Title: Generalisability of AI-based scoring systems in the ICU: a systematic review and meta-analysis

Running title: Systematic review of ICU score generalisability

Authors: Patrick Rockenschaub[1,2,3*], Ela Marie Akay[1], Benjamin Gregory Carlisle[4], Adam Hilbert[1], Falk Meyer-Eschenbach[5], Anatol-Fiete Näher[5,6], Dietmar Frey[1], Vince Istvan Madai[2,7]

Affiliations:

[1] CLAIM - Charité Lab for AI in Medicine, Charité - Universitätsmedizin Berlin, Berlin, Germany

[2] QUEST Center for Transforming Biomedical Research, Berlin Institute of Health, Berlin, Germany

[3] Fraunhofer IKS, Fraunhofer Institute for Cognitive Systems, Munich, Germany

[4] STREAM - Studies of Translation, Ethics and Medicine, School of Population and Global Health, McGill University, 2001 McGill College Ave, Montréal QC H3A 1G1 Canada

[5] Institute of Medical Informatics, Charité - Universitätsmedizin Berlin, Germany

[6] Digital Engineering Faculty, Hasso Plattner Institute, University of Potsdam, Potsdam, Germany

[7] Faculty of Computing, Engineering and the Built Environment, School of Computing and Digital Technology, Birmingham City University, Birmingham, United Kingdom

* Corresponding author: Email: rockenschaub.patrick@gmail.com; Tel: +49 173 4937526

Keywords: intensive care unit; electronic health records; machine learning; acute deterioration; external validation

Word count: 3,885

Abstract

Background

Machine learning (ML) is increasingly used to predict clinical deterioration in intensive care unit (ICU) patients through scoring systems. Although promising, such algorithms often overfit their training cohort and perform worse at new hospitals. Thus, external validation is a critical – but frequently overlooked – step to establish the reliability of predicted risk scores to translate them into clinical practice. We systematically reviewed how regularly external validation of ML-based risk scores is performed and how their performance changed in external data.

Methods

We searched MEDLINE, Web of Science, and arXiv for studies using ML to predict deterioration of ICU patients from routine data. We included primary research published in English before April 2022. We summarised how many studies were externally validated, assessing differences over time, by outcome, and by data source. For validated studies, we evaluated the change in area under the receiver operating characteristic (AUROC) attributable to external validation using linear mixed-effects models.

Results

We included 355 studies, of which 39 (11.0%) were externally validated, increasing to 17.9% by 2022. Validated studies made disproportionate use of open-source data, with two well-known US datasets (MIMIC and eICU) accounting for 79.5% of studies. On average, AUROC was reduced by -0.037 (95% CI -0.064 to -0.017) in external data, with >0.05 reduction in 38.6% of studies.

Discussion

External validation, although increasing, remains uncommon. Performance was generally lower in external data, questioning the reliability of some recently proposed ML-based scores. Interpretation of the results was challenged by an overreliance on the same few datasets, implicit differences in case mix, and exclusive use of AUROC.

Introduction

In the intensive care unit (ICU), prognostic scores are used to monitor patients' severity of illness, predict outcomes, and guide clinical decisions about interventions and resource allocation [1,2]. These scores have quickly become a fixture in modern critical care and have been adopted in hospitals worldwide [3]. Established scoring systems — such as the Acute Physiology and Chronic Health Evaluation (APACHE) [4] or the Sequential Organ Failure Assessment (SOFA) [5] — rely on a small set of carefully selected parameters to identify patients or patient groups at risk of deterioration [6]. This simplicity comes at the cost of crude prognostication and limited accuracy.

The increasing availability of detailed electronic health records (EHR) has opened the door for developing more sophisticated and personalised scores. Machine learning (ML)-based artificial intelligence (AI) has emerged as a promising tool to leverage the wealth of data [7] and ML-based scores have attracted significant interest within the critical care community [8]. A growing body of literature demonstrates improved accuracy in predicting a diverse range of outcomes including all-cause mortality [9,10], sepsis [11,12], kidney injury [13,14], respiratory failure [15], and more [16,17].

Despite their promise, ML-based scoring systems are not without risk. One notable challenge is the potential for "overfitting", where a system's performance may become overly reliant on unique characteristics of the original patient cohort used for score development. Such overfitting can lead to inaccurate predictions when the system is used in a new hospital, where the original unique characteristics are no longer present [7]. Thus, external validation on data from previously unseen hospitals is a critical first step in establishing the robustness of these systems and ensuring their reliability across different clinical environments [18,19]. Unfortunately, external validation is often disregarded in practice [8,20], raising concerns about the true potential of ML-based scores in the ICU. Indeed, when a proprietary score for the detection of sepsis was implemented in clinical practice, an independent evaluation showed that it performed much worse than anticipated [21]. There is thus potential for an emerging translational gap, where theoretical benefits and advertised gains are not realised in clinical practice.

This systematic review aims to address this issue by determining how frequently external validation is performed in the literature and whether its use has increased in recent years. We further investigated how the performances of ML-based ICU scoring systems typically changed when applied to data from new hospitals. Our work contributes to the ongoing effort of bringing reliable ML-based scores to the ICU bedside.

Methods

Eligibility criteria

Studies were included in the review if they 1) described the development of an ML-based AI model that 2) provided early warning of acute patient deterioration in 3) ICU settings based on 4) structured, routinely collected EHR data. To be included in the meta-analysis of model

performance, models further needed to 5) be externally validated on data from a geographically distinct hospital that was not part of the derivation cohort. Following Shillan et al. (2020) [8], ML was defined as “any form of automated statistical analysis or data science methodology”. Clinical events were considered “acute” if they occurred up to 7 days after the time of prediction. A model gave early warning of such an event if the event was not yet known to the treating clinician at the time of prediction. The ICU was defined as “an area with a sole function to provide advanced monitoring or support to single or multiple body systems” [8]. Models could be externally validated as part of the same study that developed the model or in a later publication.

Studies were excluded if they: predicted auxiliary outcomes such as length of stay, risk of readmission, laboratory parameters, or values for imputation; used unsupervised learning methods to identify patient subgroups (unless those subgroups were used as input for supervised prediction); included non-ICU patients without providing separate performance metrics (e.g., by including patients from a general ward); required manual note review or prospective data collection of model features; used medical images or natural language processing of free-text notes; only validated the model on data from hospitals that contributed to the development data (including temporal validation on future data); did not report performance in the development dataset.

Search strategy

We searched the bibliographic databases Ovid MEDLINE and Web of Science for all full-text, peer-reviewed articles matching our search terms published in the English language before April 29th, 2022. We additionally searched the preprint server arXiv for relevant preprints using a custom computer script (see supplementary material at https://doi.org/10.17605/OSF.IO/F7J46). We included only primary research, excluding reviews and conference abstracts (except for abstracts that were peer-reviewed and paper-length, e.g., from the International Conference on Machine Learning).

We divided our search into three sub-themes: “Machine Learning and Artificial Intelligence”, “Intensive care setting”, and “Patient deterioration”. Articles were considered for screening if they matched all three themes. Notably, no theme was defined for external validation, which was ascertained manually during screening. Details of the search strategy including all search terms can be found in the preregistered study Protocol (www.crd.york.ac.uk/prospero, RecordID: 311514).

In an attempt to identify models that were validated in a separate, subsequent publication, we further performed a reserve citation search using Dimensions AI (https://www.dimensions.ai/), looking for validation papers that referenced a screened record (see supplementary material [22]).

Study selection

Identified articles were exported from the database as RIS files and imported into the reference management software Zotero (Cooperation for Digital Scholarship; version 6.0.26), where they were deduplicated using Zotero’s semi-automated deduplication tool.

Titles and abstracts were independently screened for inclusion by four of the authors (AH, BGC, EMA, PR), with each article being seen by at least two reviewers. For all articles that remained after title and abstract screening, full texts were obtained and independently checked for eligibility by two of the authors (EMA, PR). Before each screening stage, screening was piloted on 25 randomly selected articles. Agreement between authors was assessed using Fleiss' Kappa [23]. If agreement was found to be unsatisfactory (defined as Kappa < 0.6), decisions were calibrated on another set of 25 articles. If there was uncertainty about the eligibility of an article at any stage of the screening, the article was forwarded to the next stage. Any disagreements were resolved in a consensus meeting. If multiple identified articles describe the same model – e.g., when development and external validation were published in separate articles – the article relating to model validation was included and any missing information on performance in the development dataset was supplemented from the article describing the model development.

Data collection

Limited data collection was performed for all included studies, covering information on target outcome(s), data sources, and whether or not the study was externally validated. For the subset of externally validated studies, a more detailed data collection was performed in the Numbat Systematic Review Manager [24] using a predefined extraction template (see supplementary material [22]). The template was slightly extended prior to data collection to cover all elements defined in the MINimum Information for Medical AI Reporting (MINIMAR) standard [25]. Data collection was performed independently by two authors (EMA, PR). We extracted the following information for each validated study: target population; inclusion/exclusion criteria; information on the data sources including country of origin, number of hospital sites, cohort size, patient characteristics (age, sex, race, socioeconomic status), outcome prevalence; number and type of input features (e.g., vital signs or laboratory tests); ML algorithm; strategy for dealing with missing data; data splitting; performance metrics and performance in internal and external validation; whether the authors explicitly optimised for across-hospital generalisation; and if the authors provided their computer code with the study (e.g., on GitHub). For studies that reported results for more than one algorithm, the performance of the best algorithm during internal validation was recorded. For studies that reported results for more than one outcome, the performance for both outcomes was recorded if they were sufficiently different (e.g., mortality and sepsis), otherwise the most acute outcome was chosen (e.g., mortality at 24 hours if authors reported both mortality at 24 and 48 hours). If a data item could not be ascertained from the main text or supplementary material of the article, it was recorded as missing and no attempt was made to contact study authors for additional data.

Statistical analysis

Study characteristics and extracted performance metrics were summarised using descriptive statistics and graphical analysis. Changes over time in the proportion of studies performing external validation were assessed using a Chi-square test for linear trend.

Differences in the area under the receiver-operator characteristic curve (AUROC) were analysed using a random-effects model [26]. Parameters were estimated via a Bayesian

linear regression model with a single intercept and a normally distributed random effect per study. We used weakly informative normal priors for the mean and half-Cauchy priors for the scale of the random effects [27]. Due to an observed skewed distribution that might unduly influence the results, the difference was modelled with a Cauchy likelihood, which is less sensitive to outliers [28] and is often used for robust regression [29]. Each study's sample variance was derived using Hanley's formula [30]. To explore differences in models estimating mortality — which is a well-defined and well-captured ground truth compared to inferred complications such as sepsis [31] or kidney injury [32] — a second model with a fixed effect for mortality was specified. After estimation, we further calculated the proportion of studies in which the absolute difference in AUROC was > ±0.05. A 0.05 threshold was chosen in line with previous studies [33]. No analysis of heterogeneity between studies or risk of bias was performed.

All analysis was performed in R version 4.2.2 [34]. Bayesian linear models were fitted with Hamiltonian Monte Carlo (HMC) using the rstan package version 2.21.8 [35]. All results from the database search, screening, full-text review, and data collection as well as the analysis code are available at the Open Science Framework [22]. A study protocol was pre-registered on PROSPERO (www.crd.york.ac.uk/prospero, RecordID: 311514).

Results

We identified 4,677 records from MEDLINE (2,613 records), Web of Science (1,863 records), and arXiv (201 records). A detailed flow diagram is shown in Figure 1. After deduplication, the titles and abstracts of 3,851 records were screened. Full texts were assessed for 527 manuscripts, of which 355 (67.4%) described the prediction of acute deterioration in adult ICU patients from routine data (*included studies*). The main reasons for exclusion were prospective or other non-routine data capture, non-acute outcomes, or the inclusion of image, text, or waveform data (Figure 1). Of all included studies, 39 (11.0%) were also externally validated (*validated studies;* Table 1). No additional validation studies were identified through the reverse citation search. Agreement between reviewers as measured by Fleiss' Kappa was 0.623 for screening and 0.725 for full-text review.

Trend over time

The number of both included and validated studies increased significantly over time (p=0.014) and especially after 2018, with 302 / 355 (85.1%) respectively 38 / 39 (97.4%) studies published in or after that year (Figure 2). The earliest study performing external validation was published in 2015. Between 2018 and 2022, the proportion of validated studies increased from 2 / 28 (7.1%) to 7 / 32 (17.9%; only counting studies published until April 2022).

Outcomes

A total of 214 / 355 (60.3%) included studies predicted short-term mortality. The next most commonly predicted outcome was sepsis with 53 / 355 (14.9%), followed by 37 / 355

(10.4%) studies predicting renal complications including acute kidney injury, 19 / 355 (5.4%) studies predicting respiratory complications, 16 / 355 (4.5%) studies predicting circulatory failure, and 14 / 355 (3.9%) studies predicting cardiovascular complications. At only 14 / 214 (6.5%), the rate of external validation was notably lower among studies predicting mortality compared to all studies. If studies predicting mortality were excluded, the proportion of studies that were externally validated — and therefore included in the meta-analysis — almost doubled from 11.0% (39 / 355) to 18.2% (29 / 159).

Sources of data

Externally validated studies overwhelmingly used US data, with 37 / 39 (94.9%) including studies using at least one US dataset for model development or external validation. Eight studies used Chinese data, 5 studies used European data (Netherlands, Switzerland, Denmark, France), 3 used South Korean data, and 1 used Israeli data.

The publicly available datasets MIMIC [36] and eICU [37] were overrepresented among validated studies. MIMIC was used in 29 / 39 (74.4%) of validated studies compared to 206 / 355 (58.0%) of all included studies, with 14 studies using it for model development, 10 for external validation, and 5 for both. eICU was used in 17 / 39 (43.6%) of externally validated studies compared to 57 / 355 (16.1%) of all included studies, 5 times for model development, 8 times for external validation, and 4 times in both capacities. Together, MIMIC and eICU were used in 31 / 39 (79.5%) validated studies, of which they were the only source of data in 12 / 39 (30.8%) studies. AUMCdb [38] and HiRID [16] — two further, more recent public ICU databases — were only used in 2 / 39 (5.1%) included studies each.

Performance at new hospitals

All but one of the 39 validated studies reported AUROC. After accounting for sampling variability, model performance in the external validation data was -0.037 (95% credible interval [CI] -0.064 to -0.017; p-value < 0.001) lower than estimated in the internal validation data (Figure 3). Changes in performance ranged from a maximum increase of 0.14 to a decrease of -0.32. In 38.6% of cases, performance loss was < -0.05. On the other end of the spectrum, performance *increased* by > 0.05 in 9.1% of cases – indicating differences in patient populations between train and evaluation cohorts. There was no evidence for differences between studies predicting death and those that predicted other outcomes (p-value = 0.742).

Other commonly reported metrics included specificity (18 / 39; 46.2%), sensitivity (17 / 39; 43.6%), positive predictive value (16 / 39; 41.0%), F1 score (11 / 39; 28.2%), and accuracy (10 / 39; 25.6%), although they were reported at a much lower rate than AUROC.

Discussion

This systematic review examined the generalisation of complex, ML-based ICU scoring systems to new hospitals. We considered any score that supports ICU staff through the prediction of imminent patient deterioration from routinely collected EHR data. Leveraging

EHR data in this way to improve critical care continues to attract significant research interest, as evidenced by a steady increase in research output. Yet, translating this research into widespread clinical practice — and eventually converting it into patient benefit — requires comprehensive validation of findings, including an evaluation of the scores' performance at new hospitals. We found that such external validation is still relatively uncommon. Where validation was performed, performance at the new hospital tended to be lower than in the training cohort, often notably so.

Implications for the translation of AI into clinical practice

Fueled by recent advances in natural language processing and their successful translation to consumer products, there is a reinvigorated hype around the implementation of AI in healthcare [39]. Yet, while many preliminary results keep making the headlines, the proof is in the pudding: a large majority of published results are exploratory in nature, providing only proof-of-concepts [40]. There is a continued lack of verification and clinical validation, blocking the translation of these proof-of-concepts to actual products [19]. In our review, we demonstrate that the issue of inadequate verification extends to ML-based scoring systems: the rate of retrospective external validation – a first step to establish validity and robustness – remains low. Less than 20% of identified studies that proposed new scoring systems for the ICU underwent external validation. External validation in this context is an essential step for clinical adoption. Unless a model is solely built for use in the hospital(s) it was developed at – an unlikely goal – it should be judged by its accuracy across a range of hospitals, all of which may potentially use the model in the future. When evaluated this way, we found that average model accuracy as measured by the AUROC reduced by -0.037 compared to the training hospitals. This constitutes a relative decrease of 7-23% in performance, with decreases of up to and more than 50% in some cases. Many ostensibly well-performing scores may thus no longer be suitable for use at the new hospital, a fact that would (and does) go unnoticed in the absence of external validation. To actually facilitate translation to the clinical setting, rigorous external validation must become the standard when developing ML-based scoring systems and clinical AI more generally. Retrospective external validations in particular aid the early identification of model deficiencies, highlighting the need for training and fine-tuning on a broader variety of training data [41]. While there is still a long way to go to make such external validation the default, our review at least suggests that there is a growing recognition of its importance among researchers: over 80% of all identified studies performing external validation were published in 2018 or later.

Interpretation of external validation results

The infrequent external validation of ML models for the prediction of acute events in the ICU was already noted in a 2019 systematic review, with only 7% of studies at the time using geographically independent data for model validation [8]. This has been echoed in more recent, disease-specific reviews looking at models for sepsis [20] and acute kidney injury [42]. While we showed that this percentage has somewhat improved since, we also find that challenges remain even if external validation is performed.

While we observed a tendency for reduced model performance in external data, the magnitude of reduction was milder than anticipated from previous studies [41,43–45]. This may partially be explained by the performance metric. We focused on the AUROC as the

primary effect measure, since it allowed performing a meta-analysis due to its popularity and its comparability across different levels of prevalence. However, AUROC may be less sensitive to changes in the data. For example, while the drop in AUROC in the PhysioNet CinC challenge 2019 [43] was generally mild and in line with our findings, the "utility of prediction" — a custom metric defined as a timely prediction within 12 hours before to 3 hours after the onset of sepsis — in the new hospital was worse than not predicting at all. The average reduction in performance might have been more pronounced if another metric such as utility or normalised AUPRC were used instead of AUROC. Unfortunately, it was not possible to include such metrics in a meta-analysis due to their infrequent reporting. We recommend that future validation studies systematically report multiple performance metrics that represent the performance holistically.

The observed moderate reduction in average performance may have also been driven by the non-negligible number of models whose performance *increased* during external validation. Whereas minor fluctuations may occur due to sampling variability, a model's performance shouldn't notably increase in the external validation cohort. If it does, this suggests that there may be systematic differences in case mix between the training and validation cohorts – rendering the performances incomparable. If cohorts cannot be defined well enough to ensure their comparability, we recommend also reporting the performance of a model trained solely on the validation data. This provides a (potentially overfit) upper limit on what might have been achieved in the external data [41] and thus allows readers to take any distorting effects of case mix into consideration.

Finally, although the rate of external validation is slowly rising, it appears almost exclusively confined to a few open-source validation sets, most prominently MIMIC [36] and eICU [37]. A version of MIMIC was used in almost 80% of all identified studies that performed external validation. This is potentially problematic, as studies worldwide are thus largely judged by their ability to retain performance in patients from the single US hospital included in MIMIC, which does not necessarily represent the wider ICU population. This means that users and reviewers need to closely scrutinise claims of external validation in the area of ICU scoring systems if they judge tools that are to be used outside of the specific clinical settings captured by MIMIC. This also highlights that while large open-source datasets are able to fuel a large number of publications in certain areas, they do not necessarily by themselves improve the ability to build models that generalize, limiting their impact on successful translation to the clinical setting.

Strengths and limitations

We used a thorough, pre-defined search strategy to identify all relevant studies, covering two major bibliographic databases as well as the most relevant preprint server for ML research. Inclusion criteria were carefully assessed for all identified records by at least two reviewers and we additionally performed a reverse reference search to ensure we did not miss validation results that were published as stand-alone manuscripts.

To allow for direct comparability of AUROC in the development and validation data, we limited our analysis to external validation on retrospective, routine data. We did not capture validation that was performed by prospectively collecting additional data or within clinical

trials. This has two important implications. First, the proportion of validated studies may be higher than reported here, especially in the years preceding the availability of large open-source datasets. Second, the reported performances do not imply clinical usability but rather reflect the stability of study results across different sets of data. Nevertheless, external validation in retrospective data is an invaluable first step to assess the usability of a prediction model in clinical practice and should be considered for any study developing prediction models from routine data. Existing findings are fundamental to the conception of future studies and basing future research on 'false' or non-robust results can significantly hinder genuine innovation in the field, creating a substantial drain on both time and financial resources.

Due to the anticipated heterogeneity of studies, we limited ourselves to a descriptive summary of study results and trends. We did not perform a risk of bias assessment. Previous studies that assessed study quality reported a neglect of model calibration, inappropriate internal validation, and overall lack of reproducibility [20,42], all of which may also have been presented in the studies included here. Our results also assume that there were no systematic differences between studies that did and did not get externally validated. This is a strong assumption. For example, studies that were externally validated may be more generalisable to begin with because good performance in new dataset(s) was an explicit part of the study objectives. In this case, the true performance drop among non-validated studies may be even greater than estimated here.

Conclusion

Given the increasing availability of routine data capture, open-source ICU data sources [16,36–38], and well-documented tools for data harmonisation and preprocessing [46,47], there is little standing in the way of external validation of ML-based scoring systems. External validation should thus become the default for any study developing new scoring systems. It can provide invaluable information on the robustness of newly proposed scores and their potential for widespread adoption. However, while some external validation is certainly better than none, any results derived from it will only truly be useful if the data used for validation is representative of the model's intended future use setting. The data used for validation should be carefully selected and interpreted to ensure a fair comparison and enable meaningful interpretation, taking into account shifts in data quality, patient case mix, and any other factors that may impact model performance.

Funding
This work was supported through a postdoc grant awarded to PR by the Alexander von Humboldt Foundation (Grant Nr. 1221006). This work also received funding from the European Commission via the Horizon 2020 program for PRECISE4Q (No. 777107, lead: DF).

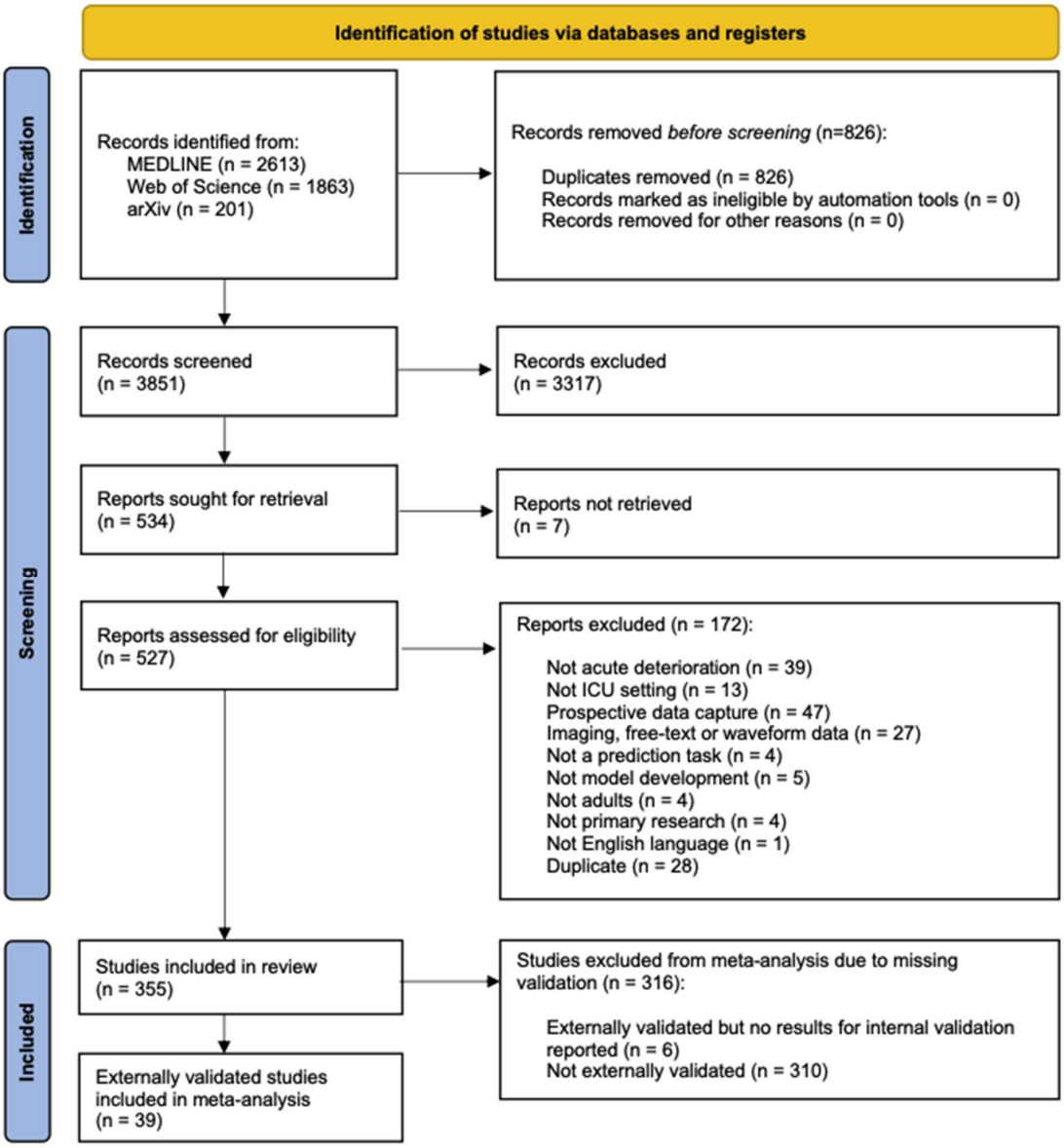

Figure 1 - PRISMA flow diagram

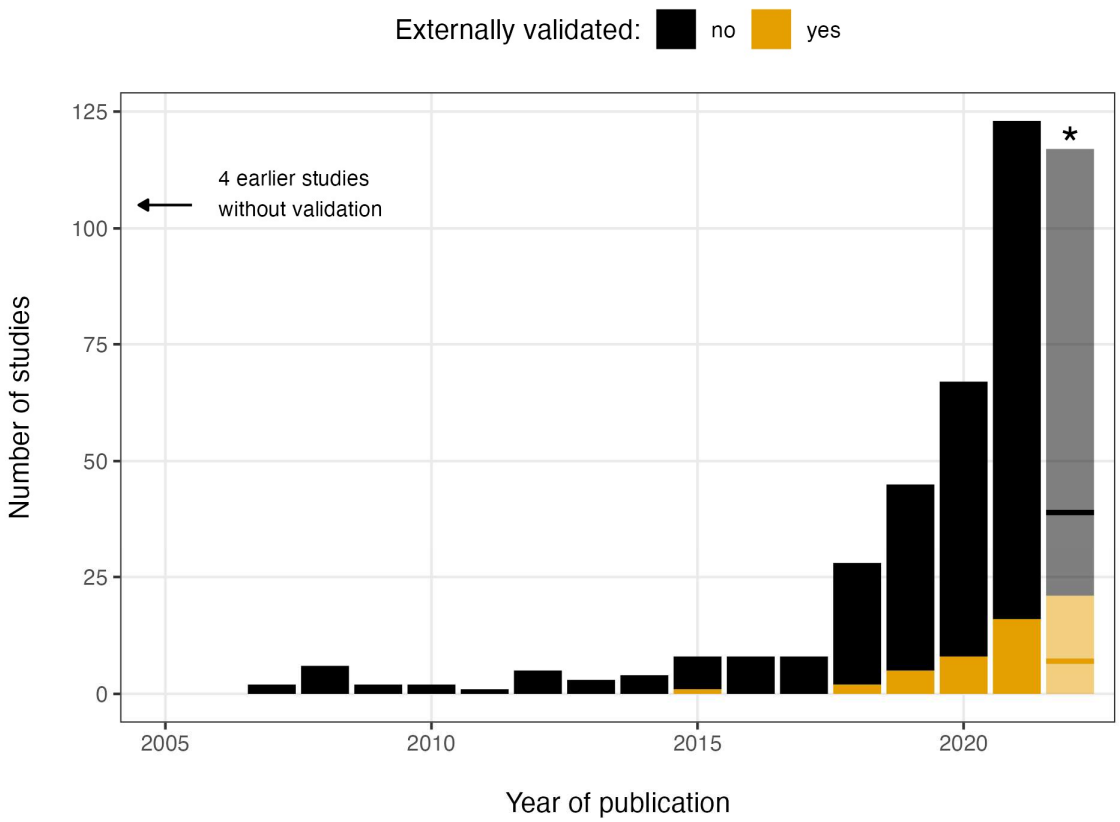

Figure 2 - Number of eligible (black) and included externally validated studies (orange) by year of publication. * For the year 2022, only studies published before May were included. Horizontal lines represent the numbers observed until the end of April, with the transparent bar representing the projected numbers for the full year.

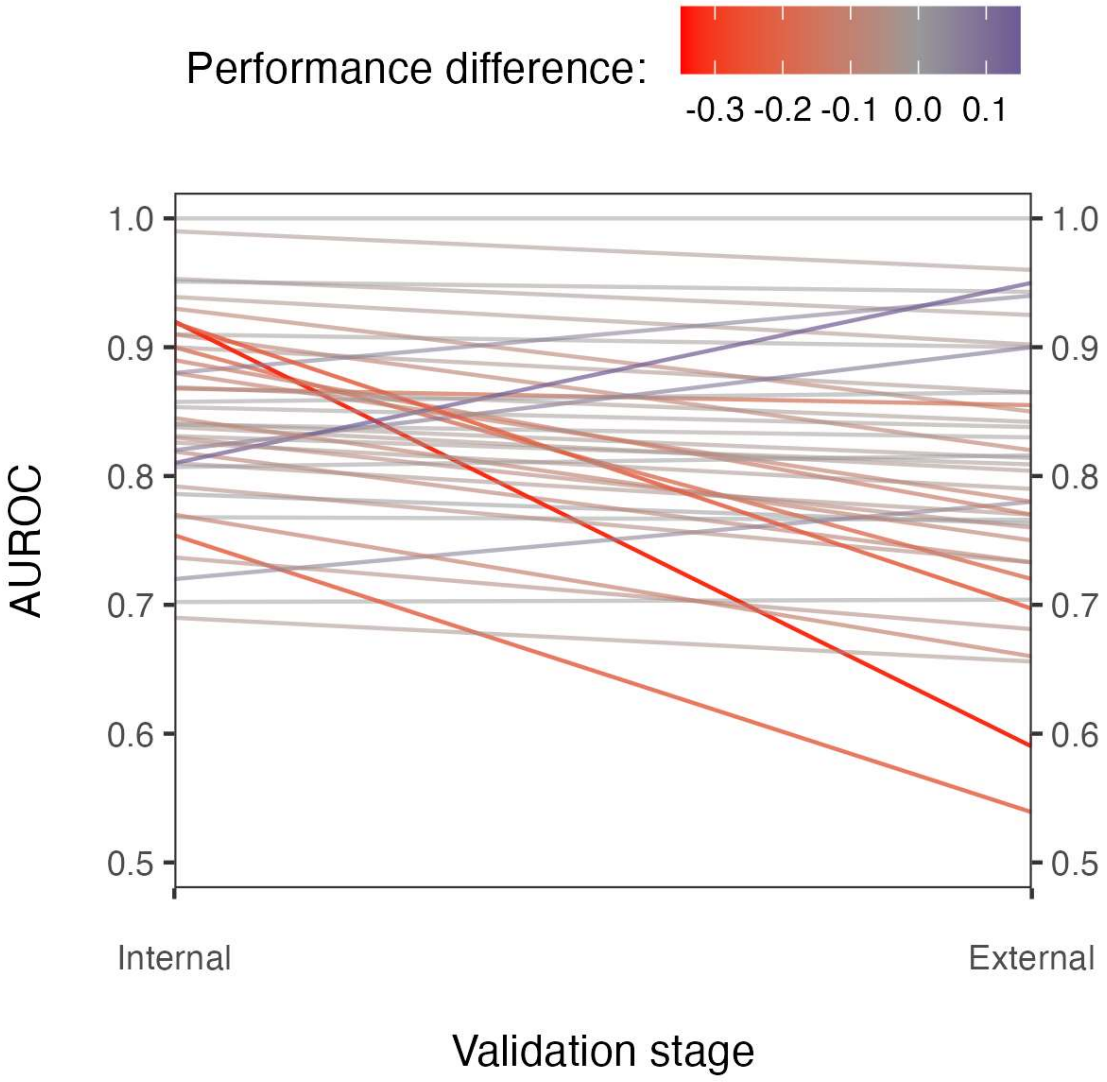

Figure 3 - Reported AUROCs for internal and external validation among (N = 39 -1) included studies. One study was omitted because they did not report AUROC.

Table 1 - Summary of externally validated studies

Authors	Year	Cohort	Outcome	Data source		Sample size		AUROC	
				Dev	Val	Dev	Val	Dev	Val
Pirracchio [9]	2015	All patients	Mortality	MIMIC	Other (France)	25k	200	.88	.94
Delahanty [48]	2018	All patients*	Mortality	Other (US)	Other (US)	147k	90k	.95	.94
Moon [49]	2018	All patients*	Neurological	Other (Korea)	Other (Korea)	3k	325	.90	.72
Huang [50]	2019	All patients*	Mortality	eICU	+,‡	28k		.74	.68
Liu [51]	2019	Sepsis	Sepsis (shock)	MIMIC	eICU	15k	?	.93	.85
Shickel [52]	2019	All patients*	Mortality	Other (US) MIMIC	+	36k 49k	+	.91 .91	.90 .90
van Wyk [53]	2019	All patients*	Sepsis	Other (US)	+	586	+	-	-
Nielsen [54]	2019	All patients*	Mortality	Other (DK)	Other (DK)	10k	2k	.79	.73
Kang [55]	2020	All patients*	Mortality	MIMIC eICU	+	21k 198k	+	.90 .87	.86 .72
Zhao [56]	2020	Sepsis	Other	MIMIC	eICU	11k	35k	.87	.84
Roimi [57]	2020	Susp. bacteraemia	Infection	MIMIC Other (Israel)	+	2k 1k	+	.89 .92	.59 .60
Hyland [16]	2020	All patients*	Circulatory	HiRID	MIMIC	36k	9k	.94	.90
Reyna [43]	2020	All patients*	Sepsis	MIMIC Other (US)	Other (US)	20k 20k	?	.82 .86	.81
Wang [58]	2020	All patients*	Renal	Other (China)	MIMIC	11k	46k	.81	.95
Liu [59]	2020	MODS	Mortality	MIMIC eICU	Other (China)	15k 34k	439	.86 .85	.84
Zhou [60]	2020	Viral pneumonia	Mortality	eICU	MIMIC	4k	937	.77	.66
Rahman [61]	2021	All patients*	Circulatory	eICU	MIMIC	216k	16k	.82	.90
Zhi [62]	2021	Sepsis	Mortality	MIMIC	Other (China)	2k	125	.75	.54
Holder[63]	2021	Sepsis	Other	Other (US)	Other (US)	9k	5k	.81	.77
Hur [17]	2021	All patients*	Neurological	Other (Korea)	MIMIC	12k	2k	.92	.70
Chen [64]	2021	All patients*	Renal	MIMIC	Other (China)	46k	226	.83	.79
He [65]	2021	Sepsis and AKI	Renal	Other (China)	MIMIC	209	509	1.0	1.0
Shashikumar [66]	2021	All patients*	Sepsis	Other (US)	Other (US)	17k	46k	.95	.93
Levi [67]	2021	GI bleeding	Other	MIMIC eICU	+	4k 10k	+	.81 .79	.76 .80
Ding [68]	2021	Sepsis and AKI	Renal	MIMIC	eICU	7k	3k	.70	.70
Huang [69]	2021	AKI	Mortality	MIMIC	eICU	4k	1k	.91	.82
Singhal [70]	2021	COVID-19	Respiratory	Other (US)	Other (US)	6k	611 77	.90	.85 .88
Sung [71]	2021	All patients*	Mortality	Other (Korea)	Other (Korea)	22k	2k	.99 .77 .84	.96 .77 .80
Liu [72]	2021	All patients*	Sepsis	Other (US)	eICU	882	6k	.72	.78
Shawwa [73]	2021	All patients*	Renal	Other (US)	MIMIC	98k	19k	.69	.66
Mamandipoor [74]	2021	All patients*	Other	eICU	MIMIC	17k	13k	.84	.83

Moor [45]	2021	All patients*	Sepsis	MIMIC eICU HiRID AUMCdb	+	37k 57k 27k 16k	+	.83 .80 .83 .92	.71 .75 .73 .81
Peng [75]	2022	CHF	Renal	MIMIC	eICU	9k	10k	.81	.82
Luo [76]	2022	CHF	Mortality	MIMIC	eICU	6k	1k	.83	.81
Kim [77]	2022	Cardiac arrest and MV	Mortality	eICU	MIMIC	2k	86	.83	.76
Fu [78]	2022	Cardiogenic shock	Renal	MIMIC	eICU	1k	1k	.82	.73
Zhang [79]	2022	Cerebrovascular disease	Renal	MIMIC	Other (China)	3k	499	.88	.78
Jiang [80]	2022	Sepsis	Other	Other (China)	MIMIC	1k	688	.89	.77
Liang [81]	2022	All patients*	Renal	Other (China) MIMIC	AUMCdb	6k 37k	15k	.86 .86	.87

* “All Patients” are defined as a general adult ICU patient population without specifying additional health conditions (e.g., admitted with sepsis). + Datasets were used alternatingly for development and validation. ‡ eICU includes data from 208 different hospitals and may thus be used for both development and validation if split by hospital.

AKI, acute kidney injury: AUROC, area under the receiver operating characteristic; CHF, congestive heart failure; DK, Denmark; GI, gastro-intestinal; MODS, Multi-organ dysfunction syndrome; MV, mechanical ventilation; US, United States

References

1. Ferreira FL, Bota DP, Bross A, Mélot C, Vincent JL. Serial evaluation of the SOFA score to predict outcome in critically ill patients. JAMA. 2001;286: 1754–1758.

2. Vincent J-L, Moreno R. Clinical review: scoring systems in the critically ill. Crit Care. 2010;14: 207.

3. Gerry S, Bonnici T, Birks J, Kirtley S, Virdee PS, Watkinson PJ, et al. Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. BMJ. 2020;369: m1501.

4. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med. 2006;34: 1297–1310.

5. 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. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med. 1996;22: 707–710.

6. Vincent J-L, de Mendonca A, Cantraine F, Moreno R, Takala J, Suter PM, et al. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: Results of a multicenter, prospective study. Crit Care Med. 1998;26: 1793.

7. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17: 195.

8. Shillan D, Sterne JAC, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Crit Care. 2019;23: 284.

9. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med. 2015;3: 42–52.

10. Silva I, Moody G, Scott DJ, Celi LA, Mark RG. Predicting In-Hospital Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012. Comput Cardiol . 2012;39: 245–248.

11. 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.

12. Moor M, Horn M, Rieck B, Roqueiro D, Borgwardt K. Early Recognition of Sepsis with Gaussian Process Temporal Convolutional Networks and Dynamic Time Warping. In: Doshi-Velez F, Fackler J, Jung K, Kale D, Ranganath R, Wallace B, et al., editors. Proceedings of the 4th Machine Learning for Healthcare Conference. PMLR; 09--10 Aug 2019. pp. 2–26.

13. Koyner JL, Carey KA, Edelson DP, Churpek MM. The Development of a Machine Learning Inpatient Acute Kidney Injury Prediction Model. Crit Care Med. 2018;46: 1070–1077.

14. Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sündermann SH, et al. Machine learning for real-time prediction of complications in critical care: a retrospective

study. Lancet Respir Med. 2018;6: 905–914.

15. Hüser M, Faltys M, Lyu X, Barber C, Hyland SL, Merz TM, et al. Early prediction of respiratory failure in the intensive care unit. arXiv [cs.LG]. 2021. Available: http://arxiv.org/abs/2105.05728

16. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med. 2020;26: 364–373.

17. Hur S, Ko R-E, Yoo J, Ha J, Cha WC, Chung CR. A Machine Learning-Based Algorithm for the Prediction of Intensive Care Unit Delirium (PRIDE): Retrospective Study. JMIR Med Inform. 2021;9: e23401.

18. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021;14: 49–58.

19. Higgins D, Madai VI. From bit to bedside: A practical framework for artificial intelligence product development in healthcare. Adv Intell Syst. 2020;2: 2000052.

20. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review. Front Med. 2021;8: 607952.

21. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med. 2021;181: 1065–1070.

22. Rockenschaub P. Supplement for “Generalisability of AI-based scoring systems in the ICU: a systematic review and meta-analysis.” Open Science Framework; 2023. doi:10.17605/OSF.IO/F7J46

23. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76: 378–382.

24. Carlisle BG. Numbat Systematic Review Manager. Berlin, Germany: The Grey Literature; 2014. Available: https://numbat.bgcarlisle.com

25. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 2020;27: 2011–2015.

26. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. J Stat Softw. 2010;36: 1–48.

27. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis, Third Edition. CRC Press; 2013.

28. Gelman A, Jakulin A, Pittau MG, Su Y-S. A weakly informative default prior distribution for logistic and other regression models. aoas. 2008;2: 1360–1383.

29. Pawitan Y. In All Likelihood: Statistical Modelling and Inference Using Likelihood. OUP Oxford; 2001.

30. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143: 29–36.

31. 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: 801–810.

32. Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. Nephron Clin Pract. 2012;120: c179–84.

33. Sullivan PG, Wallach JD, Ioannidis JPA. Meta-Analysis Comparing Established Risk Prediction Models (EuroSCORE II, STS Score, and ACEF Score) for Perioperative Mortality During Cardiac Surgery. Am J Cardiol. 2016;118: 1574–1582.

34. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: Vienna, Austria; 2018. Available: https://www.R-project.org/

35. Stan Development Team. RStan: the R interface to Stan. 2023. Available: https://mc-stan.org/

36. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3: 160035.

37. 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.

38. Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, 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. Crit Care Med. 2021;49: e563–e577.

39. Kulkarni PA, Singh H. Artificial Intelligence in Clinical Diagnosis: Opportunities, Challenges, and Hype. JAMA. 2023;330: 317–318.

40. Ben-Israel D, Jacobs WB, Casha S, Lang S, Ryu WHA, de Lotbiniere-Bassett M, et al. The impact of machine learning on patient care: A systematic review. Artif Intell Med. 2020;103: 101785.

41. Rockenschaub P, Hilbert A, Kossen T, von Dincklage F, Madai VI, Frey D. From Single-Hospital to Multi-Centre Applications: Enhancing the Generalisability of Deep Learning Models for Adverse Event Prediction in the ICU. arXiv [cs.LG]. 2023. Available: http://arxiv.org/abs/2303.15354

42. Vagliano I, Chesnaye NC, Leopold JH, Jager KJ, Abu-Hanna A, Schut MC. Machine learning models for predicting acute kidney injury: a systematic review and critical appraisal. Clin Kidney J. 2022;15: 2266–2280.

43. Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, et al. Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019. Crit Care Med. 2020;28: 210–217.

44. Song X, Yu ASL, Kellum JA, Waitman LR, Matheny ME, Simpson SQ, et al. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nat Commun. 2020;11: 5668.

45. Moor M, Bennet N, Plecko D, Horn M, Rieck B, Meinshausen N, et al. Predicting sepsis in multi-site, multi-national intensive care cohorts using deep learning. arXiv [cs.LG]. 2021. Available: http://arxiv.org/abs/2107.05230

46. van de Water R, Schmidt H, Elbers P, Thoral P, Arnrich B, Rockenschaub P. Yet Another ICU Benchmark: A Flexible Multi-Center Framework for Clinical ML. arXiv [cs.LG]. 2023. Available: http://arxiv.org/abs/2306.05109

47. Bennett N, Plečko D, Ukor I-F, Meinshausen N, Bühlmann P. ricu: R's interface to intensive care data. Gigascience. 2022;12. doi:10.1093/gigascience/giad041

48. Delahanty RJ, Kaufman D, Jones SS. Development and Evaluation of an Automated Machine Learning Algorithm for In-Hospital Mortality Risk Adjustment Among Critical Care Patients. Crit Care Med. 2018;46: e481–e488.

49. Moon K-J, Jin Y, Jin T, Lee S-M. Development and validation of an automated delirium risk assessment system (Auto-DelRAS) implemented in the electronic health record system. Int J Nurs Stud. 2018;77: 46–53.

50. Huang L, Shea AL, Qian H, Masurkar A, Deng H, Liu D. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. J Biomed Inform. 2019;99: 103291.

51. Liu R, Greenstein JL, Granite SJ, Fackler JC, Bembea MM, Sarma SV, et al. Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU. Sci Rep. 2019;9: 6145.

52. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. Sci Rep. 2019;9: 1879.

53. 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: 978–986.

54. Nielsen AB, Thorsen-Meyer H-C, Belling K, Nielsen AP, Thomas CE, Chmura PJ, et al. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. Lancet Digit Health. 2019;1: e78–e89.

55. Kang Y, Jia X, Wang K, Hu Y, Guo J, Cong L, et al. A Clinically Practical and Interpretable Deep Model for ICU Mortality Prediction with External Validation. AMIA Annu Symp Proc. 2020;2020: 629–637.

56. Zhao Q-Y, Liu L-P, Luo J-C, Luo Y-W, Wang H, Zhang Y-J, et al. A Machine-Learning Approach for Dynamic Prediction of Sepsis-Induced Coagulopathy in Critically Ill Patients With Sepsis. Front Med. 2020;7: 637434.

57. Roimi M, Neuberger A, Shrot A, Paul M, Geffen Y, Bar-Lavie Y. Early diagnosis of bloodstream infections in the intensive care unit using machine-learning algorithms. Intensive Care Med. 2020;46: 454–462.

58. Wang Y, Wei Y, Yang H, Li J, Zhou Y, Wu Q. Utilizing imbalanced electronic health records to predict acute kidney injury by ensemble learning and time series model. BMC Med Inform Decis Mak. 2020;20: 238.

59. Liu X, Hu P, Mao Z, Kuo P-C, Li P, Liu C, et al. Interpretable Machine Learning Model for Early Prediction of Mortality in Elderly Patients with Multiple Organ Dysfunction Syndrome (MODS): a Multicenter Retrospective Study and Cross Validation. arXiv [physics.med-ph]. 2020. Available: http://arxiv.org/abs/2001.10977

60. Zhou H, Cheng C, Lipton ZC, Chen GH, Weiss JC. Predicting Mortality Risk in Viral and Unspecified Pneumonia to Assist Clinicians with COVID-19 ECMO Planning. arXiv [stat.AP]. 2020. Available: http://arxiv.org/abs/2006.01898

61. Rahman A, Chang Y, Dong J, Conroy B, Natarajan A, Kinoshita T, et al. Early prediction of hemodynamic interventions in the intensive care unit using machine learning. Crit Care. 2021;25: 388.

62. Zhi D, Zhang M, Lin J, Liu P, Wang Y, Duan M. Establishment and validation of the predictive model for the in-hospital death in patients with sepsis. Am J Infect Control. 2021;49: 1515–1521.

63. Holder AL, Shashikumar SP, Wardi G, Buchman TG, Nemati S. A Locally Optimized Data-Driven Tool to Predict Sepsis-Associated Vasopressor Use in the ICU. Crit Care Med. 2021;49: e1196–e1205.

64. Chen Z, Chen M, Sun X, Guo X, Li Q, Huang Y, et al. Analysis of the Impact of Medical Features and Risk Prediction of Acute Kidney Injury for Critical Patients Using Temporal Electronic Health Record Data With Attention-Based Neural Network. Front Med. 2021;8: 658665.

65. He J, Lin J, Duan M. Application of Machine Learning to Predict Acute Kidney Disease in Patients With Sepsis Associated Acute Kidney Injury. Front Med. 2021;8: 792974.

66. Shashikumar SP, Wardi G, Malhotra A, Nemati S. Artificial intelligence sepsis prediction algorithm learns to say “I don’t know.” NPJ Digit Med. 2021;4: 134.

67. Levi R, Carli F, Arévalo AR, Altinel Y, Stein DJ, Naldini MM, et al. Artificial intelligence-based prediction of transfusion in the intensive care unit in patients with gastrointestinal bleeding. BMJ Health Care Inform. 2021;28. doi:10.1136/bmjhci-2020-100245

68. Ding C, Hu T. Development and External Verification of a Nomogram for Patients with Persistent Acute Kidney Injury in the Intensive Care Unit. Int J Gen Med. 2021;14: 5005–5015.

69. Huang H, Liu Y, Wu M, Gao Y, Yu X. Development and validation of a risk stratification model for predicting the mortality of acute kidney injury in critical care patients. Ann Transl Med. 2021;9: 323.

70. Singhal L, Garg Y, Yang P, Tabaie A, Wong AI, Mohammed A, et al. eARDS: A multi-center validation of an interpretable machine learning algorithm of early onset Acute Respiratory Distress Syndrome (ARDS) among critically ill adults with COVID-19. PLoS One. 2021;16: e0257056.

71. Sung M, Hahn S, Han CH, Lee JM, Lee J, Yoo J, et al. Event Prediction Model Considering Time and Input Error Using Electronic Medical Records in the Intensive Care Unit: Retrospective Study. JMIR Med Inform. 2021;9: e26426.

72. Liu Z, Khojandi A, Mohammed A, Li X, Chinthala LK, Davis RL, et al. HeMA: A hierarchically enriched machine learning approach for managing false alarms in real time: A sepsis prediction case study. Comput Biol Med. 2021;131: 104255.

73. Shawwa K, Ghosh E, Lanius S, Schwager E, Eshelman L, Kashani KB. Predicting acute kidney injury in critically ill patients using comorbid conditions utilizing machine learning. Clin Kidney J. 2021;14: 1428–1435.

74. Mamandipoor B, Yeung W, Agha-Mir-Salim L, Stone DJ, Osmani V, Celi LA. Prediction of blood lactate values in critically ill patients: a retrospective multi-center cohort study. J Clin Monit Comput. 2022;36: 1087–1097.

75. Peng X, Li L, Wang X, Zhang H. A Machine Learning-Based Prediction Model for Acute Kidney Injury in Patients With Congestive Heart Failure. Front Cardiovasc Med. 2022;9: 842873.

76. Luo C, Zhu Y, Zhu Z, Li R, Chen G, Wang Z. A machine learning-based risk stratification tool for in-hospital mortality of intensive care unit patients with heart failure. J Transl Med. 2022;20: 136.

77. Kim HB, Nguyen HT, Jin Q, Tamby S, Gelaf Romer T, Sung E, et al. Computational signatures for post-cardiac arrest trajectory prediction: Importance of early physiological time series. Anaesth Crit Care Pain Med. 2022;41: 101015.

78. Fu S, Wang Q, Chen W, Liu H, Li H. Development and External Validation of a Nomogram for Predicting Acute Kidney Injury in Cardiogenic Shock Patients in Intensive Care Unit. Int J Gen Med. 2022;15: 3965–3975.

79. Zhang X, Chen S, Lai K, Chen Z, Wan J, Xu Y. Machine learning for the prediction of acute kidney injury in critical care patients with acute cerebrovascular disease. Ren Fail. 2022;44: 43–53.

80. Jiang X, Wang Y, Pan Y, Zhang W. Prediction Models for Sepsis-Associated Thrombocytopenia Risk in Intensive Care Units Based on a Machine Learning Algorithm. Front Med. 2022;9: 837382.

81. Liang Q, Xu Y, Zhou Y, Chen X, Chen J, Huang M. Severe acute kidney injury predicting model based on transcontinental databases: a single-centre prospective study. BMJ Open. 2022;12: e054092.