Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review

Michael Moor [1,2*†], Bastian Rieck [1,2†], Max Horn [1,2], Catherine R. Jutzeler [1,2‡] and Karsten Borgwardt [1,2‡]

[1] Machine Learning and Computational Biology Lab, Department of Biosystems Science and Engineering, Eidgenössische Technische Hochschule Zürich (ETH Zurich), Basel, Switzerland, [2] SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

Background: Sepsis is among the leading causes of death in intensive care units (ICUs) worldwide and its recognition, particularly in the early stages of the disease, remains a medical challenge. The advent of an affluence of available digital health data has created a setting in which machine learning can be used for digital biomarker discovery, with the ultimate goal to advance the early recognition of sepsis.

Objective: To systematically review and evaluate studies employing machine learning for the prediction of sepsis in the ICU.

Data Sources: Using Embase, Google Scholar, PubMed/Medline, Scopus, and Web of Science, we systematically searched the existing literature for machine learning-driven sepsis onset prediction for patients in the ICU.

Study Eligibility Criteria: All peer-reviewed articles using machine learning for the prediction of sepsis onset in adult ICU patients were included. Studies focusing on patient populations outside the ICU were excluded.

Study Appraisal and Synthesis Methods: A systematic review was performed according to the PRISMA guidelines. Moreover, a quality assessment of all eligible studies was performed.

Results: Out of 974 identified articles, 22 and 21 met the criteria to be included in the systematic review and quality assessment, respectively. A multitude of machine learning algorithms were applied to refine the early prediction of sepsis. The quality of the studies ranged from "poor" (satisfying $\leq 40\%$ of the quality criteria) to "very good" (satisfying $\geq 90\%$ of the quality criteria). The majority of the studies ($n = 19$, 86.4%) employed an offline training scenario combined with a horizon evaluation, while two studies implemented an online scenario ($n = 2$, 9.1%). The massive inter-study heterogeneity in terms of model development, sepsis definition, prediction time windows, and outcomes precluded a meta-analysis. Last, only two studies provided publicly accessible source code and data sources fostering reproducibility.

Limitations: Articles were only eligible for inclusion when employing machine learning algorithms for the prediction of sepsis onset in the ICU. This restriction led to the exclusion of studies focusing on the prediction of septic shock, sepsis-related mortality, and patient populations outside the ICU.

Conclusions and Key Findings: A growing number of studies employs machine learning to optimize the early prediction of sepsis through digital biomarker discovery. This review, however, highlights several shortcomings of the current approaches, including low comparability and reproducibility. Finally, we gather recommendations how these challenges can be addressed before deploying these models in prospective analyses.

Systematic Review Registration Number: CRD42020200133.

Keywords: sepsis, machine learning, onset prediction, early detection, systematic review

1. INTRODUCTION

Sepsis is a life-threatening organ dysfunction triggered by dysregulated host response to infection (1) and constitutes a major global health concern (2). Despite promising medical advances over the last decades, sepsis remains among the most common causes of in-hospital deaths. It is associated with an alarmingly high mortality and morbidity, and massively burdens the health care systems world-wide (2–5). In parts, this can be attributed to challenges related to early recognition of sepsis and initiation of timely and appropriate treatment (6). A growing number of studies suggests that the mortality increases with every hour the antimicrobial intervention is delayed, further underscoring the importance of timely recognition and initiation of treatment (6–8). A major challenge to early recognition is to distinguish sepsis from disease states (e.g., inflammation) that are hallmarked by similar clinical signs (e.g., change in vitals), symptoms (e.g., fever), and molecular manifestations (e.g., dysregulated host response) (9, 10). Owing to the systemic nature of sepsis, biological and molecular correlates—also known as biomarkers—have been proposed to refine the diagnosis and detection of sepsis (5). However, despite considerable efforts to identify suitable biomarkers, there is yet no single biomarker or set thereof that is universally accepted for sepsis diagnosis and treatment, mainly due to the lack of sensitivity and specificity (11, 12).

In addition to the conventional approaches, *data-driven biomarker discovery* has gained momentum over the last decades and holds the promise to overcome existing hurdles. The goal of this approach is to mine and exploit health data with quantitative computational approaches, such as machine learning. An ever-increasing amount of data, including laboratory, vital, genetic, molecular, as well as clinical data and health history, is available in digital form and at high resolution for individuals at risk and for patients suffering from sepsis (13). This versatility of the data allows to search for digital biomarkers in a holistic fashion as opposed to a reductionist approach (e.g., solely focusing on hematological markers). Machine learning models can naturally handle the wealth and complexity of digital patient data by learning predictive patterns in the data, which in turn can be used to make accurate predictions about which patient is developing sepsis (14, 15). Searching predictive patterns is conventionally done either in a supervised or unsupervised fashion. Supervised

learning refers to algorithms that learn from labeled training data (e.g., patients have sepsis or not) to predict outcomes for unforeseen data. In contrast, in unsupervised learning, the data have no labels and the algorithm detects (known and unknown) patterns based on the data provided. Over the last decades, multiple studies have successfully employed a variety of computational models to tackle the challenge of predicting sepsis at the earliest time point possible (16–18). For instance, Futoma et al. proposed to combine multi-task Gaussian processes imputation together with a recurrent neural network in one end-to-end trainable framework (multi-task Gaussian process recurrent neural network [MGP-RNN]). They were able to predict sepsis 17 h prior to the first administration of antibiotics and 36 h before a definition for sepsis was met (19). This strategy was motivated by Li and Marlin (20), who first proposed the so-called Gaussian process adapter that combines single-task Gaussian processes imputation with neural networks in an end-to-end learning setting. A more recent study further improved predictive performance by combining the Gaussian process adapter framework with temporal convolutional networks (MGP-TCN) as well as leveraging a dynamic time warping approach for the early prediction of sepsis (21).

Considering the rapid pace at which the research in this field is moving forward, it is important to summarize and critically assess the state of the art. Thus, the aim of this review was to provide a comprehensive overview of the current state of machine learning models that have been employed for the search of digital biomarkers to aid the early prediction of sepsis in the intensive care unit (ICU). To this end, we systematically reviewed the literature and performed a quality assessment of all eligible studies. Based on our findings, we also provide some recommendations for forthcoming studies that plan to use machine learning models for the early prediction of sepsis.

2. METHODS

The study protocol was registered with and approved by the international prospective register of systematic reviews (PROSPERO) before the start of the study (registration number: CRD42020200133). We followed the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) statement (22).

2.1. Search Strategy and Selection Criteria

Five bibliographic databases were systematically searched, i.e., EMBASE, Google Scholar, PubMed/Medline, Scopus, and Web of Science, using the time range from their respective inception dates to July 20, 2020. Google Scholar was searched using the tool "Publish or Perish" (version 7.23.2852.7498) (23). Our search was not restricted by language. The search term string was constructed as ("sepsis prediction" OR "sepsis detection") AND ("machine learning" OR "artificial intelligence") to include publications focusing on (early) onset prediction of sepsis with different machine learning methods. The full search strategy is provided in **Supplementary Table 1**.

2.2. Selection of Studies

Two investigators (MM and CRJ) independently screened the titles, abstracts, and full texts retrieved from Google Scholar in order to determine the eligibility of the studies. Google Scholar was selected by virtue of its promise of an inclusive query that also captures conference proceedings, which are highly relevant to the field of machine learning but not necessarily indexed by other databases. In a second step, two investigators (MM and MH) queried EMBASE, PubMed, Scopus, and Web of Science for additional studies. Eligibility criteria were also applied to the full-text articles during the final selection. In case multiple articles reported on a single study, the article that provided the most data and details was selected for further synthesis. We quantified the inter-rater agreement for study selection using Cohen's kappa (κ) coefficient (24). All disagreements were discussed and resolved at a consensus meeting.

2.3. Inclusion and Exclusion Criteria

All full-text, peer-reviewed articles[1] using machine learning for the prediction of sepsis onset in the ICU were included. Although the 2016 consensus statement abandoned the term "severe sepsis" (1), studies published prior to the revised consensus statement targeting severe sepsis were also included in our review. Furthermore, to be included, studies must have provided sufficient information on the machine learning algorithms used for the analysis, definition of sepsis (e.g., Sepsis-3), and sepsis onset definition (e.g., time of suspicion of infection). We excluded duplicates, non-peer reviewed articles (e.g., preprints), reviews, meta-analyses, abstracts, editorials, commentaries, perspectives, patents, letters with insufficient data, studies on non-human species and children/neonates, or out-of-scope studies (e.g., different target condition). Lastly, studies focusing on the prediction of septic shock were also excluded as the septic shock was beyond the scope of this review. The extraction was performed by four investigators (MM, BR, MH, and CRJ).

2.4. Data Extraction and Synthesis

The following information was extracted from all studies: (i) publication characteristics (first author's last name, publication time), (ii) study design (retrospective, prospective data collection and analysis), (iii) cohort selection (sex, age, prevalence of sepsis), (iv) model selection (machine learning algorithm, platforms, software, packages, and parameters), (v) specifics on the data analyzed (type of data, number of variables), (vi) statistics for model performance (methods to evaluate the model, mean, measure of variance, handling of missing data), and (vii) methods to avoid overfitting as well as any additional external validation strategies. If available, we also reviewed supplementary materials of each study. A full list of extracted variables is provided in **Supplementary Table 2**.

2.5. Settings of Prediction Task

Owing to its time sensitivity, setting up the early sepsis prediction task in a clinically meaningful manner is a non-trivial issue. We extracted details on the prediction task as well as the alignment of cases and controls. Given the lack of standardized reporting, the implementation strategies and their reporting vary drastically between studies. Thus, subsequent to gathering all the information, we attempted to create new categories for the sepsis prediction task as well as the case–control alignment. The goal of this new terminology and categories is to increase the comparability between studies.

2.6. Assessment of Quality of Reviewed Machine Learning Studies

Based on 14 criteria relevant to the objectives of the review, which we adapted from Qiao (25), the quality of the eligible machine learning studies was assessed. The quality assessment comprised five categories: (1) unmet needs (limits in current machine learning or non-machine learning applications), (2) reproducibility (information on the sepsis prevalence, data and code availability, explanation of sepsis label, feature engineering methods, software/hardware specifications, and hyperparameters), (3) robustness (sample size suited for machine learning applications, valid methods to overcome overfitting, stability of results), (4) generalizability (external data validation), and (5) clinical significance (interpretation of predictors and suggested clinical use; see **Supplementary Table 3**). A quality assessment table was provided by listing "yes" or "no" of corresponding items in each category. MM, BR, MH, and CRJ independently performed the quality assessment. In case of disagreements, ratings were discussed and subsequently, final scores for each publication were determined.

2.7. Role of Funding Source

The funding sources of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

3. RESULTS

3.1. Study Selection

The results of the literature search, including the numbers of studies screened, assessments for eligibility, and articles

---

[1]This includes peer-reviewed journal articles and peer-reviewed conference proceedings.

FIGURE 1 | PRISMA flowchart of the search strategy. A total of 22 studies were eligible for the literature review and 21 for the quality assessment.
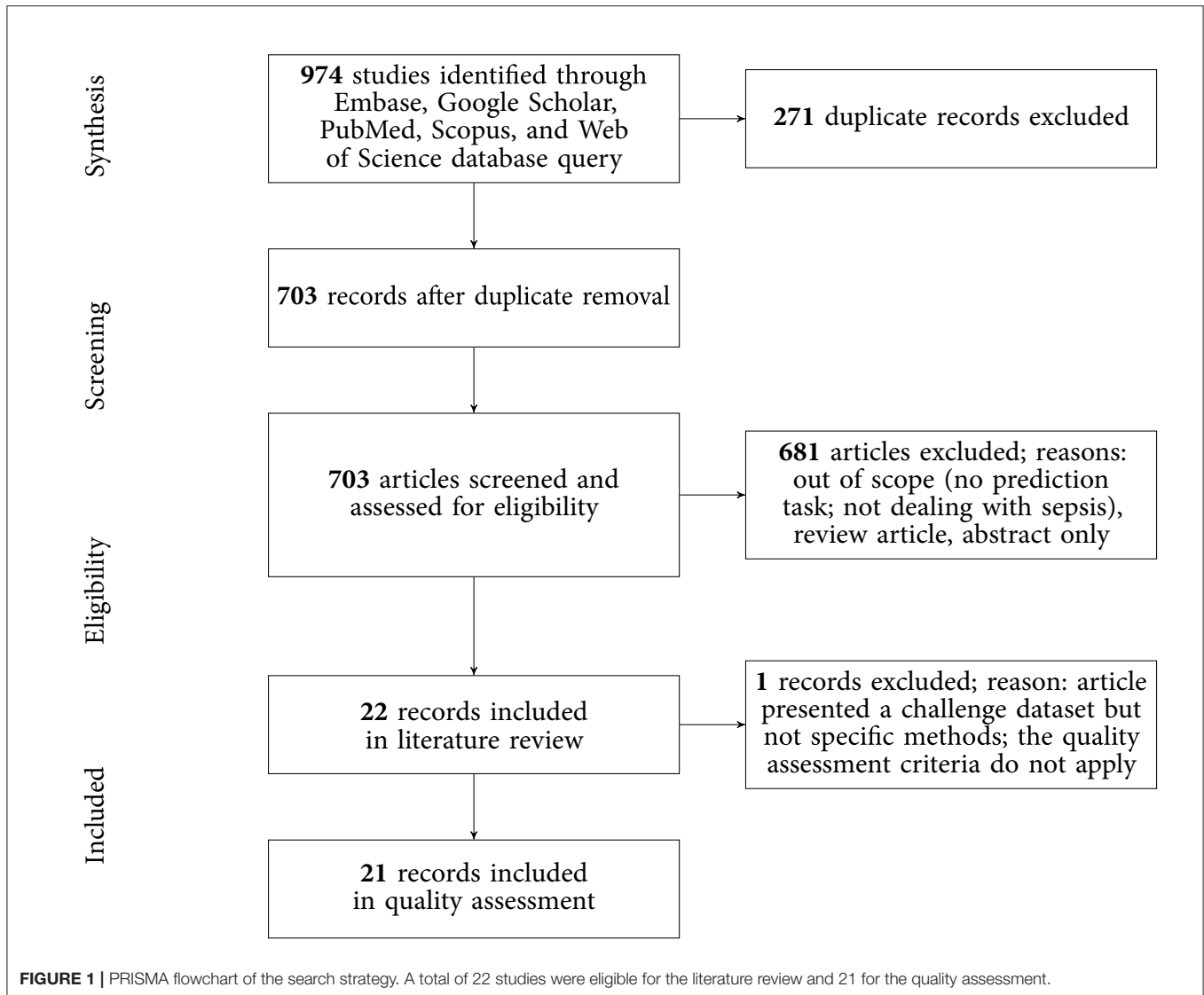
reviewed (with reasons for exclusions at each stage), are presented in **Figure 1**. Out of 974 studies, 22 studies met the inclusion criteria (16–19, 21, 26–42). The majority of excluded studies ($n = 952$) did not meet one or multiple inclusion criteria, such as studying a non-human (e.g., bovine) or a non-adult population (e.g., pediatric or neonatal), focusing on a research topic beyond the current review (e.g., sepsis phenotype identification or mortality prediction), or following a different study design (e.g., case reports, reviews, not-peer reviewed). Detailed information on all included studies are provided in **Table 1**. The inter-rater agreement was excellent ($\kappa = 0.88$).

3.2. Study Characteristics
Of the 22 included studies, 21 employed solely retrospective analyses, while one study used both retrospective and prospective analyses (16). Moreover, the most frequent data sources used to develop computational models were MIMIC-II and MIMIC-III ($n = 12$; 54.5%), followed by Emory University Hospital ($n = 5$; 22.7%). In terms of sepsis definition, the majority of the studies employed the Sepsis-2 ($n = 12$; 54.5%) or Sepsis-3 definition ($n = 9$; 40.9%). It is important to note that some studies *modified* the Sepsis-2 or Sepsis-3 definition since all existing definitions have not been intended to specify an exact sepsis onset time (e.g., the employed time window lengths have been varied) (26, 34). In one study (36), sepsis labels were assigned by trained ICU experts. Depending on the definition of sepsis used, and whether subsampling of controls was used to achieve a more balanced class ratio (facilitating the training of machine learning models), the prevalence of patients developing sepsis ranged between 3.3% (See **Table 1**) and 63.6% (**Figure 2**). One study did not report the prevalence (31). Concerning demographics, 9 studies reported the median or mean age, 12 the prevalence of female patients, and solely 1 the ethnicity of the investigated cohorts (**Supplementary Table 4**).

TABLE 1 | Overview of included studies.

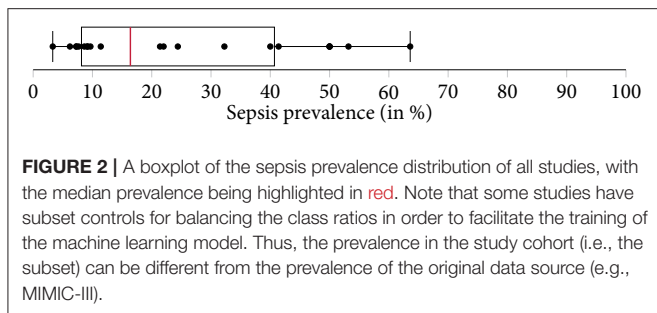|	References	Dataset	Sepsis definition	Number of sepsis encounters	Prevalence (%)	Used cohort available	Code for analysis	Code for label	Model	AUROC	Hours before onset	External validation	Data types	Number of variables
1	Abromavičius et al. (26)	Emory University Hospital, MIMIC-III	Sepsis-3 (with modified time windows)	2,932	7.3	Yes	No	No	AdaBoost and Discriminant Subspace Learning	–	–	No	Demographics, labs, vitals	11
2	Barton et al. (17)	MIMIC-III, UCSF	Sepsis-3	3,673	3.3	No	No	No	XGBoost	0.88	0	No	Vitals	6
3	Bloch et al. (27)	RMC	Sepsis-2 related	300	50.0	No	No	No	Neural Networks, SVM, logistic regression	0.88	4	No	Vitals	4
4	Calvert et al. (28)	MIMIC-II	Sepsis-2 related	159	11.4	No	No	No	InSight Algorithm	0.92	3	No	Demographics, labs, vitals	9
5	Desautels et al. (29)	MIMIC-III	Sepsis-3	1,840	9.7	No	No	No	InSight Algorithm	0.88	0	No	Demographics, vitals	8
6	Futoma et al. (19)	Duke University Health System	Sepsis-2 related	11,064	21.4	No	No	No	MGP-RNN	0.91	0	No	Comorbidities, demographics, labs, medications, vitals	77
7	Kaji et al. (18)	MIMIC-III	Sepsis-2 related	36,176	63.6	Yes	Yes	Yes	LSTM	0.88	"Next day"	No	Demographics, labs, medications, vitals	119
8	Kam and Kim (30)	MIMIC-II	Sepsis-2 related	360	6.2	No	No	No	SepLSTM	0.99	0	No	Demographics, labs, vitals	9
9	Lauritsen et al. (31)	Danish EHR	Sepsis-2 related	–	–	No	No	No	CNN-LSTM	0.88	0.25	No	Diagnoses, labs, imaging, medications, vitals, procedures	–
10	Lukaszewski et al. (32)	Queen Alexandra Hospital	Sepsis-2 related	25	53.2	No	No	No	MLP	–	–	No	Clinical parameters, cytokine mRNA expression	–
11	Mao et al. (33)	MIMIC-III, UCSF	Sepsis-2 related	1,965	9.1	Yes	No	No	InSight Algorithm	0.92	0	Yes	Vitals	30
12	McCoy and Das (16)	CRMC	Sepsis-3, Severe Sepsis	407	24.4	No	No	No	InSight Algorithm	0.91	–	–	Labs, vitals	–
13	Moor et al. (21)	MIMIC-III	Sepsis-3	570	9.2	Yes	Yes	Yes	MGP-TCN	0.91	0	No	Labs, vitals	44

(Continued)

TABLE 1 | Continued

	References	Dataset	Sepsis definition	Number of sepsis encounters	Prevalence (%)	Used cohort available	Code for analysis	Code for label	Model	AUROC	Hours before onset	External validation	Data types	Number of variables
14	Nemati et al. (34)	Emory Healthcare system, MIMIC-III	Sepsis-3 (modified time windows)	2,375	8.6	No	No	No	Weilbull-Cox proportional hazards model	0.85	4	Yes	Demographics, vitals	48
15	Reyna et al. (35)	Emory University Hospital, MIMIC-III	Sepsis-3 (modified time windows)	2,932	7.3	Yes	No	No	–	–	–	Yes	Demographics, labs, vitals	40
16	Schamoni et al. (36)	University Medical Centre Mannheim	Sepsis tag by ICU clinicians	200	32.3	No	No	No	Non-linear ordinal regression	0.84	4	No	Comorbidities, demographics, labs, vitals	55
17	Scherpf et al. (37)	MIMIC-III	Sepsis-2 related	2,724	7.7	No	No	No	RNN-GRU	0.81	3	No	Labs, vitals	10
18	Shashikumar et al. (38)	Emory Healthcare system	Sepsis-3	242	22.0	No	No	No	ElasticNet	0.78	4	No	Comorbidities, clinical context, demographics, vitals	17
19	Shashikumar et al. (39)	Emory Healthcare system	Sepsis-3	100	40.0	No	No	No	SVM	0.8	4	No	Demographics, comorbidity, clinical context, vitals	2
20	Sheetrit et al. (40)	MIMIC-III	Sepsis-2 related	1,034	41.4	No	No	No	Temporal Probabilistic Profiles	–	–	No	Demographics, labs, vitals	–
21	van Wyk et al. (41)	MLH System	Sepsis-2 related	–	50.0	No	No	No	Random Forests, RNN	–	–	No	Labs, vitals	7
22	van Wyk et al. (42)	MLH System	Sepsis-2 related	377	50.0	No	No	No	Random Forests	0.79	0	No	Vitals	7

Only if area under the Receiver Operating Characteristic Curve (AUROC) was reported in an early prediction setup, the performance and the corresponding prediction window is reported (in hours before onset). As these windows were highly heterogeneous, to achieve more comparability, we report the minimal hour before onset that was reported. Notably, due to heterogeneous sepsis definition implementations and experimental setups, these metrics likely have low comparability between studies, which is why we deemed a quantitative meta-analysis to be inappropriate.

AUROC, area under the ROC curve; CNN-LSTM, convolutional neural network long short-term memory; EHR, electronic health record; ICU, intensive care unit; LSTM, long short-term memory; MGP-RNN, multi-task Gaussian process recurrent neural network; MGP-TCN, multi-task Gaussian process temporal convolutional network; MIMIC, medical information mart for intensive care; MLH, Methodist Le Bonheur Healthcare System; MLP, multilayer perceptron; RMC, Rabin Medical Center; RNN-GRU, recurrent neural net gated recurrent unit; SepLSTM, proper name for LSTM for sepsis; SVM, support vector machine; USCF, University of California San Francisco Health System.

FIGURE 2 | A boxplot of the sepsis prevalence distribution of all studies, with the median prevalence being highlighted in red. Note that some studies have subset controls for balancing the class ratios in order to facilitate the training of the machine learning model. Thus, the prevalence in the study cohort (i.e., the subset) can be different from the prevalence of the original data source (e.g., MIMIC-III).

3.3. Overview of Machine Learning Algorithms and Data

As shown in **Table 1**, a wide range of predictive models was employed for the early detection of sepsis, with some models being specifically developed for the respective application. Most prominently, various types of neural networks ($n = 9$; 40.9%) were used. This includes recurrent architectures, such as long short-term memory (LSTM) (43) or gated recurrent units (GRU) (44), convolutional networks (45), as well as temporal convolutional networks, featuring causal, dilated convolutions (46, 47). Furthermore, several studies employed boosted tree models ($n = 4$; 18.2%), including XGBoost (48) or random forest (49). As for the data analyzed, the most common data type were vitals ($n = 21$; 95.5%), followed by laboratory values ($n = 13$; 59.1%), demographics ($n = 12$; 54.5%), and comorbidities ($n = 4$; 18.2%). The number of variables included in the respective models ranged between 2 (38) and 119 (18). While reporting the type of variables, four studies failed to report the number of variables included in the models (16, 31, 32, 40).

3.4. Model Validation

Approximately 80% of the studies employed one type of cross-validation (e.g., 5-fold, 10-fold, or leave-one-out cross-validation) to avoid overfitting. Additional validation of the models on out-of-distribution ICU data (i.e., external validation) was only performed in three studies (33–35). Specifically, Mao et al. (33) used a dataset provided by the UCSF Medical Center as well as the MIMIC-III dataset to train, validate, and test the *InSight* algorithm. Aiming at developing and validating the *Artificial Intelligence Sepsis Expert (AISE)* algorithm, Nemati et al. (34) created a development cohort using ICU data of over 30,000 patients admitted to two Emory University hospitals. In a subsequent step, the *AISE* algorithm was externally validated on the publicly available MIMIC-III dataset (at the time containing data from over 52,000 ICU stays of more than 38,000 unique patients) (34). Last, the study by Reyna et al. (35) describes the protocol and results of the *PhysioNet/Computing in Cardiology Challenge 2019*. Briefly, the aim of this challenge was to facilitate the development of automated, open-source algorithms for the early detection of sepsis. The PhysioNet/Computing in Cardiology Challenge provided sequestered real-world datasets to the participating researchers for the training, validation, and testing of their models.

3.5. Experimental Design Choices for Sepsis Onset Prediction

In this review, we identified two main approaches of implementing sepsis prediction tasks on ICU data. The most-frequent setting ($n = 19$; 86.4%) combines "offline" training with a "horizon" evaluation. Briefly, offline training refers to the fact that the models have access to the entire feature window of patient data. For patients with sepsis, this feature window ranges from hospital admission to sepsis onset, while for the control subjects the endpoint is a matched onset. Alternatively, a prediction window (i.e., a *gap*) between the feature window and the (matched) onset has been employed (27). As for the "horizon" evaluation, the purpose is to determine how early the fitted model would recognize sepsis. To this end, all input data gathered up to n h before onset is provided to the model for the sepsis prediction at a horizon of n h. For studies employing only a single horizon, i.e., predictions preceding sepsis onset by a fixed number of hours, we denote their task as "offline" evaluation in **Table 2**, since there are no sequentially repeated predictions over time. This experimental setup, offline training plus horizon evaluation, is visualized in **Figure 3**. In the second most-frequently used sepsis prediction setting ($n = 2$; 9.1%), both the training and evaluation occur in an "online" fashion. This means that the model is presented with all the data that have been collected until the time point of prediction. The amount of data depends on the spacing of data collection. In order to incentivize early predictions, these timepoint-wise labels can be shifted into the past: in the case of the PhysioNet Challenge dataset, already timepoint-wise labels 6 h before onset are assigned to the positive (sepsis) class (35). For an illustration of an online training and evaluation scenario, refer to **Figure 4**.

Selecting the "onset" for controls (i.e., case–control alignment) is a crucial step in the development of models predicting the onset of sepsis (19). Surprisingly, the majority of the studies ($n = 16$; 72.7%) did not report any details on how the onset matching was performed. For the six studies (27.3%) providing details, we propose the following classification: four employed *random onset matching*, one *absolute onset matching*, and one *relative onset matching* (**Figure 3**, top). As the name indicates, during random onset matching, the onset time of a control is set at a random time of the ICU stay. Often, this time has to satisfy certain additional constraints, such as not being too close to the patient's discharge. The absolute onset matching refers to taking the absolute time since admission until sepsis onset for the case and assigning it as the matched onset time for a control (21). Finally, the relative onset matching is when the matched onset time is defined as the relative time since ICU admission until sepsis onset for the case (50).

3.6. Quality of Included Studies

The results of the quality assessment are shown in **Table 3**. One study (35), showcasing the results of the *PhysioNet/Computing in Cardiology Challenge 2019*, was excluded from the quality assessment, which was intended to assess the quality of the implementation and reporting of *specific* prediction models. The quality of the remaining 21 studies ranged from poor (satisfying

TABLE 2 | An overview of experimental details: the used sepsis definition, the exact prediction task, and which type of temporal case–control alignment was used (if any).

References	Prediction task	Sepsis definition	Case–control alignment	Inclusion criteria
1 Abromavičius et al. (26)	Online training, online evaluation	Sepsis-3 (with modified time windows)	–	–
2 Barton et al. (17)	Offline training, horizon evaluation	Sepsis-3	Random onset matching	Inpatients, age ≥18 years, at least one observation per measurement, prediction times between 7 and 2,000 h
3 Bloch et al. (27)	Offline training, horizon evaluation	Sepsis-2 related: SIRS criteria plus diagnosis of infection	Random onset matching (at least 12 h after admission to the ICU)	age >18 years, admitted to ICU; minimum stay of 12 h in the ICU; patients did not meet SIRS criteria at time of admission to the ICU; Continuous documented measurements were available for at least 12 h for vital signs
4 Calvert et al. (28)	Offline training, horizon evaluation	Sepsis-2 related: ICD-9 code 995.9 and a 5-h persisting window of fulfilled SIRS	–	Medical ICU, age >18 years, SIRS not fulfilled upon admission, measurements for set of nine variables available
5 Desautels et al. (29)	Offline training, horizon evaluation, but retrained for each prediction horizon	Sepsis-3	–	Age ≥15 years, any measurements present, Metavision logging, for cases: sepsis onset between 7 and 500 h after ICU admission, all variables at least once measured, excluded patients that received antibiotics before ICU
6 Futoma et al. (19)	Offline training, horizon evaluation	Sepsis-2 related: SIRS fulfilled and blood culture drawn and 1 abnormal vital (time windows not stated)	Relative onset matching	Entire EHR cohort included
7 Kaji et al. (18)	Offline training, horizon evaluation	Sepsis-2 related: SIRS criteria plus ICD-9 code consistent with infection	Fixed length of 14 days in ICU (truncation if longer, zero filling, and masking if shorter)	Individual patient ICU admissions 2 days or longer were identified
8 Kam and Kim (30)	Offline training, horizon evaluation	Sepsis-2 related: ICD-9 code 995.9 and the first 5-h persisting window of fulfilled SIRS	insufficient detail: during training, 5-h windows are randomly extracted from case before sepsis and entire control stay, during testing it is not stated which data are used for controls	Medical ICU, age >18 years, patient can be checked for 5-h SIRS window plus ICD-9 995.9 code (if only one of the two was available, patients were excluded)
9 Lauritsen et al. (31)	Offline training, horizon evaluation	Sepsis-2 related: SIRS criteria plus clinically suspected infection	Random onset matching (excluding the first and last 3 h)	Inpatients, admissions ≥3 h, hospital departments with sepsis prevalence ≥2%, ≥1 observations for each vital sign measurement
10 Lukaszewski et al. (32)	Offline training, offline evaluation (fixed 24-h horizon)	Sepsis-2 related: SIRS criteria plus positive microbiological culture	Insufficient detail (but age-matching between cases and controls; healthy volunteers used as controls)	Blood samples taken daily; last sample on day of diagnosis or last stay in ICU
11 Mao et al. (33)	Offline training, offline evaluation (single fixed 4-h horizon)	Sepsis-2 related (suspected infection and first hour of fulfilled SIRS criteria), Severe Sepsis: ICD-9 plus SIRS plus organ dysfunction criteria; Septic Shock: ICD-9 plus manually defined conditions	–	Inpatients, age ≥18 years, ≥1 observations for each vital sign measurement, prediction time between 7 and 2,000 h
12 McCoy and Das (16)	Offline training, evaluation on retrospective dataset, prospective evaluation implemented as risk score	Sepsis-3, Severe Sepsis (SIRS criteria plus 2 organ dysfunction lab values)	–	Age >18 years; two or more sirs criteria during stay (hard to tell "Patient encounters were included in the sepsis-related outcome metrics if they met two or more SIRS criteria at some point during their stay." Is this an inclusion criterion or their label definition?)
13 Moor et al. (21)	Offline training, horizon evaluation	Sepsis-3	Absolute onset matching	Age ≥15 years, chart data including ICU admission/discharge time available, Metavision logging, cases: onset at least 7 h into ICU stay

(Continued)

TABLE 2 | Continued

References	Prediction task	Sepsis definition	Case–control alignment	Inclusion criteria
14Nemati et al. (34)	Offline training, horizon evaluation	Sepsis-3 (with modified time windows)	–	Age ≥18 years; sepsis onset not earlier than 4 h within ICU admission
15Reyna et al. (35)	Online training, online evaluation	Sepsis-3 (with modified time windows)	–	≥8 h of measurements
16Schamoni et al. (36)	Offline training, horizon evaluation as well as prediction of severity (ordinal regression)	Sepsis tag by ICU clinicians via electronic questionnaire	–	Sepsis onset not earlier than on the second day after ICU admission
17Scherpf et al. (37)	Offline training, horizon evaluation	Sepsis-2 related: ICD-9 codes plus SIRS criteria	Random onset matching via drawing fixed size time windows	Age ≥18 years, at least one measurement for SIRS parameters, no sepsis on admission, at least 5 h plus prediction time of measurements
18Shashikumar et al. (38)	Offline training, Offline prediction (single fixed 4-h horizon)	Sepsis-3	–	–
19Shashikumar et al. (39)	Offline training, Offline prediction (single fixed 4-h horizon)	Sepsis-3	–	–
20Sheetrit et al. (40)	Offline training, horizon evaluation on two prediction windows (12 and 1 h)	Sepsis-2 related: ICD-9 Codes 995.91 or 995.92 plus antibiotics administered. Onset time is defined as the earliest of either antibiotics prescription or fulfilled qSOFA criteria	Insufficient detail: the paper uses the "equivalent time" as the feature window of the control group	ICU admission, age ≥15 years, for sepsis cases: onset not before third day
21van Wyk et al. (41)	Offline training, horizon evaluation	Sepsis-2 related: SIRS criteria plus suspicion of infection, indicated by the presence of a blood culture and the administration of antibiotics during the encounter, along with relevant ICD10	Insufficient detail: the paper uses "a given 6-h observational period" for the control group	At least 8 h of continuous data, absence of cardiovascular disease
22van Wyk et al. (42)	Offline training, horizon evaluation	Sepsis-2 related: SIRS criteria plus suspicion of infection, indicated by the presence of a blood culture and the administration of antibiotics during the encounter, along with relevant ICD10	Insufficient detail: the paper uses "a given 3-h observational period" for the control group	Age >18 years, physiological data available for at least 3 or 6 h, respectively; absence of cardiovascular disease

Abbreviations: EHR, electronic health record; ICD-9, International Classification of Disease Version 9; ICU, intensive care unit; qSOFA, quick Sequential Organ Failure Assessment; SIRS, Systemic Inflammatory Response Syndrome.

≤ 40% of the quality criteria) to very good (satisfying ≥ 90% of the quality criteria). None of the studies fulfilled all 14 criteria. A single criterion was met by 100% of the studies: all studies highlighted the limits in current non-machine-learning approaches in the introduction. Few studies provided the code used for the data cleaning and analysis ($n = 2$; 9.5%), provided data or code for the reproduction of the exact sepsis labels and onset times ($n = 2$; 9.5%), and validated the machine learning models on an external dataset ($n = 3$; 14.3%). For the interpretation, power, and validity of machine learning methods, considerable sample sizes are required. With the exception of one study (32), all studies had sample sizes larger than 50 sepsis patients.

4. DISCUSSION

In this study, we systematically reviewed the literature for studies employing machine learning algorithms to facilitate early prediction of sepsis. A total of 22 studies were deemed eligible

for the review and 21 were included in the quality assessment. The majority of the studies used data from the MIMIC-III database (13), containing deidentified health data associated with ≈ 60,000 ICU admissions and/or data from Emory University Hospital[2]). With the exception of one, *all* studies used internationally acknowledged guidelines for sepsis definitions, namely Sepsis-2 (51) and Sepsis-3 (1). In terms of the analysis, a wide range of machine learning algorithms were chosen to leverage the patients' digital health data for the prediction of sepsis. Driven by our findings from the reviewed studies, this section first highlights four major challenges that the literature on machine learning driven sepsis prediction is currently facing: (i) asynchronicity, (ii) comparability, (iii) reproducibility, and (iv) circularity. We then discuss the limitations of this study,

[2]The dataset was not publicly available. However, with the 2019 PhysioNet Computing in Cardiology Challenge, a pre-processed dataset from Emory University Hospital has been published (35).

provide some recommendations for forthcoming studies, and conclude with an outlook.

4.1. Asynchronicity

While initial studies employing machine learning for the prediction of sepsis have demonstrated promising results (28–30), the literature since has been diverging on which are the most pressing open challenges that need to be addressed to further the goal of early sepsis detection. On the one hand, corporations have been propelling the deployment of the first interventional studies (52, 53), while on the other hand, recent findings have cast doubt on the validity and meaningfulness of the experimental pipeline that is currently being implemented in most retrospective analyses (36). This can be partially attributed to circular prediction settings (for more details, please refer to section 4.4). Ultimately, only the demonstration of favorable outcomes in large prospective randomized controlled trials (RCTs) will pave the way for machine learning models entering the clinical routine. Nevertheless, not every possible choice of model architecture can be tested prospectively due to the restricted sample sizes (and therefore, number of study arms). Rather, the development of these models is generally assumed to occur retrospectively. However, precisely those retrospective studies are facing multiple obstacles, which we are going to discuss next.

FIGURE 3 | (A) Offline training scenario and case–control matching. Every case has a specific sepsis onset. Given a random control, there are multiple ways of determining a matched onset time: (i) *relative* refers to the relative time since intensive care unit (ICU) admission (here, 75% of the ICU stay); (ii) *absolute* refers to the absolute time since ICU admission; (iii) *random* refers to a pseudo-random time during the ICU stay, often with the requirement that the onset is not too close to ICU discharge. **(B)** Horizon evaluation scenario. Given a case and control, with a matched relative sepsis onset, the look-back horizon indicates how *early* a specific model is capable of predicting sepsis. As the (matched) sepsis onset is approached, this task typically becomes progressively easier. Notice the difference in the prediction targets (labels) (shown in red for predicting a case, and blue for predicting a control).

4.2. Comparability

Concerning the comparability of the reviewed studies, we note that there are several challenges that have yet to be overcome, namely the choice of (i) prediction task, (ii) case–control onset matching, (iii) sepsis definition, (iv) implementation of a given sepsis definition, and (v) performance measures. We subsequently discuss each of these challenges.

4.2.1. Prediction Task

As described in section 3.5, we found that the vast majority of the included papers follow one of two major approaches when implementing the sepsis onset prediction task: Either an offline training step was followed by a horizon evaluation, or both the training and the evaluation were conducted in an online fashion. As one of our core findings, we next highlight the strengths but also the intricacies of these two setups. Considering the most frequently used strategy, i.e., offline training plus horizon evaluation, we found that the horizon evaluation provides valuable information about how early (in hours before sepsis onset) the machine learning model is able to recognize sepsis. However, in order to train such a classifier, the choice of a meaningful time window (and matched onset) for controls is an essential aspect of the study design (for more details, please refer to section 4.2.2). By contrast, the online strategy does not require a matched onset for controls (see **Figure 4**), but it removes the convenience of easily estimating predictive performance for a given prediction horizon (i.e., in hours before sepsis onset). Nevertheless, models trained and evaluated in an online fashion may be more easily deployed in practice, as they are by construction optimized for continuously predicting sepsis while new data arrive. Meanwhile, in the offline setting, the entire classification task is retrospective because all input data are extracted right up until a previously known sepsis onset. Whether a model trained this way would generalize to a prospective setup in terms of predicting sepsis *early* remains to be analyzed in forthcoming studies. In this review, the only study featuring prospective analysis focused on (and improved) prospective targets other than sepsis onset, namely mortality, length of stay, and hospital readmission. Finally, we observed that the online setting also contains a non-obvious design choice, which is absent in the offline/horizon approach: How many hours *before* and *after* a sepsis onset should a positive prediction be considered a true positive or rather a false positive? In other words, how long before or after the onset should a model be incentivized to raise an alarm for sepsis? Reyna et al. (35) proposed a clinical utility score that customizes a clinically motivated reward system for a given positive or negative prediction with respect to a potential sepsis onset. For example, it reflects that late true positive predictions are of little to no clinical importance, whereas late false negatives predictions can indeed be harmful. While such a hand-crafted score may account for a clinician's diagnostic demands, the resulting score remains highly sensitive to the exact specifications for which there is currently neither an internationally accepted standard nor a consensus. Furthermore, in its current form, the proposed clinical utility score is hard to interpret.
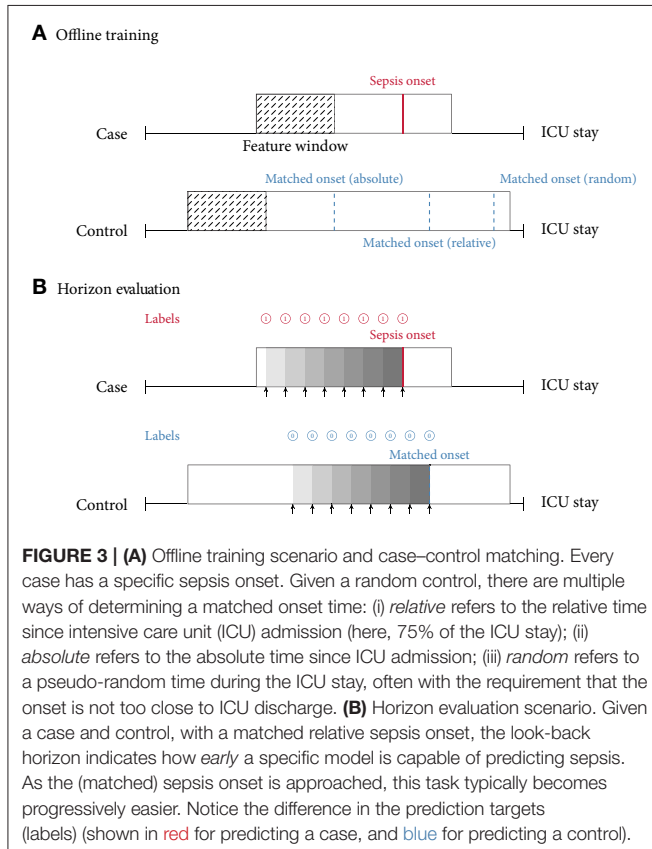
4.2.2. Case–Control Onset Matching

Futoma et al. (19) observed a drastic drop in performance upon introducing their (relative) case–control onset matching scheme as compared to an earlier version of their study, where the classification scenario compares sepsis onsets with the discharge time of controls (50). Such a matching can be seen as an implicit onset matching, which studies that do not account for this issue tend to default to. This suggests that comparing the data distribution of patients at the time of sepsis onset with the one of controls when being discharged could systematically underestimate the difficulty of the relevant clinical task at hand, i.e., identifying sepsis in an ICU stay. Futoma et al. (19) also remarked that "for non-septic patients, it is not very clinically relevant to include all data up until discharge, and compare predictions about septic encounters shortly before sepsis with predictions about non-septic encounters shortly before discharge. This task would be too easy, as the controls before discharge are likely to be clinically stable." The choice of a matched onset time is therefore crucial and highlights the need for a more uniform reporting procedure of this aspect in the literature. Furthermore, Moor et al. (21) proposed to match the *absolute* sepsis onset time (i.e., perform absolute onset matching) to prevent biases that could arise from systematic differences in the length of stay distribution of sepsis cases and controls (in the worst case, a model could merely re-iterate that one class has shorter stays than the other one, rather than pick up an actual signal in their time series). Finally, **Table 2** lists four studies that employed random onset matching. Given that sepsis onsets are not uniformly distributed over the length the ICU stay (for more details, please refer to section 4.4), this strategy could result in overly distinct data distributions between sepsis cases and non-septic controls.

4.2.3. Defining and Implementing Sepsis

A heterogeneous set of existing definitions (and modifications thereof) was implemented in the reviewed studies. The choice of sepsis definition will affect studies in terms of the prevalence of patients with sepsis and the level of difficulty of the prediction task (due to assigning earlier or later sepsis onset times). We

note that it remains challenging to fully disentangle all of these factors: on the one side, a larger absolute count of septic patients is expected to be beneficial for training machine learning models (in particular deep neural networks). On the other side, including more patients could make the resulting sepsis cohort a less severe one and harder to distinguish from non-septic ICU patients. Then again, a more inclusive sepsis labeling would result in a higher prevalence (i.e., class balance), which would be beneficial for the training stability of machine learning models. To further illustrate the difficulty of defining sepsis, consider the prediction target *in-hospital mortality*. Even though in-hospital mortality rates (and therefore any subsequent prediction task) vary between cohorts and hospitals, their *definition* typically does not. Sepsis, by contrast, is inherently hard to define, which over the years has led to multiple refinements of clinical criteria (Sepsis 1–3) for trying to capture sepsis in one easy-to-follow, rule-based definition (1, 51, 54). It has been previously shown that applying different sepsis definitions to the same dataset results in largely dissimilar cohorts (55). Furthermore, this specific study found that using Sepsis-3 is too inclusive, resulting in a large cohort showing mild symptoms. By contrast, practitioners have reported that Sepsis-3 is indeed too *restrictive* in that sepsis cannot occur without organ dysfunction any more (55). This suggests that even *within* a specific definition of sepsis, substantial heterogeneity and disagreement in the literature prevails. On top of that, we found that even applying the same definition on the same dataset has resulted in dissimilar cohorts. Most prominently, in **Table 1**, this can be confirmed for studies employing the MIMIC-III dataset. However, the determining factors cannot be easily recovered, as the code for assigning the labels is not available in 19 out of 21 (90.4%) studies employing computer-derived sepsis labels.

Another factor exacerbating comparability is the heterogeneous sepsis prevalence. This is partially influenced by the training setup of a given study, because certain studies prefer balanced datasets for improving the training stability of the machine learning model (27, 41, 42), while others preserve the observed case counts to more realistically reflect how their approach would fare when being deployed in ICU. Furthermore,
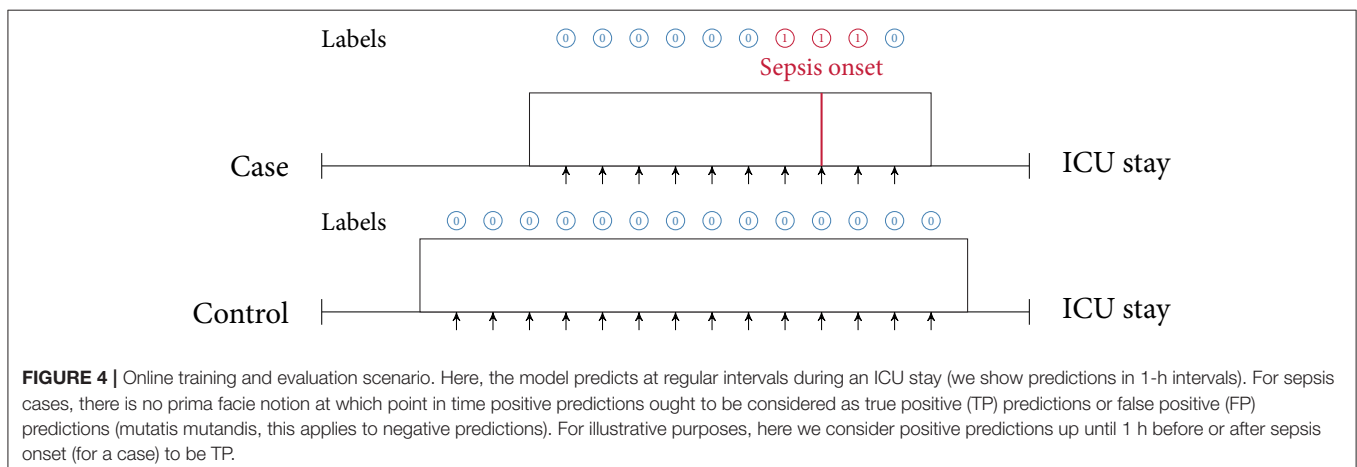

FIGURE 4 | Online training and evaluation scenario. Here, the model predicts at regular intervals during an ICU stay (we show predictions in 1-h intervals). For sepsis cases, there is no prima facie notion at which point in time positive predictions ought to be considered as true positive (TP) predictions or false positive (FP) predictions (mutatis mutandis, this applies to negative predictions). For illustrative purposes, here we consider positive predictions up until 1 h before or after sepsis onset (for a case) to be TP.

TABLE 3 | Quality assessment of all studies.

	Study	Limits of current approaches	Prevalence of sepsis reported	Data availability	Feature engineering methods	Code for analysis	Code for label generation	Platforms/packages reported	Hyperparameters reported	Sample size > 50	Valid methods to prevent overfitting	Stability of results reported	External data validation	Explanation of predictors	Suggested clinical use	Total
1	Abromavičius et al. (26)	✓	✓	✓	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	50%
2	Barton et al. (17)	✓	✓	✗	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓	✓	57%
3	Bloch et al. (27)	✓	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	71%
4	Calvert et al. (28)	✓	✓	✗	✓	✗	✗	✗	✗	✓	✓	✓	✗	✗	✓	43%
5	Desautels et al. (29)	✓	✓	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓	✗	✓	50%
6	Futoma et al. (19)	✓	✓	✗	✓	✗	✗	✗	✓	✓	✓	✗	✗	✗	✓	50%
7	Kaji et al. (18)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	93%
8	Kam and Kim (30)	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗	36%
9	Lauritsen et al. (31)	✓	✗	✗	✓	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓	57%
10	Lukaszewski et al. (32)	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	43%
11	Mao et al. (33)	✓	✓	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓	✗	✓	64%
12	McCoy and Das (16)	✓	✓	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	✗	✓	36%
13	Moor et al. (21)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	93%
14	Nemati et al. (34)	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✗	✓	50%
15	Schamoni et al. (36)	✓	✓	✗	✓	✗	✗	✗	✗	✓	✓	✓	✗	✓	✓	57%
16	Scherpf et al. (37)	✓	✓	✗	✓	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗	43%
17	Shashikumar et al. (38)	✓	✓	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✓	✓	50%
18	Shashikumar et al. (39)	✓	✓	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✓	✓	50%
19	Sheetrit et al. (40)	✓	✓	✗	✓	✗	✗	✓	✗	✓	✓	✓	✗	✗	✗	43%
20	van Wyk et al. (41)	✓	✓	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	✗	✓	36%
21	van Wyk et al. (42)	✓	✓	✗	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✓	43%
		100%	95%	19%	81%	10%	10%	19%	29%	95%	81%	62%	14%	38%	86%	
Study		Unmet need	Reproducibility							Stability			Generalizability	Clinical significance		Total

We excluded Reyna et al. (35) from the assessment because it does presents a dataset challenge rather than a single method, making most of the categories not applicable.
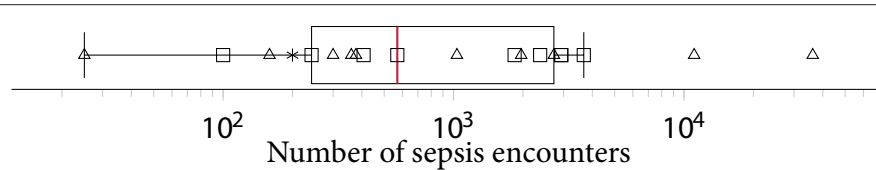
FIGURE 5 | A boxplot of the number of sepsis encounters reported by all studies, with the median number of encounters being highlighted in red. Since the numbers feature different orders of magnitude, we employed logarithmic scaling. The marks indicate which definition or modification thereof was used. Sepsis-3: squares, Sepsis-2: triangles, domain expert label: asterisk.

the exact sepsis definition used as well as the applied data pre-processing and filtering steps influence the resulting sepsis case count and therefore the prevalence (21, 55). **Figure 2** depicts a boxplot of the prevalence values of all studies. Of the 22 studies, 10 report prevalences $\leq$ 10%, with the maximum reported prevalence being 63.6% (18). In addition, **Figure 5** depicts the distribution of all sepsis encounters, while also encoding the sepsis definition (or modification thereof) that is being used.

4.2.4. Performance Measures
The last obstacle impeding comparability is the choice of performance measures. This is entangled with the differences in sepsis prevalence: simple metrics, such as accuracy are directly impacted by class prevalence, rendering a comparison of two studies with different prevalence values moot. Some studies report the area under the receiver operating characteristic curve (AUROC, sometimes also reported as AUC). However, AUROC also depends on class prevalence and is known to be less informative if the classes are highly imbalanced (56, 57). The area under the precision–recall curve (AUPRC, sometimes also referred to as average precision) should be reported in such cases, and we observed that $n = 6$ studies already do so. AUPRC is also affected by prevalence but permits a comparison with a random baseline that merely "guesses" the label of a patient. AUROC, by contrast, can be high even for classifiers that fail to properly classify the minority class of sepsis patients. This effect is exacerbated with increasing class imbalance. Recent research suggests reporting the AUPRC of models, in particular in clinical contexts (58), and we endorse this recommendation.

4.2.5. Comparing Studies of Low Comparability
Our findings indicate that quantitatively comparing studies concerned with machine learning for the prediction of sepsis in the ICU is currently a nigh-impossible task. While one would like to perform meta-analyses in these contexts to aggregate an overall trend in performance among state-of-the-art models, at the current stage of the literature this would carry little meaning. Therefore, we currently cannot ascertain the best performing approaches by merely assessing numeric results of performance measures. Rather, we had to resort to *qualitatively* assess study designs in order identify underlying biases, which could lead to overly optimistic results.

4.3. Reproducibility
Reproducibility, i.e., the capability of obtaining similar or identical results by independently repeating the experiments

described in a study, is the foundation of scientific accountability. In recent years, this foundation has been shaken by the discovery of failures to reproduce prominent studies in several disciplines (59). Machine learning in general is no exception here, and despite the existence of calls to action (60), the field might face a reproducibility crisis (61). The interdisciplinary nature of digital medicine comes with additional challenges for reproducibility (62), foremost of which is the issue of dealing with sensitive data (whereas for many theoretical machine learning papers, benchmark datasets exist), but also the issue of algorithmic details, such as pre-processing. Our quality assessment highlights a lot of potential for improvement here: only two studies (18, 21), both from 2019, share their analysis code and the code for generating a "label" (to distinguish between cases or controls within the scenario of a specific paper). This amounts to < 10% of the eligible studies. In addition, only four studies (18, 21, 26, 33) report results on publicly available datasets (more precisely, the datasets are available for research after accepting their terms and conditions). This finding is surprising, given the existence of high-quality, freely accessible databases, such as MIMIC-III (13) or eICU (63). An encouraging finding of our analysis is that a considerable number of studies ($n = 6$) report hyperparameter details of their models. Hyperparameter refers to any kind of parameter that is model specific, such as the regularization constant and the architecture of a neural network (64). This information is crucial for everyone who attempts to reproduce computational experiments.

4.4. Circularity
Considering that the exact sepsis onset is usually unknown, most of the existing works have approximated a plausible sepsis onset via clinical criteria, such as Sepsis-3 (1). However, these criteria comprise a set of rules to apply to vital and laboratory measurements. Schamoni et al. (36) pointed out that using clinical measurements for predicting a sepsis label, which was itself derived from clinical measurements, could potentially be circular (a statistical term referring to the fact that one uses the same data for the selection of a model and its subsequent analysis). This runs the risk being unable to discover unknown aspects of the data, since classifiers may just confirm existing criteria instead of helping to generate new knowledge. In the worst case, a classifier would merely reiterate the guidelines used to define sepsis *without* being able to detect patterns that permit an earlier discovery. To account for this, Schamoni et al. chose a questionnaire-based definition of sepsis and clinical experts

manually labeled the cases and controls. While this strategy may reduce the problem of circularity, a coherent and comprehensive definition of sepsis cannot be easily guaranteed. Notably, Schamoni et al. (36) report very high inter-rater agreement. They assign, however, only *daily* labels, which is in contrast to automated Sepsis-3 labels that are typically extracted in an hourly resolution. Furthermore, it is plausible that even with clinical experts in the loop, some level of (indirect) circularity could still take place, because a clinician would also consult the patients' vital and laboratory measurements in order to assign the sepsis tag, it would merely be less explicit. Since Schamoni et al. (36) proposed a way to circumvent the issue of circularity, this also means that no existing work has empirically assessed the existence (or the relevance) of circularity in machine learning-based sepsis prediction. For Sepsis-3, if the standard 72-h window is used for assessing an increase in SOFA (sequential organ failure assessment score) score, i.e., starting 48 h before suspected infection time until 24 h afterwards, and if the onset happens to occur at the very end of this window, then measurements that go 72 h into the past have influenced this label. Since the SOFA score aggregates the most abnormal measurements of the preceding 24 h (65), Sepsis-3 could even "reach" 96 h into the past. Meanwhile, the distribution of onsets using Sepsis-3 tends to be highly right-skewed, as can be seen in Moor et al. (21), where removing cases with an onset during the first 7 h drastically reduced the resulting cohort size. Therefore, we conjecture that with Sepsis-3, it could be virtually impossible to strictly separate data that are used for assigning the label from data that are used for prediction, without overly reducing the resulting cohort. Finally, the relevance of an ongoing circularity may be challenged given first promising results (in terms of mortality reduction) of the first interventional studies applying machine learning for sepsis prediction prospectively (52), without explicitly accounting for circularity.

4.5. Limitations of This Study

A limitation of this review is that our literature search was restricted to articles listed in Embase, Google Scholar, PubMed/Medline, Scopus, and Web of Science. Considering the pace at which the research in this area—in particular, in the context of machine learning—is moving forward, it is likely that the findings of the publications described in this paper will be quickly complemented by further research. The literature search also excluded gray literature (e.g., preprints and reports), the importance of which to this topic is unknown[3], and thus might have introduced another source of search bias. The lack of studies reporting poor performance of machine learning algorithms regarding sepsis onset prediction suggests high probability of publication bias (66, 67). Publication bias is likely to result in studies with more positive results being preferentially submitted and accepted for publication (68). Finally, our review specifically focused on machine learning applications for the prediction of sepsis and severe sepsis. We therefore used a stringent search term that potentially excluded studies pursuing a classical statistical approach of early detection and sepsis prediction.

[3]In the machine learning community, for example, it is common practice to use preprints to disseminate knowledge about novel methods early on.

5. RECOMMENDATIONS

This section provides recommendations how to harmonize experimental designs and reporting of machine learning approaches for the early prediction of sepsis in the ICU. This harmonization is necessary to warrant meaningful comparability and reproducibility of different machine learning models, ensure continued model development as opposed to starting from scratch, and establish benchmark models that constitute the state-of-the-art.

As outlined above, only few studies score highly with respect to reproducibility. This is concerning, as reproducibility remains one of the cornerstones of scientific progress (62). The lack of comparability of different studies impedes progress because a priori, it may not be clear which method is suitable for a specific scenario if different studies lack common ground (see also the aforementioned issues preventing a meta-analysis). The way out of this dilemma is to improve reproducibility of a *subset* of a given study. We suggest the following approach: (i) picking an openly available dataset (or a subset thereof) as an additional validation site, (ii) reporting results on this dataset, and (iii) making the code for this analysis available (including models and labels). This suggestion is flexible and still enables authors to showcase their work on their respective private datasets. We suggest that code sharing—within reasonable bounds—should become the *default* for publications as modern machine learning research is increasingly driven by implementations of complex algorithms. Therefore, a prerequisite of being able to replicate the results of any study, or to use it in a comparative setting, is having access to the raw code that was used to perform the experiment. This is crucial, as any pseudocode description of an algorithm permits many different implementations with potentially different runtime behavior and side effects. With only two studies sharing code, method development is stymied. We thus encourage authors to consider sharing their code, for example via platforms, such as GitHub (https://github.com). Even sharing only parts of the code, such as the label generation process, would be helpful in many scenarios and improve comparability. The availability of numerous open source licenses (70) makes it possible to satisfy the constraints of most authors, including companies that want to protect their intellectual property. A recent experiment at the International Conference of Machine Learning (ICML) demonstrated that reviewers and area chairs react favorably to the inclusion of code (71). If code sharing is *not* possible, for example because of commercial interests, there is the option to share binaries, possibly using virtual machines or "containers" (72). Providing containers would satisfy all involved parties: intellectual property rights are retained but additional studies can compare their results.

As for the datasets used in a study, different rules apply. While some authors suggest that peer-reviewed publications should be come with a waiver agreement for open access data (73), we are aware of the complications of sharing clinical data. We think that a reasonable middle ground can be reached by following the suggestion above, i.e., using existing benchmark datasets, such as MIMIC-III (13) to report performance.

BOX 1 | Recommendations for the practitioner.

Recommendation	Remarks	Details
Make code publicly available or usable	A prerequisite of being able to replicate the results of any study, or to use any model in a comparative setting, is having access to the raw code or a binary variant thereof that was used to perform the experiments. Authors are encouraged to share their code, for example via platforms, such as GitHub, or their binaries using container technologies like Docker.	GitHub, Docker
Use external validation for the machine learning model	External validation of a classifier is crucial for assessing the model's generalizability. Several publicly available data sources exist that can be used for this purpose.	MIMIC-II, MIMIC-III, eICU, HiRID
Provide exact definition of sepsis label	Implementations vary drastically in terms of prevalence and number of sepsis encounters. Thus, reporting the label generation process is essential, particularly when labels deviate from the international definitions of sepsis. For instance, when using the eICU dataset, microbiology measurements are under-reported for defining suspected infection, yet the exact modifications of sepsis implementations have not explicitly been stated (69).	Provide code of how sepsis label was determined.
Provide an detailed description of a control and, if applicable, its matched onset	While there is a defined point in time for an event in the sepsis cohorts, it is much more challenging to determine at what time to extract data for a control case when was the non-event. For transparency and replication reasons, it is crucial to provide details on how controls were defined and how the onset was determined.	Provide code of how a control was defined and, if applicable, its matched onset was determined.
Make data available	If possible and in compliance with international data protection laws, data sources should be made accessible to bona fide researchers. There are multiple data repositories, which researchers can use to make their data accessible, while complying with data protection laws.	Harvard Dataverse, PhysioNet, Zenodo
Ensure comparability of models and their performances	To advance the field, it is important that researchers compare their models to existing models in order to evaluate and compare the performance across different studies. This necessitates improvements in prevalence reporting as well as the choice of different performance metrics.	Report prevalence and AUPRC in addition to other metrics.
Use licenses for code	Licenses protect the creators and the users of code. Numerous open source licenses exist, making it possible to satisfy the constraints of most authors, including companies that want to protect their intellectual property.	Apache license, BSD licenses, GPL

Moreover, we urge authors to report additional details of their experimental setup, specifically the selection of cases and controls and the label generation/calculation process. As outlined above, the case–control matching is crucial as it affects the difficulty (and thus the significance) of the prediction task. We suggest to either follow the absolute onset matching procedure (21), which is simple to implement and prevents biases caused by differences in the length of stay distribution. In any case, forthcoming work should always report their choice of case–control matching. As for the actual prediction task, given the heterogeneous prediction horizons that we observed, we suggest that authors always report performance for a horizon of 3 h or 4 h (in addition to any other performance metrics that are reported). This reporting should always use the AUPRC metric as it is the preferred metric for rare prevalences (74). Last, we want to stress that a description of the inclusion process of patients is essential in order to ensure comparability.

6. CONCLUSIONS AND FUTURE DIRECTIONS

This study performed a systematic review of publications discussing the early prediction of sepsis in the ICU by means of machine learning algorithms. Briefly, we found that the majority

of the included papers investigating sepsis onset prediction in the ICU are based on data from the same center, MIMIC-II or MIMIC-III (13), two versions of a high-quality, publicly available critical care database. Despite the data agreement guidelines of MIMIC-III stating that code using MIMIC-III needs to be published (paragraph 9 of the current agreement reads "If I openly disseminate my results, I will also contribute the code used to produce those results to a repository that is open to the research community."), only two studies (18, 21) make their code available. This leaves a lot of room for improvement, which is why we recommend code (or binary) sharing (**Box 1**). Of 22 included studies, only *one* reflects a non-Western (i.e., neither North-American nor European) cohort, pinpointing toward a significant dataset bias in the literature (see **Supplementary Table 4** for an overview of demographical information). In addition to demographic aspects, such as ethnicity, differing diagnostic, and therapeutic policies as well as the availability of input data for prediction are known to impact the generation of the sepsis labels. This challenge hampers additional benchmarking efforts unless more diverse cohorts are included. Moreover, since the prediction task is highly sensitive to minor changes in study specification (including, but not limited to, the sepsis definition and the case–control alignment), the majority of investigated papers do not permit a straightforward reproduction/replication and comparison of their employed cohorts and their presented prediction task. Meta-analyses are therefore impossible, as the reported metrics pertain to different, incomparable scenarios: both prevalence and case counts are highly variable, even on the same dataset, and previous work (19) indicated that minor changes in the experimental setup can substantially affect the difficulty of the prediction task. As a consequence, we are currently not able to identify the most predictive method for recognizing sepsis early, which then ought to be further investigated in prospective trials. All in all, we found this state of the art to leave lots of room for improvement; it would be beneficial to be able to compare different models as to their generalizability, in particular when deploying machine learning algorithms in a prospective study. We see our paper as a "call to arms" for the community and hope that our recommendations are taken in the spirit of improving this task together.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://github.com/BorgwardtLab/sepsis-prediction-review.

AUTHOR CONTRIBUTIONS

MM, BR, and CJ contributed substantially to the data acquisition, extraction, analysis (i.e., quality assessment), and interpretation. Furthermore, they drafted the review article. MH made a substantial contributions to data interpretation (i.e., quality assessment) and participated in revising the review article critically for important intellectual content. KB made a significant contributions to the study conception and revised the review article critically for important intellectual content. All authors contributed to the article and approved the submitted version.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2021.607952/full#supplementary-material

REFERENCES

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*. (2016) 315:801–10. doi: 10.1001/jama.2016.0287

2. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *Lancet*. (2020) 395:200–11. doi: 10.1016/S0140-6736(19)32989-7

3. Dellinger RP, Levy MM, Rhodes A, Annane D, Gerlach H, Opal SM, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock 2012. *Crit Care Med*. (2013) 41:580–637. doi: 10.1097/CCM.0b013e31827e83af

4. Kaukonen KM, Bailey M, Suzuki S, Pilcher D, Bellomo R. Mortality related to severe sepsis and septic shock among critically ill patients in Australia and New Zealand, 2000–2012. *J Am Med Assoc*. (2014) 311:1308–16. doi: 10.1001/jama.2014.2637

5. Hotchkiss RS, Moldawer LL, Opal SM, Reinhart K, Turnbull IR, Vincent JL. Sepsis and septic shock. *Nat Rev Dis Primers*. (2016) 2:16045. doi: 10.1038/nrdp.2016.45

6. Ferrer R, Martin-Loeches I, Phillips G, Osborn TM, Townsend S, Dellinger RP, et al. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. *Crit Care Med*. (2014) 42:1749–55. doi: 10.1097/CCM.0000000000000330

7. Weiss SL, Fitzgerald JC, Balamuth F, Alpern ER, Lavelle J, Chilutti M, et al. Delayed antimicrobial therapy increases mortality and organ

dysfunction duration in pediatric sepsis. *Crit Care Med*. (2014) 42:2409. doi: 10.1097/CCM.0000000000000509

8. Pruinelli L, Westra BL, Yadav P, Hoff A, Steinbach M, Kumar V, et al. Delay within the 3-hour surviving sepsis campaign guideline on mortality for patients with severe sepsis and septic shock. *Crit Care Med*. (2018) 46:500. doi: 10.1097/CCM.0000000000002949

9. Lever A, Mackenzie I. Sepsis: definition, epidemiology, and diagnosis. *BMJ*. (2007) 335:879–883. doi: 10.1136/bmj.39346.495880.AE

10. Al Jalbout N, Troncoso Jr R, Evans JD, Rothman RE, Hinson JS. Biomarkers and molecular diagnostics for early detection and targeted management of sepsis and septic shock in the emergency department. *J Appl Lab Med*. (2019) 3:724–9. doi: 10.1373/jalm.2018.027425

11. Parlato M, Philippart F, Rouquette A, Moucadel V, Puchois V, Blein S, et al. Circulating biomarkers may be unable to detect infection at the early phase of sepsis in ICU patients: the CAPTAIN prospective multicenter cohort study. *Intensive Care Med*. (2018) 44:1061–70. doi: 10.1007/s00134-018-5228-3

12. Faix JD. Biomarkers of sepsis. *Crit Rev Clin Lab Sci*. (2013) 50:23–36. doi: 10.3109/10408363.2013.764490

13. Johnson AE, Pollard TJ, Shen L, Li-Wei HL, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. (2016). doi: 10.1038/sdata.2016.35

14. Fleuren LM, Klausch TL, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med*. (2020) 46:383–400. doi: 10.1007/s00134-019-05872-y

15. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit Health*. (2020) 2:179–91. doi: 10.1016/S2589-7500(20)30018-2

16. McCoy A, Das R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ Open Qual*. (2017) 6:e000158. doi: 10.1136/bmjoq-2017-000158

17. Barton C, Chettipally U, Zhou Y, Jiang Z, Lynn-Palevsky A, Le S, et al. Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Comput Biol Med*. (2019) 109:79–84. doi: 10.1016/j.compbiomed.2019.04.027

18. Kaji DA, Zech JR, Kim JS, Cho SK, Dangayach NS, Costa AB, et al. An attention based deep learning model of clinical events in the intensive care unit. *PLoS ONE*. (2019) 14:e0211057. doi: 10.1371/journal.pone.0211057

19. Futoma J, Hariharan S, Heller K, Sendak M, Brajer N, Clement M, et al. *An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection. Vol. 68 of Proceedings of Machine Learning Research*. PMLR (2017). p. 243–54.

20. Li SCX, Marlin BM. A scalable end-to-end gaussian process adapter for irregularly sampled time series classification. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R. editors. *Advances in Neural Information Processing Systems*. Curran Associates, Inc., (2016). 29, p. 1804–12.

21. Moor M, Horn M, Rieck B, Roqueiro D, Borgwardt K. *Early Recognition of Sepsis with Gaussian Process Temporal Convolutional Networks and Dynamic Time Warping. Vol. 106 of Proceedings of Machine Learning Research*. PMLR (2019). p. 2–26.

22. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. (2015) 4:1. doi: 10.1186/2046-4053-4-1

23. Harzing A. *Publish or Perish Software*. (2007). Available online at: https://harzingcom/resources/publish-or-perish

24. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med*. (2005) 37:360–3.

25. Qiao N. A systematic review on machine learning in sellar region diseases: quality and reporting items. *Endocr Connect*. (2019) 8:952–60. doi: 10.1530/EC-19-0156

26. Abromavičius V, Plonis D, Tarasevičius D, Serackis A. Two-stage monitoring of patients in intensive care unit for sepsis prediction using non-overfitted machine learning models. *Electronics*. (2020) 9:1133. doi: 10.3390/electronics9071133

27. Bloch E, Rotem T, Cohen J, Singer P, Aperstein Y. Machine learning models for analysis of vital signs dynamics: a case for sepsis onset prediction. *J Healthc Eng*. (2019) 2019:5930379. doi: 10.1155/2019/5930379

28. Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, et al. A computational approach to early sepsis detection. *Comput Biol Med*. (2016) 74:69–73. doi: 10.1016/j.compbiomed.2016.05.003

29. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med Inform*. (2016) 4:e28. doi: 10.2196/medinform.5909

30. Kam HJ, Kim HY. Learning representations for the early detection of sepsis with deep neural networks. *Comput Biol Med*. (2017) 89:248–55. doi: 10.1016/j.compbiomed.2017.08.015

31. Lauritsen SM, Kalør ME, Kongsgaard EL, Lauritsen KM, Jørgensen MJ, Lange J, et al. Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artif Intell Med*. (2020) 104:101820. doi: 10.1016/j.artmed.2020.101820

32. Lukaszewski RA, Yates AM, Jackson MC, Swingler K, Scherer JM, Simpson A, et al. Presymptomatic prediction of sepsis in intensive care unit patients. *Clin Vacc Immunol*. (2008) 15:1089–94. doi: 10.1128/CVI.00486-07

33. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open*. (2018) 8:e017833. doi: 10.1136/bmjopen-2017-017833

34. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med*. (2018) 46:547. doi: 10.1097/CCM.0000000000002936

35. Reyna MA, Josef C, Seyedi S, Jeter R, Shashikumar SP, Westover MB, et al. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Crit Care Med*. (2019). 48:210–217. doi: 10.1097/CCM.0000000000004145

36. Schamoni S, Lindner HA, Schneider-Lindner V, Thiel M, Riezler S. Leveraging implicit expert knowledge for non-circular machine learning in sepsis prediction. *Artif Intell Med*. (2019) 100:101725. doi: 10.1016/j.artmed.2019.101725

37. Scherpf M, Gräßer F, Malberg H, Zaunseder S. Predicting sepsis with a recurrent neural network using the MIMIC III database. *Comput Biol Med*. (2019) 113:103395. doi: 10.1016/j.compbiomed.2019.103395

38. Shashikumar SP, Li Q, Clifford GD, Nemati S. Multiscale network representation of physiological time series for early prediction of sepsis. *Physiol Meas*. (2017) 38:2235. doi: 10.1088/1361-6579/aa9772

39. Shashikumar SP, Stanley MD, Sadiq I, Li Q, Holder A, Clifford GD, et al. Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. *J Electrocardiol*. (2017) 50:739–43. doi: 10.1016/j.jelectrocard.2017.08.013

40. Sheetrit E, Nissim N, Klimov D, Shahar Y. Temporal probabilistic profiles for sepsis prediction in the ICU. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY (2019). p. 2961–9. doi: 10.1145/3292500.3330747

41. Van Wyk F, Khojandi A, Kamaleswaran R. Improving prediction performance using hierarchical analysis of real-time data: a sepsis case study. *IEEE J Biomed Health Inform*. (2019) 23:978–86. doi: 10.1109/JBHI.2019.2894570

42. van Wyk F, Khojandi A, Mohammed A, Begoli E, Davis RL, Kamaleswaran R. A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier. *Int J Med Inform*. (2019) 122:55–62. doi: 10.1016/j.ijmedinf.2018.12.002

43. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. (1997) 9:1735–80. doi: 10.1162/neco.1997.9.8.1735

44. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv*. (2014) 14061078. doi: 10.3115/v1/D14-1179

45. Fukushima K, Miyake S, Ito T. Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Trans Syst Man Cybernet*. (1983) 36:826–34. doi: 10.1109/TSMC.1983.6313076

46. Lea C, Flynn MD, Vidal R, Reiter A, Hager GD. Temporal convolutional networks for action segmentation and detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2017). p. 156–65. doi: 10.1109/CVPR.2017.113

47. Oord Avd, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, et al. Wavenet: a generative model for raw audio. *arXiv*. (2016) 160903499.

48. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY (2016). p. 785–94. doi: 10.1145/2939672.2939785

49. Kam HT. Random decision forest. In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition*. (1995). p. 278–82.

50. Futoma J, Hariharan S, Heller K. Learning to detect sepsis with a multitask Gaussian process RNN classifier. *arXiv*. (2017) 170604152.

51. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, et al. 2001 SCCM/ESICM/ACCP/ATS/SIS international sepsis definitions conference. *Intensive Care Med*. (2003) 29:530–8. doi: 10.1007/s00134-003-1662-x

52. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir Res*. (2017) 4:234. doi: 10.1136/bmjresp-2017-000234

53. Burdick H, Pino E, Gabel-Comeau D, McCoy A, Gu C, Roberts J, et al. Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. *BMJ Health Care Inform*. (2020) 27:e100109. doi: 10.1136/bmjhci-2019-100109

54. Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, Knaus WA, et al. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*. (1992) 101:1644–55. doi: 10.1378/chest.101.6.1644

55. Johnson AE, Aboab J, Raffa JD, Pollard TJ, Deliberato RO, Celi LA, et al. A comparative analysis of sepsis identification methods in an electronic database. *Crit Care Med*. (2018) 46:494. doi: 10.1097/CCM.0000000000002965

56. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. *Glob Ecol Biogeogr*. (2008) 17:145–51. doi: 10.1111/j.1466-8238.2007.00358.x

57. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*. (2015) 10:e0118432. doi: 10.1371/journal.pone.0118432

58. Pinker E. Reporting accuracy of rare event classifiers. *NPJ Digit Med*. (2018) 1:56. doi: 10.1038/s41746-018-0062-0

59. Baker M. 1,500 Scientists lift the lid on reproducibility. *Nature*. (2016) 533:452–4. doi: 10.1038/533452a

60. Crick T, Hall BA, Ishtiaq S. "Can i implement your algorithm?" A model for reproducible research software. *arXiv*. (2014) 1407.5981.

61. Hutson M. Artificial intelligence faces reproducibility crisis. *Science*. (2018) 359:725–6. doi: 10.1126/science.359.6377.725

62. Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. *NPJ Digit Med*. (2019) 2:2. doi: 10.1038/s41746-019-0079-z

63. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data*. (2018) 5:180178. doi: 10.1038/sdata.2018.178

64. Wu J, Chen XY, Zhang H, Xiong LD, Lei H, Deng SH. Hyperparameter optimization for machine learning models based on Bayesian optimization. *J Electron Sci Technol*. (2019) 17:26–40. doi: 10.11989/JEST.1674-862X.80904120

65. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. *The SOFA (Sepsis-related Organ Failure Assessment) Score to Describe Organ Dysfunction/Failure*. Heidelberg: Springer (1996). doi: 10.1007/s001340050156

66. Dickersin K, Chalmers I. Recognizing, investigating and dealing with incomplete and biased reporting of clinical research: from Francis Bacon to the WHO. *J R Soc Med*. (2011) 104:532–8. doi: 10.1258/jrsm.2011.11k042

67. Kirkham JJ, Altman DG, Williamson PR. Bias due to changes in specified outcomes during the systematic review process. *PLoS ONE*. (2010) 5:e9810. doi: 10.1371/journal.pone.0009810

68. Joober R, Schmitz N, Annable L, Boksa P. Publication bias: what are the challenges and can they be overcome? *J Psychiatry Neurosci*. (2012) 37:149. doi: 10.1503/jpn.120065

69. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. (2018) 24:1716–20. doi: 10.1038/s41591-018-0213-5

70. Rosen L. *Open Source Licensing: Software Freedom and Intellectual Property Law*. Upper Saddle River, NJ: Prentice Hall (2004).

71. Chaudhuri K, Salakhutdinov R. *The ICML 2019 Code-at-Submit-Time Experiment*. (2019). Available online at: https://medium.com/@kamalika_19878/the-icml-2019-code-at-submit-time-experiment-f73872c23c55

72. Elmenreich W, Moll P, Theuermann S, Lux M. Making computer science results reproducible–a case study using Gradle and Docker. *PeerJ*. (2018) 6:e27082v1. doi: 10.7287/peerj.preprints.27082v1

73. Hrynaszkiewicz I, Cockerill MJ. Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals. *BMC Res Notes*. (2012) 5:494. doi: 10.1186/1756-0500-5-494

74. Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J Clin Epidemiol*. (2015) 68:855–9. doi: 10.1016/j.jclinepi.2015.02.010

75. Moor M, Rieck B, Horn M, Jutzeler C, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: a systematic review. *medRxiv*. (2020). doi: 10.1101/2020.08.31.20185207