 %
8	Wang et al. [71]	✓[3]	–	Patients	–	inf.[4]	[139,367]	36,046	25.86 %
9	Li et al. [72]		2.0	Patients	–	ATB+Lab	[139,367]	31,737	22.77 %
10	Zhao et al. [73]		–	ICU Stays	–	ATB+Lab	200,859	41,667	20.74 %
11	Lin et al. [74]		[2.0]	Patients	–	con. inf.[4]	[139,367]	23,136	16.60 %
12	Lin et al. [75]		[2.0]	Patients	–	ICD-9	[139,367]	23,136	16.60 %
13	Shen et al. [76]		2.0	Patients	–	–	[139,367]	23,136	16.60 %
14	Chang et al. [77]		–	Patients	–	ICD-9	[139,367]	23,136	16.60 %
15	Yang et al. [78]	✓[1]	–	Patients	–	kno./sus. inf.[4]	[139,367]	22,864	16.41 %
16	Kasugai et al. [79]		–	[ICU Stays]	–	doc./sus. inf.[4]	200,859	30,114	14.99 %
17	Wang et al. [80]	✓[2]	2.0	ICU Stays	–	–	200,859	28,667	14.27 %
18	Li et al. [81]		2.0	Patients	–	–	[139,367]	19,430	13.94 %
19	Sheng et al. [82]		2.0	[ICU Stays]	[18]	–	200,859[5]	26,360	13.12 %
20	Takkavatakarn et al. [83]	✓[3]	[2.0]	ICU Stays	[18]	ATB+Lab	[200,859][5]	19,575	9.75 %

The 20 eICU-CRD studies showed rates ranging from 9.8 % to 47.9 % (median: 18.7 %, IQR: 16.1 %–28.7 %). Comparing the two databases, MIMIC-III studies showed higher median detection rates (26.4 % vs. 18.7 %) and wider variability (range: 61.8 vs. 38.1 percentage points; IQR: 25.0 vs. 12.6 percentage points). The higher median in MIMIC-III may reflect institutional characteristics of a single academic medical center, while the wider variability likely results from MIMIC-III's longer availability and more extensive use in sepsis research (n = 44 vs. n = 20 studies), providing more opportunity for methodological divergence.

Variability persisted even among studies analyzing identical source populations within each database. In MIMIC-III, five studies deriving cohorts from the patients table without age restrictions (n=46,520) reported detection rates ranging from 16.9 % to 42.2 %, while three studies using patients aged ≥ 16 years (n=38,641) identified between 4.9 % and 62.0 %. Similarly, in eICU-CRD, ten studies analyzing the full patients table (n=139,367) without age restrictions reported detection rates from 13.9 % to 31.4 %. Since

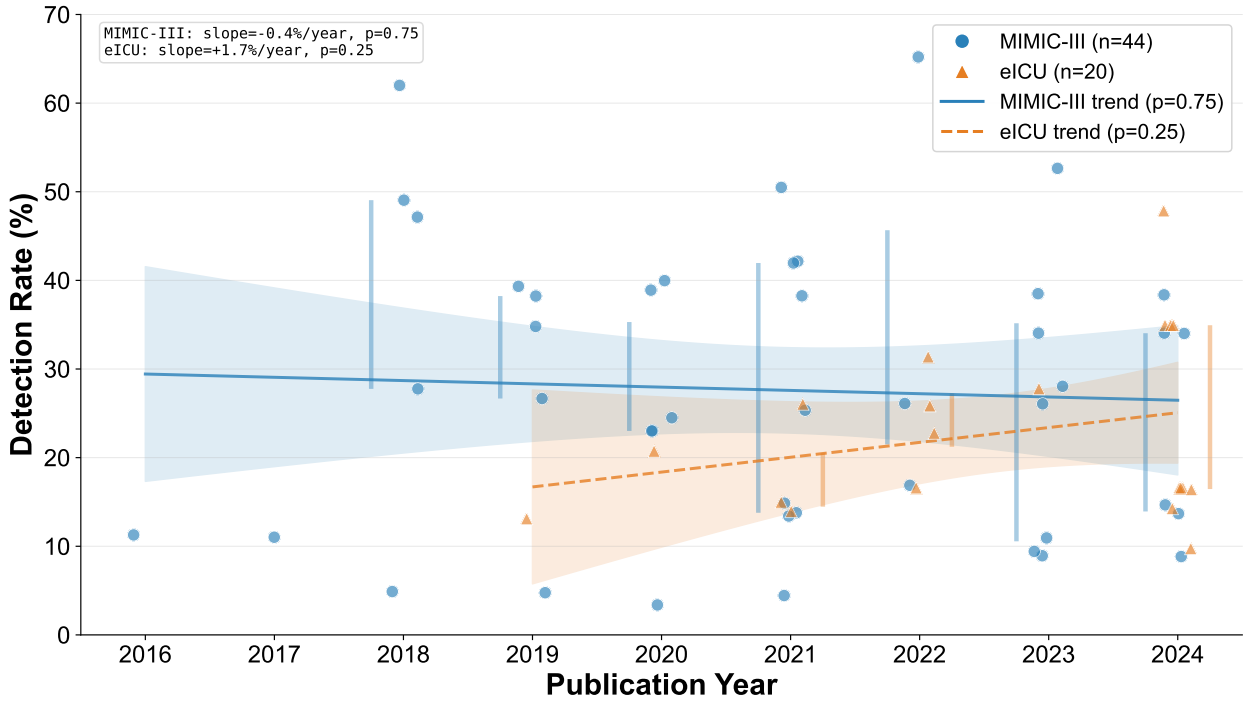

Figure 2. Detection rates by publication year across MIMIC-III (n=44, blue circles) and eICU (n=20, orange triangles) studies (2016–2024). Each point represents one study; points are jittered horizontally for visibility. Lines show linear regression trends with 95 % confidence bands (shaded areas). Vertical bars indicate interquartile ranges for years with ≥ 3 studies.

these studies operated on identical source data within each database, this within-population variability demonstrates that implementation decisions, rather than patient selection, contribute to detection rate differences. This variability was not limited to early implementations following publication of the current sepsis definition. Detection rates showed spread across all publication years from 2016 to 2024, with no trend toward convergence over time in either database (Figure 2; MIMIC-III: slope = –0.4 %/year, p = 0.75; eICU-CRD: slope = +1.7 %/year, p = 0.25).

Among MIMIC-III studies, ten (22.7 %) provided links to publicly accessible source code, with code availability notably higher among Set B studies (8/20, 40.0 %) compared to Set A (2/24, 8.3 %). Two eICU-CRD studies (10.0 %) provided source code, three additional studies were linked to an upstream MIT-LCP repository through dependency tracing (Table 2). Additionally, 14 MIMIC-III studies (31.8 %) did not explicitly report cohort sizes (6 not reported, 8 inferred from standard reference populations), and 15 (34.1 %) provided limited documentation of infection detection methods, limiting reproducibility assessment.

Methodological Heterogeneity Across Publications

Analysis of preprocessing methodology across 44 MIMIC-III/sepsis studies revealed documentation gaps and methodological heterogeneity across all five domains examined (D2–D6). Parameter coverage (D1) could not be reliably assessed from publication text alone and is addressed through source code analysis in the next section. Documentation rates, defined as explicit reporting of methods applicable to sepsis label calculation, were uniformly low: SOFA calculation parameters (34.1 %, n=15), infection detection (31.8 %, n=14), temporal windows (15.9 %, n=7), aggregation methods (11.4 %, n=5), and missing data handling (6.8 %, n=3). A significant proportion of studies described preprocessing steps where applicability to label calculation versus prediction model development remained ambiguous (D2: 15.9 %, D3: 20.5 %, D4: 47.7 %), complicating reproducibility assessment. Regarding temporal resolution (D2), among the 7 studies with explicit temporal documentation for label calculation, aggregation window sizes ranged from 1 hour to 24 hours, with 4-hour windows most common (n=3), followed by 24-hour (n=1) and 1-hour (n=1). Window step parameters (offset between consecutive calculations) were documented in 3 of these studies (1-hour steps: n=2; 24-hour: n=1). The combination of window size and step, critical for understanding temporal resolution, was rarely specified together. For aggregation methods (D3), among the 5 studies with explicit documentation, strategies divided between parameter-specific approaches (n=4), applying different methods to different measurement types (e.g., mean for vital signs, sum for urine output, or worst value per SOFA component), and uniform aggregation (n=1) applying mean to all parameters. For missing data handling (D4), only 3 studies explicitly documented methods for label calculation, all combining forward-fill with default-to-normal assumptions for missing SOFA components – a practice that may reduce SOFA scores and thus underestimate sepsis detection rates. An additional 21 studies (47.7 %) documented imputation methods where applicability to label calculation versus prediction model development remained unclear, limiting reproducibility assessment for this domain. Regarding SOFA calculation (D5), among the 15 studies with explicit documentation for label calculation, 12 stated the ≥ 2 threshold required by the current sepsis definition. However, baseline SOFA determination, critical for detecting the 2-point increase rather than absolute score, was documented in only 2 studies: one assumed zero baseline (healthy pre-ICU state), the other used admission SOFA. One study specified a 6-hour window for detecting the required change. For infection detection (D6), among the 14 studies with explicit documentation, the antibiotic-culture temporal matching criteria [12] predominated (n=11), with ICD-9 diagnostic codes used in 2 studies and one study requiring both methods simultaneously.

Figure 3 illustrates the relationship between documented preprocessing methods and sepsis detection rates across all domains. Substantial variability exists within each methodological category, suggesting that documented preprocessing choices alone do not explain the observed detection rate heterogeneity.

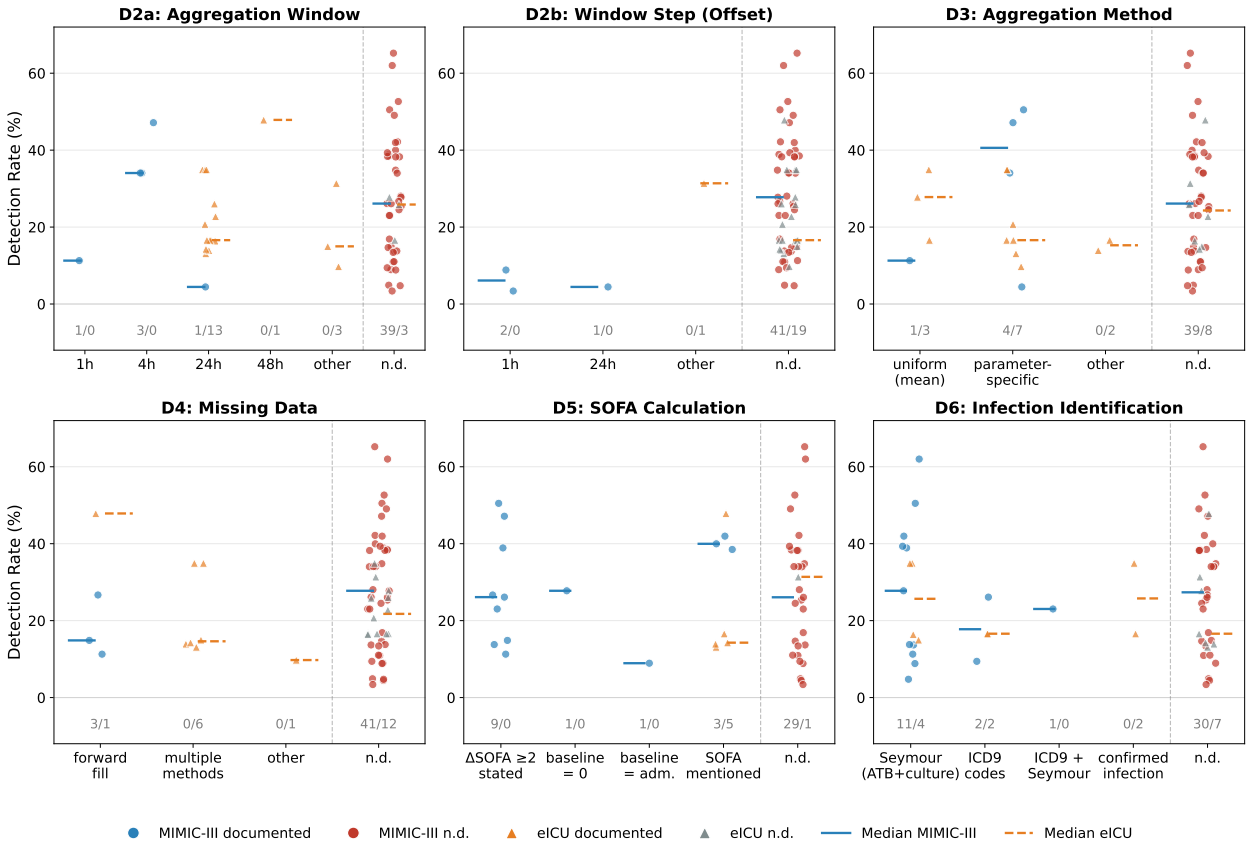

Figure 3. Detection rate variability across preprocessing methodology in MIMIC-III (n=44) and eICU-CRD (n=20) studies. Panels show temporal windows (D2a/b), aggregation methods (D3), missing data handling (D4), SOFA calculation (D5), and infection detection (D6). Blue circles: MIMIC-III documented; red circles: MIMIC-III not documented (n.d.); orange triangles: eICU-CRD documented; gray triangles: eICU-CRD n.d. Solid lines: MIMIC-III medians; dashed lines: eICU-CRD medians. Sample sizes shown as MIMIC-III/eICU-CRD above each category.

Documentation rates in eICU-CRD studies (n=20) were considerably higher: SOFA calculation (95.0 %, n=19), infection detection (65.0 %, n=13), temporal windows (85.0 %, n=17), aggregation methods (60.0 %, n=12), and missing data handling (40.0 %, n=8). Despite documentation, detection rate variability persisted (range: 38.1 percentage points), confirming that methodological heterogeneity rather than documentation alone drives outcome differences. Across both databases (n=64), combined documentation rates were: SOFA calculation (53.1 %), infection detection (42.2 %), temporal windows (37.5 %), aggregation methods (26.6 %), and missing data handling (17.2 %).

Source Code Analysis

Of the 64 shortlisted articles, 12 provided references to openly accessible source code, though several referenced identical repositories. After deduplication and screening for complete Sepsis-3 label calculation implementations, six repositories remained for detailed analysis: Johnson et al. [48] (sepsis3-

mimic; 18 files, 110 decisions), Komorowski et al. [55] (AI_Clinician; 3 files, 36 decisions), Johnson et al. [20] (mimic-code; 34 files, 50 decisions), and Rosnati et al. [37] (MIMIC-III-sepsis-3-labels; 25 files, 68 decisions) for MIMIC-III; Chen et al. [84] (16 files, 34 decisions) and Yang et al. [78] (1 file, 23 decisions) for eICU-CRD. Additional repositories with public code (e.g., Moor et al. [57]) were identified as re-implementations building on Johnson et al. [48] and were therefore traced back to the original implementation rather than analyzed separately, consistent with our stated methodology. Chen et al. [84] was not captured by the systematic literature search as the publication does not reference the current sepsis definition, but was identified through dependency tracing as the upstream source of Wang et al. [80], whose code derives from this repository. Each repository underwent systematic profiling through line-by-line source code review to extract implementation decisions across all domains (D1–D6), repository metadata, and dependency relationships. This analysis identified 264 MIMIC-III and 57 eICU-CRD implementation decisions within the domains D1–D6 across 80 and 17 source files respectively, revealing substantial methodological heterogeneity unreported in the associated publications. Complete technical documentation is provided in Supplement S2.

Implementation dependencies exist among MIMIC-III repositories: For instance, Komorowski et al. [55] builds upon Johnson et al. [48], which derives from MIT-LCP. This dependency structure suggests potential propagation of methodological decisions, whether intentional design choices or incidental defaults, across a substantial portion of published sepsis research. The two eICU-CRD repositories were independently developed with no direct code dependencies.

For parameter coverage (D1), analysis across MIMIC-III repositories identified 149 unique ITEMIDs with variation in cardiovascular monitoring (18–28 ITEMIDs): mimic-code (Johnson et al. [20]) provided the most comprehensive coverage of vasoactive substances (28 ITEMIDs, seven types of vasoactive substances including milrinone), while Rosnati et al. [37] used the fewest (18 ITEMIDs, four types of vasoactive substances – dopamine, norepinephrine, epinephrine, and dobutamine – excluding phenylephrine, vasopressin, and milrinone). Respiratory (74–82) and renal (25–27) parameters varied modestly, while coagulation, liver, and CNS components remained identical. Total parameter coverage ranged from 125 ITEMIDs (Komorowski et al. [55]) to 143 ITEMIDs (Johnson et al. [20]). In eICU-CRD, parameter identification ranged from 14 to 32 string identifiers, with Chen et al. [84] using dual data sources (infusiondrug and medication tables with HICL codes) and tracking 7 vasopressors, while Yang et al. [78] used simpler patterns covering 5 vasopressors with additional parameters from external pre-calculated tables whose methodology remains invisible.

Temporal window definitions (D2) varied across 5–16 implementation decisions per MIMIC-III repository. Rosnati et al. [37] calculated SOFA scores relative to ICU admission with hourly resolution throughout hospitalization, while Johnson et al. [48] and Komorowski et al. [55] calculated within infection-relative windows (48-hour retrospective, 24-hour prospective around suspected infection onset). This temporal alignment difference alters baseline SOFA determination and the population eligible for sep-

sis classification. Komorowski et al. [55] additionally employed 4-hour discretization blocks rather than hourly resolution. In eICU-CRD, Chen et al. [84] implemented 0–24h first-day windows while Yang et al. [78] applied 72-hour laboratory windows (−24h to +48h), with ICU LOS requirements differing 3-fold (≥ 72h vs. > 24h). Aggregation strategies (D3) required 1–6 implementation decisions per MIMIC-III repository. All repositories implemented worst-case aggregation within evaluation windows (minimum for platelets/GCS, maximum for creatinine/bilirubin), but differed in neurological assessment timing and urine output calculation methods. Rosnati et al. [37] applied 24-hour rolling sums for urine output, while other repositories used window-specific totals, affecting oliguria detection sensitivity. Yang et al. [78] imported aggregated values from external tables, precluding verification of aggregation methodology. Missing data strategies (D4) showed the highest decision complexity (8–16 decisions per MIMIC-III repository). Johnson et al. [48] and MIT-LCP employed zero-imputation as primary strategy, reducing SOFA scores when measurements were absent. In contrast, Komorowski et al. [55] implemented multi-stage imputation procedures combining linear interpolation, k-nearest neighbor imputation, and variable-specific defaults. Rosnati et al. [37] applied best case neurological imputation (missing GCS assigned normal values). In eICU-CRD, Chen et al. [84] used zero-imputation while Yang et al. [78] relied on external table values with undocumented imputation logic. For SOFA calculation (D5), baseline definition varied fundamentally: Johnson et al. [48], MIT-LCP, Komorowski et al. [55], and Chen et al. [84] defined baseline as zero (interpreting the definition as absolute SOFA ≥ 2), while Rosnati et al. [37] implemented dynamic baseline using the lowest observed SOFA, a choice that alters case inclusion for patients with pre-existing organ dysfunction. Yang et al. [78] imported pre-calculated sofa_total from an external eICU-CRD table, precluding verification of calculation methodology and baseline definition. For infection detection (D6), all four MIMIC-III repositories implemented antibiotic-culture temporal matching criteria [12], though Komorowski et al. [55] used asymmetric temporal windows deviating from the symmetric windows specified in the original criteria. Neither eICU-CRD repository implemented antibiotic-culture matching, both relied on APACHE related diagnosis codes, reflecting database constraints where multi-hospital structures provide less granular timing data than single-institution designs. Additionally, Yang et al. [78] contained an implementation error where base excess validation incorrectly converts severely negative values to positive (≤ −30 to +30).

The practical impact is illustrated by MIMIC-III *MetaVision* cohorts: Johnson et al. [48] reported 49.1 % detection rate, while overlapping implementations using the same source population reported 8.8 % – 42.0 % (Table 1). In eICU-CRD, differing temporal windows (24h vs. 72h), parameter coverage (32 vs. 14 identifiers), ICU-LOS requirements (≥ 72h vs. > 24h), and infection criteria (APACHE and ICD vs. APACHE only) would produce different cohorts from identical source data.

Discussion

This systematic review of 44 MIMIC-III and 20 eICU-CRD studies reveals that the current definition of sepsis, despite providing a standardized clinical framework, may lead to variability in detection rates. Detection rates ranged from 3.4 % to 65.2 % in MIMIC-III (median 26.4 %) and from 9.8 % to 47.9 % in eICU-CRD (median 18.7 %), reflecting a degree of heterogeneity that cannot be attributed to stochastic variation alone but instead indicates divergences in computational implementation decisions. Documentation rates were low across all domains examined, and source code analysis of six major repositories identified 321 specific implementation decisions that would produce different patient cohorts from identical data. Notably, this variability persists in current literature, underscoring the ongoing significance of these methodological challenges for research and machine learning applications. The observed variability stems from three interconnected mechanisms rather than random methodological choices:

First, publicly available ICU databases present data quality challenges requiring numerous undocumented decisions. MIMIC-III contains over 12,000 unique parameter names (e.g., > 100 names for ventilation parameters), physiologically implausible outliers (e.g., mean arterial pressure: 989999 mmHg), dual information systems with different documentation practices (*CareVue* 2001–2008, *MetaVision* 2008–2012), and de-identification artifacts affecting temporal analyses. These challenges are not database-specific; systematic comparison across five major ICU databases confirms that inconsistent diagnostic coding, variable temporal resolution, and heterogeneous parameter naming represent barriers to harmonization [85].

Second, missing data handling for SOFA components varies across implementations. Our source code analysis revealed diverse approaches including zero-imputation (assuming normal organ function), multi-stage imputation procedures combining interpolation with k-nearest neighbor methods, and variable-specific defaults. Zero-imputation, while computationally simple, may underestimate dysfunction in critically ill patients where missing measurements may indicate deterioration rather than normality. Infection detection similarly requires choosing between antibiotic-culture temporal matching [12] (temporal precision but limited to treated infections) and ICD-9 diagnostic codes (broader coverage but dependent on variable coding practices). These choices influence detection rates yet remain poorly documented.

Third, implementation choices may be shared across research groups. The similar median detection rates between Set A and Set B studies (24.9 % vs. 27.9 %) suggest that within-study methodological choices have greater impact than between-study patient population differences. As the following sections demonstrate, this replication occurs through both documented code repositories and professional data analytics services – that have not been described in this context before.

Our analysis further reveals patterns consistent with shared computational approaches across studies. In the eICU-CRD dataset, we observed two distinct clusters of studies reporting identical detection rates and cohort sizes, across research groups.

Cluster 1 (34.94 % detection rate): Three publications by Zhang and colleagues [65–67] report identical cohorts (48,694 septic cases from 139,367 patients) while investigating different research questions. Author affiliation analysis confirms all three publications originate from the same institution, consistent with shared data extraction procedures applied across studies.

Cluster 2 (16.60 % detection rate): Four publications—Lin et al. [74, 75], Shen et al. [76], and Chang et al. [77]—report identical cohorts (23,136 septic cases from 139,367 patients) across three geographically distinct institutions. Author analysis reveals a common co-author affiliated with a professional data analytics service across three of four publications [74, 75, 77]. This pattern suggests that shared data extraction and preprocessing services may represent a mechanism through which external biostatistical services replicate implementation decisions across independently operating research groups without shared code documentation.

This pattern of identical detection rates extending beyond single research groups indicates that implementation decisions are transmitted not only through documented code repositories (as shown for MIMIC-III where Komorowski et al. [55] builds upon Johnson et al. [48], which derives from mimic-code, initiated by Johnson et al. [20]), but also through professional data analysis services and adaptation of methodological choices from influential publications. While reuse of established computational methods can accelerate research, undocumented dependencies may lead to systematic propagation of specific analytic decisions. Where the foundational implementation contains methodologically specific decisions (e.g., zero-imputation of missing values, particular temporal windows), these choices may become *de facto* standards without independent evaluation, potentially resulting in methodologically homogeneous subgroups within a diverse literature and narrowing the methodological space explored by subsequent research.

Analysis of detection rates over time (Figure 2) reveals persistent heterogeneity without convergence toward methodological consensus. Despite nearly a decade of research following publication of the current sepsis definition, detection rates show no trend toward standardization in either database (MIMIC-III: slope = –0.4 %/year, $p = 0.75$; eICU-CRD: slope = +1.7 %/year, $p = 0.25$), with the full range of variability (3.4 %–65.2 % in MIMIC-III, 9.8 %–47.9 % in eICU-CRD) persisting across all publication years. This suggests that awareness of implementation heterogeneity alone does not drive methodological consensus, and that the mechanisms of methodological transmission documented above—shared computational dependencies and professional data analytics services—continue to operate despite growing recognition of reproducibility challenges. These findings have implications for future sepsis definitions.

The recent introduction of SOFA-2 [86], with additional parameters (SpO_2/FiO_2 ratio, high-flow nasal

cannula, continuous positive airway pressure, bilevel positive airway pressure, noninvasive ventilation, ECMO, expanded vasopressor definitions, mechanical cardiovascular support, drugs for delirium treatment) and updated scoring thresholds, will likely increase the implementation heterogeneity documented here. Our analysis demonstrates that even the current SOFA framework generates detection rate variations spanning 3.4 % to 65.2 %. Additional complexity of SOFA-2 introduces new decision points across all six methodological domains (D1–D6), from defining "missing" versus "not clinically indicated" parameters to handling selective clinical application of contemporary variables and accommodating variable temporal resolution across monitoring devices. Retrospective application of SOFA-2 to existing ICU datasets spanning different time periods (MIMIC-III 2001–2012, MIMIC-IV 2008–2019 [16], eICU-CRD 2014–2015, AmsterdamUMCdb 2003–2016 [17]) may present challenges due to limited availability of newer parameters such as high-flow nasal cannula and mechanical cardiovascular support devices. This necessitates additional assumptions and imputation strategies that exceed the complexity we documented for current SOFA implementations. Without standardized approaches to handling absent parameters in historical data, researchers will likely implement divergent strategies, further fragmenting the literature. To prevent these challenges repeating, we recommend that SOFA-2 deployment include proactive standardization: reference implementations with comprehensive D1–D6 documentation released simultaneously with clinical definitions, standardized test datasets for validation, mandatory implementation reporting integrated into existing guidelines (TRIPOD [87], CONSORT [88]), and community-driven consensus on edge case handling before widespread adoption.

This study provides the first systematic multi-level analysis combining publication review with source code examination to characterize variability in sepsis case detection. The six-domain framework (D1–D6) offers a structured approach for assessing and reporting methodological decisions. Our analysis of 321 implementation decisions across six major repositories quantifies previously undocumented sources of heterogeneity. Unlike previous reviews documenting heterogeneity [10, 89, 90], our approach systematically characterizes the specific implementation decisions contributing to this variability.

Several limitations warrant consideration. First, our inclusion criteria restricted the corpus to open access publications, introducing a potential selection bias: studies behind paywalls may employ different implementation approaches or report different detection rates. However, this restriction was necessary to enable full-text extraction of methodological details, which are rarely available from abstracts alone. Second, our analysis focused on MIMIC-III and eICU-CRD; generalizability to other datasets (MIMIC-IV, AmsterdamUMCdb, institutional databases) requires further research. Implementation challenges may differ with dataset structure, documentation practices, and available parameters. Third, the low documentation rates in publications (> 65 % undocumented for most domains) precluded regression analyses quantifying individual decisions' contribution to detection rate variability. Fourth, source code analysis was limited to six repositories meeting inclusion criteria; additional repositories may em-

ploy different approaches. Fifth, we focused on methodological heterogeneity without clinical outcome validation; prospective evaluation using expert-adjudicated ground truth labels would be valuable to determine whether higher or lower detection rates better capture clinically meaningful sepsis.

The observed heterogeneity has potential implications for sepsis research reproducibility. The detection rate variation (3.4 %–65.2 %) among studies was rather substantial, especially when considering that identical data were used. We could demonstrate that algorithmic implementation details, rather than the conceptual Sepsis-3 framework, drive this variability. This challenges the assumption that standardized definitions lead to comparable results across implementations. Furthermore, the predominance of undocumented methodological choices (> 65 % for most domains) points to a documentation gap that limits validation, comparison, and clinical implementation of complex multi-stage data processing methods. These challenges are not unique to sepsis research. A systematic review of 152 machine learning-based clinical prediction models revealed comparable methodological patterns, with 82 % failing to report sample size calculations and 95 % omitting calibration metrics [11]. For machine learning applications, predictive algorithms trained on differently labeled cohorts may achieve comparable performance metrics while predicting distinct clinical phenotypes, compromising cross-study comparability and clinical translation.

The 321 implementation decision points identified provide a foundation for targeted standardization efforts. We recommend: (1) mandatory reporting of implementation details across D1–D6 domains as supplementary material, (2) publication of version-controlled source code, (3) development of consensus reference implementations for common databases, and (4) integration of sepsis label derivation standards into existing reporting guidelines, analogous to CONSORT for clinical trials and TRIPOD for prediction models. Without such standardization, the clinical utility of sepsis prediction models will remain limited by implementation variability rather than model performance, with direct implications for the development of reliable clinical decision support systems.

Conclusion

The methodological challenges identified here likely extend beyond cases defined on the Sepsis-3 framework to other disease phenotypes derived from EHR that require complex operationalization from clinical data [91]. Future research should address systematic experimentation using configurable data processing architectures that enable controlled variation of individual implementation decisions across D1–D6 domains to quantify their impact on detection rates, clinical outcome validation comparing implementations against expert-adjudicated ground truth and mortality outcomes, and extension of this framework to assess whether implementation heterogeneity represents a broader challenge in computational phenotyping. Ultimately, the path to reliable clinical decision support systems depends not on developing more sophisticated prediction models, but on establishing the foundational

infrastructure, standardized implementations, transparent documentation, and reproducible data processing workflows, necessary to ensure that these models predict clinically meaningful and consistent disease phenotypes across research contexts.

Data Availability. All data supporting the findings of this study are available in a Zenodo repository (https://doi.org/10.5281/zenodo.18612351). The repository contains: literature search exports from PubMed and Web of Science (RIS files), PRISMA screening records with multi-reviewer assessments, detection rate and methodology classification data underlying Tables 1–2 and Figures 2–3, source code analysis results including 321 implementation decisions across six repositories, ITEMID and parameter cross-repository mappings, and Python scripts for reproducing all figures. The six analyzed source code repositories are publicly accessible, URLs and commit hashes are documented in the repository. Raw screening spreadsheets are available from the corresponding author upon reasonable request.

References

[1] K. E. Rudd et al. "Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study". In: *Lancet Lond. Engl.* 395 (2020), pp. 200–211. DOI: 10.1016/S0140-6736(19)32989-7.

[2] UN, World Population Prospects (2024) – processed by Our World in Data. *Number of deaths, total – UN WPP [dataset]*. United Nations, "World Population Prospects" [original data]. 2024.

[3] Y. Sakr et al. "Sepsis in Intensive Care Unit Patients: Worldwide Data From the Intensive Care over Nations Audit". In: *Open Forum Infect. Dis.* 5 (2018), ofy313. DOI: 10.1093/ofid/ofy313.

[4] C. Fleischmann-Struzek, D. Schwarzkopf, and K. Reinhart. "Inzidenz der Sepsis in Deutschland und weltweit: Aktueller Wissensstand und Limitationen der Erhebung in Abrechnungsdaten". In: *Med. Klin. - Intensivmed. Notfallmedizin* 117 (2022), pp. 264–268. DOI: 10.1007/s00063-022-00906-2.

[5] C. Fleischmann et al. "Hospital Incidence and Mortality Rates of Sepsis". In: *Dtsch. Arzteblatt Int.* 113 (2016), pp. 159–166. DOI: 10.3238/arztebl.2016.0159.

[6] C. J. Paoli, M. A. Reynolds, M. Sinha, M. Gitlin, and E. Crouser. "Epidemiology and Costs of Sepsis in the United States—An Analysis Based on Timing of Diagnosis and Severity Level". In: *Crit. Care Med.* 46 (2018), pp. 1889–1897. DOI: 10.1097/CCM.0000000000003342.

[7] C. Rhee et al. "Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014". In: *Jama* 318 (2017), pp. 1241–1249. DOI: 10.1001/jama.2017.13836.

[8] F. M. Brunkhorst et al. "S3-Leitlinie Sepsis–Prävention, Diagnose, Therapie und Nachsorge–Update 2025". In: *Med. Klin.-Intensivmed. Notfallmedizin* (2025), pp. 1–69. DOI: 10.1007/s00063-024-01210-9.

[9] M. Singer et al. "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)". In: *JAMA* 315 (2016), pp. 801–810. DOI: 10.1001/jama.2016.0287.

[10] S. Lambden, P. F. Laterre, M. M. Levy, and B. Francois. "The SOFA score—development, utility and challenges of accurate assessment in clinical trials". In: *Crit. Care* 23 (2019), p. 374. DOI: 10.1186/s13054-019-2663-7.

[11] C. L. Andaur Navarro et al. "Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models". In: *J. Clin. Epidemiol.* 154 (2023), pp. 8–22. DOI: 10.1016/j.jclinepi.2022.11.015.

[12] C. W. Seymour et al. "Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)". In: *JAMA* 315 (2016), pp. 762–774. DOI: 10.1001/jama.2016.0288.

[13] A. E. W. Johnson et al. "MIMIC-III, a freely accessible critical care database". In: *Sci. Data* 3 (2016), p. 160035. DOI: 10.1038/sdata.2016.35.

[14] N. Bennett, D. Plečko, I.-F. Ukor, N. Meinshausen, and P. Bühlmann. "ricu: R's interface to intensive care data". In: *GigaScience* 12 (2022), giad041. DOI: 10.1093/gigascience/giad041.

[15] T. J. Pollard et al. "The eICU Collaborative Research Database, a freely available multi-center database for critical care research". In: *Sci. Data* 5 (2018), p. 180178. DOI: 10.1038/sdata.2018.178.

[16] A. E. W. Johnson et al. "MIMIC-IV, a freely accessible electronic health record dataset". In: *Sci. Data* 10 (2023), p. 1. DOI: 10.1038/s41597-022-01899-x.

[17] P. J. Thoral et al. "Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example". In: *Crit. Care Med.* 49 (2021), e563–e577. DOI: 10.1097/CCM.0000000000004916.

[18] T. C. Desautels et al. "Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach". In: *JMIR Med. Inform.* 4 (2016), e28. DOI: 10.2196/medinform.5909.

[19] Anthropic. *Claude: A family of highly capable AI assistants*. Model: Claude Opus 4.5, accessed January 2026. 2025. URL: https://www.anthropic.com/claude.

[20] A. E. Johnson, D. J. Stone, L. A. Celi, and T. J. Pollard. "The MIMIC Code Repository: enabling reproducibility in critical care research". In: *J Am Med Inf. Assoc* (2018). DOI: 10.1093/jamia/ocx084.

[21] Y.-G. Wang, J.-C. Zhou, and K.-S. Wu. "High 28-day mortality in critically ill patients with sepsis and concomitant active cancer". In: *J Int Med Res* (2018). DOI: 10.1177/0300060518789040.

[22] S. Parbhoo et al. "Improving counterfactual reasoning with kernelised dynamic mixing models". In: *PLoS One* (2018). DOI: 10.1371/journal.pone.0205839.

[23] J.-C. Peng et al. "Favorable Outcomes of Anticoagulation With Unfractioned Heparin in Sepsis-Induced Coagulopathy: A Retrospective Analysis of MIMIC-III Database". In: *Front Med Lausanne* (2021). DOI: 10.3389/fmed.2021.773339.

[24] H. Chen et al. “Central venous pressure measurement is associated with improved outcomes in septic patients: an analysis of the MIMIC-III database”. In: *Crit Care* (2020). DOI: 10.1186/s13054-020-03109-9.

[25] R.-Q. Yao et al. “A Machine Learning-Based Prediction of Hospital Mortality in Patients With Post-operative Sepsis”. In: *Front Med Lausanne* (2020). DOI: 10.3389/fmed.2020.00445.

[26] Y. Chen et al. “Association between chronic obstructive pulmonary disease and 28-day mortality in patients with sepsis: a retrospective study based on the MIMIC-III database”. In: *BMC Pulm Med* (2023). DOI: 10.1186/s12890-023-02729-5.

[27] C. Drudi, M. Mollura, L. H. Lehman, and R. Barbieri. “A Reinforcement Learning Model for Optimal Treatment Strategies in Intensive Care: Assessment of the Role of Cardiorespiratory Features”. In: *IEEE Open J. Eng. Med. Biol.* 5 (2024), pp. 806–815. DOI: 10.1109/OJEMB.2024.3454958.

[28] M. Huang, M. Han, W. Han, and L. Kuang. “Proton pump inhibitors versus histamine-2 receptor blockers for stress ulcer prophylaxis in patients with sepsis: a retrospective cohort study”. In: *J Int Med Res* (2021). DOI: 10.1177/03000605211025130.

[29] H. Chen, C. Zhao, Y. Wei, and J. Jin. “Early lactate measurement is associated with better outcomes in septic patients with an elevated serum lactate level”. In: *Crit Care* (2019). DOI: 10.1186/s13054-019-2625-0.

[30] Y. Choi et al. “Deep reinforcement learning extracts the optimal sepsis treatment policy from treatment records”. In: *Commun. Med.* 4 (2024), p. 245. DOI: 10.1038/s43856-024-00668-2.

[31] Q. Chen et al. “Transferability and interpretability of the sepsis prediction models in the intensive care unit”. In: *BMC Med Inf. Decis Mak* (2022). DOI: 10.1186/s12911-022-02090-3.

[32] C. Wang, J. Zheng, J. Wang, L. Zou, and Y. Zhang. “Cox-LASSO Analysis for Hospital Mortality in Patients With Sepsis Received Continuous Renal Replacement Therapy: A MIMIC-III Database Study”. In: *Front Med Lausanne* (2021). DOI: 10.3389/fmed.2021.778536.

[33] Z. Shahn, N. I. Shapiro, P. D. Tyler, D. Talmor, and L.-W. H. Lehman. “Fluid-limiting treatment strategies among sepsis patients in the ICU: a retrospective causal analysis”. In: *Crit Care* (2020). DOI: 10.1186/s13054-020-2767-0.

[34] K. Zhang et al. “Development and Validation of a Sepsis Mortality Risk Score for Sepsis-3 Patients in Intensive Care Unit”. In: *Front Med Lausanne* (2020). DOI: 10.3389/fmed.2020.609769.

[35] N. Hou et al. “Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost”. In: *J Transl Med* (2020). DOI: 10.1186/s12967-020-02620-5.

[36] Y. Ren et al. “Risk factor analysis and nomogram for predicting in-hospital mortality in ICU patients with sepsis and lung infection”. In: *BMC Pulm Med* (2022). DOI: 10.1186/s12890-021-01809-8.

[37] M. Rosnati and Vincent F. “MGP-AttTCN: An interpretable machine learning model for the prediction of sepsis”. In: *PloS One* 16 (2021), e0251248. DOI: 10.1371/journal.pone.0251248.

[38] L. Zhou, M. Shao, C. Wang, and Y. Wang. "An early sepsis prediction model utilizing machine learning and unbalanced data processing in a clinical context". In: *Prev. Med. Rep.* 45 (2024), p. 102841. DOI: 10.1016/j.pmedr.2024.102841.

[39] M. Peng, F. Deng, and D. Qi. "Development of a nomogram model for the early prediction of sepsis-associated acute kidney injury in critically ill patients". In: *Sci Rep* (2023). DOI: 10.1038/s41598-023-41965-x.

[40] D. Yan et al. "U-shaped association between serum calcium levels and 28-day mortality in patients with sepsis: a retrospective analysis of the MIMIC-III database". In: *Shock* (2023). DOI: 10.1097/SHK.0000000000002203.

[41] M. Apalak and K. Kiasaleh. "Advancing Early Detection of Sepsis With Temporal Convolutional Networks Using ECG Signals". In: *IEEE Access* 12 (2024), pp. 3417–3427. DOI: 10.1109/ACCESS.2023.3349019.

[42] P. Lan et al. "High Serum Iron level is Associated with Increased Mortality in Patients with Sepsis". In: *Sci Rep* (2018). DOI: 10.1038/s41598-018-29353-2.

[43] T. Wang et al. "Risk Potential for Organ Dysfunction Associated With Sodium Bicarbonate Therapy in Critically Ill Patients With Hemodynamic Worsening". In: *Front Med Lausanne* (2021). DOI: 10.3389/fmed.2021.665907.

[44] N. Sakib, S. I. Ahamed, R. A. Khan, P. M. Griffin, and M. M. Haque. "Unpacking Prevalence and Dichotomy in Quick Sequential Organ Failure Assessment and Systemic Inflammatory Response Syndrome Parameters: Observational Data-Driven Approach Backed by Sepsis Pathophysiology". In: *JMIR Med Inf.* (2020). DOI: 10.2196/18352.

[45] S. Lin, D. Lai, and W. He. "Association between hyperglycemia and adverse clinical outcomes of sepsis patients with diabetes". In: *Front Endocrinol Lausanne* (2022). DOI: 10.3389/fendo.2022.1046736.

[46] X. Zhong et al. "Association between different MAP levels and 30-day mortality in sepsis patients: a propensity-score-matched, retrospective cohort study". In: *BMC Anesth.* (2023). DOI: 10.1186/s12871-023-02047-7.

[47] S. Wang, S. Ding, H. Luo, and X. Chai. "International Normalized Ratio to Albumin Ratio (PTAR): An Objective Risk Stratification Tool in Patients with Sepsis". In: *Int J Gen Med* (2021). DOI: 10.2147/IJGM.S305085.

[48] A. E. W. Johnson et al. "A Comparative Analysis of Sepsis Identification Methods in an Electronic Database". In: *Crit. Care Med.* 46 (2018), pp. 494–499. DOI: 10.1097/CCM.0000000000002965.

[49] H. Wang et al. "Prognostic Value of the Red Cell Distribution Width in Patients with Sepsis-Induced Acute Respiratory Distress Syndrome: A Retrospective Cohort Study". In: *Markers* (2021). DOI: 10.1155/2021/5543822.

[50] R. Liu et al. “Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU”. In: *Sci Rep* (2019). DOI: 10.1038/s41598-019-42637-5.

[51] P. F. Ghalati et al. “Critical Transitions in Intensive Care Units: A Sepsis Case Study”. In: *Sci Rep* (2019). DOI: 10.1038/s41598-019-49006-2.

[52] G. Rhodes, M. Davidian, and W. Lu. “Dynamic prediction of residual life with longitudinal covariates using long short-term memory networks”. In: *Ann Appl Stat* (2023). DOI: 10.1214/22-aoas1706.

[53] G. Rhodes, M. Davidian, and W. Lu. “Estimation of optimal treatment regimes with electronic medical record data using the residual life value estimator”. In: *Biostatistics* 25 (2024), pp. 933–946. DOI: 10.1093/biostatistics/kxae002.

[54] R. Liu, K. M. Hunold, J. M. Caterino, and P. Zhang. “Estimating treatment effects for time-to-treatment antibiotic stewardship in sepsis”. In: *Nat Mach Intell* (2023). DOI: 10.1038/s42256-023-00638-0.

[55] M. Komorowski, L. Celi, O. Badawi, A. Gordon, and A. Faisal. “The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care”. In: *Nat. Med.* 24 (2018), p. 1716. DOI: 10.1038/s41591-018-0213-5.

[56] Q. Chen, L. Zhang, S. Ge, W. He, and M. Zeng. “Prognosis predictive value of the Oxford Acute Severity of Illness Score for sepsis: a retrospective cohort study”. In: *PeerJ* (2019). DOI: 10.7717/peerj.7083.

[57] M. B. Moor et al. “Predicting sepsis using deep learning across international sites: a retrospective development and validation study”. In: *EClinicalMedicine* 62 (2023), p. 102124. DOI: 10.1016/j.eclinm.2023.102124.

[58] B. He and Z. Qiu. “Development and validation of an interpretable machine learning for mortality prediction in patients with sepsis”. In: *Front. Artif. Intell.* 7 (2024), p. 1348907. DOI: 10.3389/frai.2024.1348907.

[59] X. Wei et al. “Admission Blood Glucose Is Associated With the 30-Days Mortality in Septic Patients: A Retrospective Cohort Study”. In: *Front Med Lausanne* (2021). DOI: 10.3389/fmed.2021.757061.

[60] M. Peng, F. Deng, D. Qi, Z. Hu, and L. Zhang. “The Hyperbilirubinemia and Potential Predictors Influence on Long-Term Outcomes in Sepsis: A Population-Based Propensity Score-Matched Study”. In: *Front Med Lausanne* (2021). DOI: 10.3389/fmed.2021.713917.

[61] D.-K. Li, X.-T. Wang, and D.-W. Liu. “Association between elevated central venous pressure and outcomes in critically ill patients”. In: *Ann Intensive Care* (2017). DOI: 10.1186/s13613-017-0306-1.

[62] J. Solís-García, B. Vega-Márquez, J. A. Nepomuceno, J. C. Riquelme-Santos, and I. A. Nepomuceno-Chamorro. “Comparing artificial intelligence strategies for early sepsis detection in the ICU: an experimental study”. In: *Appl. Intell.* 53 (2023), pp. 30691–30705. DOI: 10.1007/s10489-023-05124-z.

[63] C. Barton et al. “Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs”. In: *Comput Biol Med* (2019). DOI: 10.1016/j.compbiomed.2019.04.027.

[64] Y. Fang, Y. Zhang, and X. Zhang. "Serum phosphate levels and the development of sepsis associated acute kidney injury: evidence from two independent databases". In: *Front Med Lausanne* (2024). DOI: 10.3389/fmed.2024.1367064.

[65] G. Zhang et al. "Predicting sepsis in-hospital mortality with machine learning: a multi-center study using clinical and inflammatory biomarkers". In: *Eur J Med Res* (2024). DOI: 10.1186/s40001-024-01756-0.

[66] G. Zhang et al. "The neutrophil-to-lymphocyte ratio levels over time correlate to all-cause hospital mortality in sepsis". In: *Heliyon* (2024). DOI: 10.1016/j.heliyon.2024.e36195.

[67] G. Zhang et al. "U-shaped correlation of lymphocyte count with all-cause hospital mortality in sepsis and septic shock patients: a MIMIC-IV and eICU-CRD database study". In: *Int J Emerg Med* (2024). DOI: 10.1186/s12245-024-00682-6.

[68] L. Zhang et al. "Developing an ensemble machine learning model for early prediction of sepsis-associated acute kidney injury". In: *iScience* (2022). DOI: 10.1016/j.isci.2022.104932.

[69] Q. Yang et al. "Association between the β-blocker use and patients with sepsis: a cohort study". In: *Front Med Lausanne* (2023). DOI: 10.3389/fmed.2023.1272871.

[70] L. Zhang et al. "Influence of fluid balance on the prognosis of patients with sepsis". In: *BMC Anesth.* (2021). DOI: 10.1186/s12871-021-01489-1.

[71] Z. Wang et al. "The relationship between hematocrit and serum albumin levels difference and mortality in elderly sepsis patients in intensive care units—a retrospective study based on two large database". In: *BMC Infect Dis* (2022). DOI: 10.1186/s12879-022-07609-7.

[72] Y. Li et al. "Conservative oxygen therapy in critically ill and perioperative period of patients with sepsis-associated encephalopathy". In: *Front Immunol* (2022). DOI: 10.3389/fimmu.2022.1035298.

[73] Q.-Y. Zhao et al. "A Machine-Learning Approach for Dynamic Prediction of Sepsis-Induced Coagulopathy in Critically Ill Patients With Sepsis". In: *Front Med Lausanne* (2020). DOI: 10.3389/fmed.2020.637434.

[74] Y. Lin et al. "Association between 24-hour urine volume and 28-day intensive care unit mortality in sepsis patients: a multi-center retrospective cohort study". In: *Front Med Lausanne* (2024). DOI: 10.3389/fmed.2024.1486232.

[75] W. Lin et al. "Association between the serum albumin-creatinine ratio and 28-day intensive care unit mortality among patients with sepsis: a multicenter retrospective cohort study". In: *Front Med Lausanne* (2024). DOI: 10.3389/fmed.2024.1484370.

[76] L. Shen et al. "Interpretable machine learning-based prediction of 28-day mortality in ICU patients with sepsis: a multicenter retrospective study". In: *Front Cell Infect Microbiol* (2024). DOI: 10.3389/fcimb.2024.1500326.

[77] L. Chang, X. Chen, and C. Lian. "The association between the non-HDL-cholesterol to HDL-cholesterol ratio and 28-day mortality in sepsis patients: a cohort study". In: *Sci Rep* (2022). DOI: 10.1038/s41598-022-07459-y.

[78] M. Yang et al. "Enhancing Patient Selection in Sepsis Clinical Trials Design Through an AI Enrichment Strategy: Algorithm Development and Validation". In: *J Med Internet Res* (2024). DOI: 10.2196/54621.

[79] D. Kasugai et al. "Relative platelet reductions provide better pathophysiologic signatures of coagulopathies in sepsis". In: *Sci Rep* (2021). DOI: 10.1038/s41598-021-93635-5.

[80] K. Wang, D. Lu, and F. Wang. "Subphenotypes of platelet count trajectories in sepsis from multicenter ICU data". In: *Sci Rep* (2024). DOI: 10.1038/s41598-024-71186-9.

[81] Y. Li et al. "Association Between Red Cell Distribution Width and Hospital Mortality in Patients with Sepsis". In: *J Int Med Res* (2021). DOI: 10.1177/03000605211004221.

[82] S. Sheng et al. "Association between hemoglobin and in-hospital mortality in critically ill patients with sepsis: evidence from two large databases". In: *BMC Infect Dis* (2024). DOI: 10.1186/s12879-024-10335-x.

[83] K. Takkavatakarn et al. "Machine learning derived serum creatinine trajectories in acute kidney injury in critically ill patients with sepsis". In: *Crit Care* (2024). DOI: 10.1186/s13054-024-04935-x.

[84] J. Chen et al. "Association of longitudinal platelet count trajectory with ICU mortality: A multicohort study". In: *Front. Immunol.* 13 (2022), p. 936662. DOI: 10.3389/fimmu.2022.936662.

[85] Y. Cao et al. "Public critical care databases for clinical research: challenges and opportunities". In: *Crit. Care* 30 (2026), p. 26. DOI: 10.1186/s13054-025-05280-z.

[86] O. T. Ranzani et al. "Development and Validation of the Sequential Organ Failure Assessment (SOFA)-2 Score". In: *JAMA* 334 (2025), p. 2090. DOI: 10.1001/jama.2025.0104.

[87] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons. "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement". In: *Br. J. Surg.* 102 (2015), pp. 148–158. DOI: 10.1002/bjs.9736.

[88] the CONSORT Group, K. F. Schulz, D. G. Altman, and D. Moher. "CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials". In: *BMC Med.* 8 (2010), p. 18. DOI: 10.1186/1741-7015-8-18.

[89] S. N. Cohen et al. "Subtle variation in Sepsis-III definitions markedly influences predictive performance within and across methods". In: *Sci. Rep.* 14 (2024), p. 1920. DOI: 10.1038/s41598-024-51426-6.

[90] T. Edinburgh, S. J. Eglen, P. Thoral, P. Elbers, and A. Ercole. "Sepsis-3 criteria in AmsterdamUMCdb: open-source code implementation". In: *Gigabyte* 2022 (2022), pp. 1–7. DOI: 10.46471/gigabyte.55.

[91] S. D. Boie et al. "A scalable approach for critical care data extraction and analysis in an academic medical center". In: *Int. J. Med. Inf.* 192 (2024), p. 105611. DOI: 10.1016/j.ijmedinf.2024.105611.