Review Article

Healthc Inform Res. 2023 October;29(4):301-314.
https://doi.org/10.4258/hir.2023.29.4.301
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Machine Learning for Benchmarking Critical Care Outcomes

Louis Atallah[1], Mohsen Nabian[1], Ludmila Brochini[2], Pamela J. Amelung[3]

[1]Clinical Integration and Insights, Philips, Cambridge, MA, USA
[2]Clinical Integration and Insights, Philips, Eindhoven, The Netherlands
[3]EMR & Care Management, Philips, Cambridge, MA, USA

Objectives: Enhancing critical care efficacy involves evaluating and improving system functioning. Benchmarking, a retrospective comparison of results against standards, aids risk-adjusted assessment and helps healthcare providers identify areas for improvement based on observed and predicted outcomes. The last two decades have seen the development of several models using machine learning (ML) for clinical outcome prediction. ML is a field of artificial intelligence focused on creating algorithms that enable computers to learn from and make predictions or decisions based on data. This narrative review centers on key discoveries and outcomes to aid clinicians and researchers in selecting the optimal methodology for critical care benchmarking using ML. **Methods:** We used PubMed to search the literature from 2003 to 2023 regarding predictive models utilizing ML for mortality (592 articles), length of stay (143 articles), or mechanical ventilation (195 articles). We supplemented the PubMed search with Google Scholar, making sure relevant articles were included. Given the narrative style, papers in the cohort were manually curated for a comprehensive reader perspective. **Results:** Our report presents comparative results for benchmarked outcomes and emphasizes advancements in feature types, preprocessing, model selection, and validation. It showcases instances where ML effectively tackled critical care outcome-prediction challenges, including nonlinear relationships, class imbalances, missing data, and documentation variability, leading to enhanced results. **Conclusions:** Although ML has provided novel tools to improve the benchmarking of critical care outcomes, areas that require further research include class imbalance, fairness, improved calibration, generalizability, and long-term validation of published models.

Keywords: Benchmarking, Critical Care, Length of Stay, Machine Learning, Mortality, Ventilation

Submitted: December 9, 2022
Revised: August 23, 2023
Accepted: September 25, 2023

Corresponding Author
Louis Atallah
Clinical Integration and Insights, Philips, 222 Jacobs Street, 7th Floor, Cambridge, MA 02141, USA. Tel: +1-617-798-8244, E-mail: louis.atallah@philips.com (https://orcid.org/0000-0002-6657-319X)

I. Introduction

Performance comparison is an important aspect of benchmarking in critical care, whether to observe a critical care unit over time or to compare units, hospitals, or even health systems across geographic regions [1,2]. Benchmarking outcomes in critical care, such as mortality or length of stay, allows a risk-adjusted comparison with healthcare leaders as a proxy for quality and efficacy of care. Risk adjustment models have been the cornerstone for benchmarking outcomes in critical care. These models allow the prediction of outcomes to enable the benchmarking or comparison of actual versus

predicted outcomes among peers. Outcomes are difficult to interpret unless they are risk-stratified for diagnosis groups, severity of illness, and other patient characteristics [3].

Several taskforces worldwide have recommended the use of quality indicators that are measurable, comparable, and relevant across critical care units [1,4-6]. Regarding outcomes, several measures have been proposed [3,7]. An example is mortality, which is utilized as a quality indicator in intensive care units (ICUs) due to its direct reflection of patient outcomes; it serves to measure the effectiveness of medical interventions and the overall quality of care provided. Mortality is usually assessed using the standardized mortality ratio, which compares actual hospital mortality to predicted mortality through risk-adjusted scoring systems. Morbidity and complications, such as acute renal failure, hemodialysis, and prolonged mechanical ventilation, are more prevalent than mortality events and are also used as outcome measures [8]. Length of stay, encompassing both hospital and ICU durations, is commonly employed as an indicator of cost and efficiency; however, it is influenced by variables like structural factors and patient transfers [9]. Variation in ICU readmissions can also highlight opportunities for enhancement and is potentially influenced by ICU discharge practices [10]. Ventilation outcomes, including mechanical ventilation duration [11] and probability [12], facilitate the comparison of ventilator practices across ICUs. Ventilation outcomes are also valuable for controlling patient disparities in clinical trials or weaning techniques and for advancing quality improvement endeavors. Patient-reported outcomes, used to a lesser degree, cover a range of aspects, such as cognition, fatigue, pain, psychological well-being, activities of daily living, sleep, appetite, and alcohol consumption [13].

Machine learning (ML) constitutes a field within computer science where statistical techniques are employed to analyze data, which facilitates classification, prediction, and optimization by leveraging past data observations. It can help address issues such as imbalanced classes (such as deaths versus surviving patients), missing data, and variation in documentation. This narrative review is meant for clinicians and scientists who would like to understand some of the most important directions in developing these models for benchmarking clinical outcomes. We also highlight the most important sources of bias and variations in performance, aiming to give researchers a concrete list of factors to consider when planning benchmarking studies.

II. Methods

This article reviews ML approaches for benchmarking clinical outcomes in the ICU with a focus on mortality, length of stay, and mechanical ventilation. The literature search was conducted on PubMed, including all articles and reviews between January 1, 2003 and August 1, 2023. Search terms for mortality were "mortality," "ICU" AND ("machine learning" OR "artificial Intelligence"). For length of stay, they were "length of stay," "ICU" AND ("machine learning" OR "artificial Intelligence"). For ventilation, the search terms were "ventilation," "ICU" AND ("machine learning" OR "artificial Intelligence"). Only articles related to adult critical care in English were included. The searches above were also conducted in Google Scholar to ensure that relevant works were not excluded, and to add any missing articles. The initial search yielded 592 articles on mortality, 143 on length of stay, and 195 on ventilation. After a meticulous review, 26, 12, and nine pertinent papers were chosen for each respective domain. For mortality and length of stay, we eliminated articles focusing on specific patient groups and focused on approaches applicable to all critical care patients. An added condition for mortality was the use of a dataset of more than 10,000 patients to enable a fair comparison of results between different studies. In this narrative review, we focus on the important directions for ML in each outcome area rather than providing an exhaustive listing of prior work.

III. Outcome Benchmarking with ML

1. Mortality Benchmarking with ML

Mortality prediction models are applied to critical care patients for benchmarking and stratification into different risk categories. The most widely used models are the Acute Physiology and Chronic Health Evaluation (APACHE) models, the Simplified Acute Physiology Score (SAPS) I–III, and the Mortality Prediction Model (MPM) [14]. However, other models have been developed for improved calibration in particular regions, such as the Intensive Care National Audit & Research Centre (ICNARC) in the UK [15].

Several reviews have covered mortality models: Keuning et al. [16] surveyed predictive mortality models and focused mostly on statistical linear models. An earlier review by Strand et al. [17] reviewed articles focusing on prognostic, single-organ failure, trauma scores and organ dysfunction scores. Siontis et al. [18] evaluated predictive mortality models with a focus on specific patient groups. Promising approaches for particular groups such as brain injury [19] and

https://doi.org/10.4258/hir.2023.29.4.301

coronavirus disease 2019 (COVID-19) patients [20] have also been explored. The 2012 PhysioNet/Computing in Cardiology Challenge focused on the prediction of in-hospital mortality of ICU patients leading to several new prediction models [21]. In a more recent review by Barboi et al. [22], the authors highlighted that ML-based models can accurately predict ICU mortality as an alternative to traditional scoring models. However, they concluded that the results cannot be generalized due to the high degree of heterogeneity and that clinicians should only select models with sufficient validation for use in a practice environment.

Table 1 summarizes several relevant articles on ML for predicting mortality [15,23-46]. Note that there are several aspects of mortality prediction: at the ICU level or hospital level, within 48 or 72 hours after discharge, and 28-day and 90-day mortality, among others. The periods used are variable. For example, mortality may be predicted on admission to the ICU, during the first 6 hours [32], 24 hours after arrival (similar to APACHE), during the last ICU day [28], or even in a continuous manner [36].

Although it is challenging to compare the approaches in Table 1, since many were developed on different datasets and predict different types of mortality (ICU or in-hospital versus post-release), we may summarize the main observations:

Improved interpretability: Numerous ML algorithms have faced criticism for their "black box" nature, which limits interpretability. This concern is particularly evident in deep learning models, where a balance between predictive accuracy and interpretability must be struck. Deep learning, a subset of ML, organizes algorithms into layers, forming an artificial neural network capable of learning from data. Methods such as Shapley values [47], used in Thorsen-Meyer et al. [23] and Caicedo-Torres et al. [24], can convey the importance (or weighting) that the deep model assigns to each input feature, which offers improved interpretability for these networks.

Features used: The approaches summarized vary between models that use features similar to existing models (APACHE) as well as some novel features. The benefits of using simple features such as demographics, labs, and vitals are their availability, reliability, and ease of use. Even a reduced set of features, such as the 15 selected by Kim et al. [31], showed a good area under the curve (AUC) when used with ML models. However, when combined with static features, physiological time series such as vitals and interventions offer an improved means of continuous mortality prediction [23]. Another promising direction is to use semi-structured data, such as those present in diagnosis and inspection reports [37]. Methods such as topic modeling from clinical notes can be added to traditional variables to improve prediction [37,45]. Grnarova et al. [26] proposed a convolutional document-embedding approach applied to clinical notes showing high AUC values. However, variations in clinical annotation practices across health systems may affect how benchmarking may be applied to this type of model. Purushotham et al. [46] compared hand-picked features (such as those used for SAPS-II), raw values of features, and inputs without pre-processing. They showed that when using models that can learn data representations (such as deep learning models), unprocessed inputs provided the best results. Although premorbid functional status and diagnosis are known predictors of ICU–relevant study outcomes, they are not regularly implemented in established scoring systems. Moser et al. [34] included this information showing increased predictive model performance compared to predictions from established risk scoring systems.

Model choice: A one-fits-all model is unlikely, since model selection depends on the type of features used (raw data or clinical notes versus hand-picked clinical features) and outcomes required (continuous mortality prediction versus in-hospital and post-discharge). However, several promising approaches have addressed mortality prediction in different ways. Purushotham et al. [46] benchmarked the performance of deep learning models with respect to ensemble ML models and prognostic scoring systems, showing improved performance of deep learning models. Deep learning also offered promising results in Caicedo-Torres et al. [24], who used multi-scale deep convolutional neural networks. It was also used in Aczon et al. [25] regarding pediatric mortality risk. A convolutional document-embedding approach based on the textual content of clinical notes was proposed by Grnarova et al. [26]. Another popular approach is that of using ensemble classifiers to leverage the power of different groups of classifiers. Guo et al. [30] proposed a dynamic ensemble-learning algorithm based on k-means (DELAK) for mortality prediction. They used k-means sampling to generate several data subsets on which base classifiers could learn the classification boundary. El-Rashidy et al. [35] used a stacking ensemble classifier, leading to a high AUC for in-hospital mortality, whereas Awad et al. [32] used an ensemble-learning random forest model.

Class imbalance: A common problem with mortality prediction is that of class imbalance, with a rather low mortality versus survival rate. Several strategies are deployed commonly, either to pre-process imbalanced data (re-sampling, optimizing feature space) or to provide new algorithms that can address this problem. Bhattacharya et al. [29], for ex-

Table 1. Summary of studies that use machine learning to predict mortality

Study, year	Outcome	Number of patients/stays	Method	Main results
Ferrando-Vivas et al. [15], 2017	Acute hospital mortality, including deaths that occurred after transfer of the patient from the original hospital to another acute hospital.	Training: 155,239 admissions; Validation: 90,017 admissions	Multivariate logistic regression	AUC = 0.8853
Thorsen-Meyer et al. [23], 2020	90-day mortality	14,190 admissions of 11,492 patients	Recurrent neural network trained on a temporal resolution of 1 hour	AUC = 0.73 at admission; AUC = 0.82 after 24 hours; AUC = 0.85 after 72 hours; AUC = 0.88 at the time of discharge
Caicedo-Torres et al. [24], 2019	ICU mortality	22,413 patients	Multi-scale deep convolutional neural network	AUC = 0.8735
Aczon et al. [25], 2021	Pediatric mortality risk 12 hours after admission and prior to discharge	9,070 children	Recurrent neural network	AUC = 0.94
Grnarova et al. [26], 2016	In-hospital, 30-day and 1 year mortality	46,520 patients	Convolutional document embedding approach based on textual content of clinical notes	In-hospital (AUC = 0.963); 30-day (AUC = 0.858); 1-year mortality (AUC = 0.853)
Purushotham et al. [46], 2018	In-hospital, short term (2-3 days), 30-day and 1-year post discharge	58,576 admissions	Benchmarked the performance of deep learning models with respect to machine learning models and prognostic scoring systems	For deep learning models: AUC = 0.92 (in-hospital mortality); AUC = 0.8872 (1-year post-discharge)
Badawi et al. [28], 2012	Mortality within 48 hours of release from the ICU	469,976 patients (development); 234,987 patients (validation)	Multivariable logistic regression	AUC = 0.92
Bhattacharya et al. [29], 2017	In-hospital mortality	4,000 patients from the PhysioNet 2012 challenge [21]	A binary classifier consisting of skewness-based transformation of input features and statistical hypothesis tests to obtain the final classification (aiming to address class imbalance).	AUC = 0.867 (0.031)
Ghassemi et al. [27], 2015	In-hospital mortality on discharge and 1-year post-discharge	10,202 patients	Multi-task Gaussian process (MTGP) Transforming ICU patient clinical notes into timeseries and using MTGP hyperparameters from these time-series as features to predict mortality probability.	AUC = 0.812 (hospital mortality); AUC = 0.686 (1-year post-discharge)

Continued on the next page.

Table 1. Continued

Study, year	Outcome	Number of patients/stays	Method	Main results
Guo et al. [30], 2021	72 hours mortality, in-hospital mortality, 30 days mortality and 1 year mortality	42,145 patients	Dynamic ensemble learning algorithm based on K-means	AUC = 0.87 (72 hours mortality); AUC = 0.842 (in-hospital mortality); AUC = 0.861 (30 days mortality); AUC = 0.829 (1-year mortality)
Kim et al. [31], 2011	Mortality at ICU discharge	38,474 admissions	Decision tree (DT) algorithm, artificial neural network (ANN), and support vector machine (SVM)	DT (AUC= 0.892); ANN (AUC = 0.874); SVM (AUC = 0.876)
Awad et al. [32], 2017	In-hospital mortality using only the first 6 hours in the ICU	11,722 patients	Ensemble learning random forest model	AUC = 0.82
Pirracchio et al. [33], 2015	In-hospital mortality	24,508 patients	Bayesian additive regression tree (BART)	AUC = 0.88
Moser et al. [34], 2021	In-hospital mortality	61,224 admissions	Hierarchical logistic regression model	AUC = 0.886
El-Rashidy et al. [35], 2020	In-hospital mortality at 24 hours	10,664 patients	Stacking ensemble classifier	AUC = 0.933
Badawi et al. [36], 2018	Mortality within 24 hours in the ICU	563,470 patients	Multivariable logistic regression	AUC = 0.84
Chiu et al. [37], 2022	In-hospital, within 48 or 72 hours, 30 days, 1 year	46,520 patients	Latent Dirichlet allocation (LDA) model to classify text in the semi-structured data of some particular topics, followed by classification (gradient boosting)	AUC = 0.93 for 48 hours mortality; AUC = 0.87 for 30-day mortality
Iwase et al. [38], 2022	ICU mortality	12,747 patients	Random forest classifier	AUC = 0.945
Pang et al. [39], 2022	ICU mortality	67,748 patients	Boosting (XGBoost)	AUC = 0.918
Safaei et al. [40], 2022	Mortality on discharge (analyzed per disease group)	200,000 patients	Boosting	AUC = 0.86–0.92
Stenwig et al. [41], 2022	Hospital mortality	129,794 patients	Random forest (among other comparison methods)	AUC = 0.87
Zhao et al. [42], 2023	Mortality within 1 week	12,393 patients	Boosting (XGBoost)	AUC =0 .97
Meiring et al. [43], 2018	ICU mortality	22,514 patients	Deep learning	AUC = 0.883 (after 1st day); AUC = 0.895 (after 2nd day)
Davoodi et al. [44], 2018	ICU mortality	10,972 patients	Deep rule based fuzzy classifier	AUC = 0.739 (using first 48 hours)
Marafino et al. [45], 2018	In-hospital mortality	101,196 patients	Augmenting labs and vitals with clinical trajectory and NLP-derived terms	AUC = 0.922

The area under the curve (AUC) metric is widely used to assess the performance of mortality prediction models.

ICU: intensive care unit, NLP: natural language processing.

ample, proposed a binary classifier consisting of skewness-based transformation of input features and statistical hypothesis tests to obtain the final classification. The balanced random forest (BRF) algorithm was used by Li et al. [48] to address class imbalance with promising results.

2. Length of Stay Benchmarking Using ML

Length of stay (LOS) predictions can be used for planning, identifying individuals with unexpectedly long (or short) LOS, and benchmarking. The models for LOS prediction enable case-mix correction when comparing LOS between ICUs, hospitals, or even health systems across geographic regions. Verburg et al. conducted a systematic review of models that can be used for predicting LOS [9]. They identified 11 studies describing the development of 31 prediction models and three studies describing the external validation of one of these models. For benchmarking, they concluded that none of the models satisfied their criteria for performance with the exception of the original [49] and the second-order recalibration of APACHE [50]. However, none of the models considered fulfilled their requirements for moderate calibration. It is worth noting that the models reviewed were multivariable linear models, which assume linearity between LOS and its covariates or predictors. This assumption might not capture the complexity of the relationship; although patients with more severe illness tend to have a longer LOS, they also have a higher mortality risk, which could lead to shorter stays. Another important observation highlighted by Verburg et al. [9] is that LOS distributions are asymmetrical (right-skewed) and present multimodality, since patient discharge tends to occur at particular times of day.

Given the above, nonlinear models have shown promising results in LOS prediction versus linear/statistical models. The recent review by Peres et al. [51] covers several approaches that have been proposed to address some disadvantages of using linear models. Another recent review [52] focused on the use of ML for predicting medical inpatient LOS with a focus on non-ICU patients. In this report, we focus on the use of ML for predicting LOS for ICU patients.

It is important to highlight that methods can be categorized into two types: regression, which involves predicting LOS as a continuous result, and classification, which revolves around categorizing patients into distinct groups. These categories might encompass distinctions like extended stays versus brief ones. The results for LOS regression models are usually assessed using the R-squared error, root-mean-squared error (RMSE), and mean absolute error (MAE). The concordance correlation coefficient is also presented in some studies. The classification results (a long stay, for example) are usually presented using the AUC metric, as well as sensitivity, specificity, and prediction accuracy. Recent studies have proposed the use of classification models that convert length of stay into a binary or multi-class problem and classify LOS into smaller buckets [53].

Table 2 shows a set of different ML approaches for the prediction of both ICU and hospital LOS for critical care patients [38,53-63]. Below we summarize some of our observations:

Preprocessing: To deal with the asymmetric nature of the LOS distribution, preprocessing in some studies included log transformation [54] and Z-score normalization. In a regression application, a log transformation can be seen as modeling LOS via a Poisson or negative binomial regression model, among others. Studies in other areas have used methods that can deal with a skewed distribution; an example is the use of gamma mixture models that were applied to maternal hospital LOS [64].

Features: In most of the studies above, the features or predictors used focused on data readily available in the ICU, such as labs and patient demographics. However, Houthouft et al. [54] combined the raw data available in the first 5 ICU days with sequential organ failure assessment scores, as well as sub-scores created to assess the performance of different physiological systems, such as renal, cardiovascular, and respiratory systems. Recently, Peres et al. [65] surveyed risk factors that have been used in ICU LOS prediction and suggested that a list of risk factors should be considered in prediction models for ICU LOS. These factors included severity scores, mechanical ventilation, hypomagnesemia, delirium, malnutrition, infection, trauma, red blood cell count, and PaO_2:FiO_2 ratios.

Models: As in the mortality prediction case, it is difficult to compare model performance because the datasets used were different, with different sizes, patient groups, and geographic regions. However, turning the problem into a classification issue showed better results. Harutyunyan et al. [53] showed an AUC of 0.84 for predicting ICU LOS >7 days using channel-wise long short-term memory units (LSTMs) and multi-task training, whereas Ma et al. [58] had an AUC of 0.85 for predicting LOS >10 days using just-in-time learning (JITL) and one-class extreme learning machine (note that their study included only 4000 patients). In the studies of Iwase et al. [38] and Peres et al. [62], the authors used random forest models and achieved good classification accuracy for short and long ICU stays with an AUC larger than 0.87. Houthouft et al. [54] approached the issue by initially transforming it

 https://doi.org/10.4258/hir.2023.29.4.301

Table 2. Summary of studies using machine learning to predict length of stay (LOS)

Study, year	Outcome	Number of patients/stays	Method	Main results
Houthouft et al. [54], 2015	Long LOS* (over 10 days) plus ICU LOS prediction**	14,480 patients	SVM: This work uses data from the first 5 days of ICU stay	For predicting patient mortality and a prolonged stay (>10 days), the best performing model is a SVM with an AUC = 0.82. In terms of LOS regression, the best performing model is support vector regression, with MAE of 1.79 days for patients surviving a non-prolonged stay.
Li et al. [55], 2019	ICU LOS**	1,214 unplanned ICU admissions	Least absolute shrinkage and selection operator (LASSO) algorithm	0.88 day for RMSE, 0.87 day for MAE, and 0.35 ± 0.09 for R-squared
Sotoodeh et al. [57], 2019	ICU LOS**	4,000 ICU patients	Hidden Markov models	RMSE = 9.48 days
Harutunyan et al. [53], 2019	ICU LOS**	42,276 ICU stays of 33,798 unique patients	Recurrent neural network framework (channel-wise LSTMs and multitask training)	AUC = 0.84 for predicting extended LOS (>7 days) at 24 hours after admission
Gentimis et el. [57], 2017	ICU LOS* (>5 days), or short (≤5 days)	31,018 patients	Neural networks	80% prediction accuracy
Ma et al. [58], 2020	Hospital LOS* (more or less than 10 days)	4,000 patients	Just-in-time learning (JITL) and one-class extreme learning machine	AUC = 0.85 (accuracy, specificity, and sensitivity were 0.82, 1, and 0.62 respectively)
Muhlestein et al. [59], 2019	Hospital LOS** following brain surgery	41,222 patients	Ensemble model: Top-performing algorithms were the gradient-boosted tree (GBT) and SVR; these models were combined with an elastic net to create an ensemble model	The ensemble model predicted LOS with RMSE of 0.56 days on internal validation and 0.63 days on external validation
Wu et al. [60], 2021	ICU LOS*	139,367 patients (eICU dataset), external validation (MIMIC); 38,597 adult patients	Comparison-best results obtained by a gradient boosting decision tree	AUC = 0.742
Iwas et al. [38], 2022	ICU LOS*	12,747 patients	Random forest	Predictive value for long ICU stays (AUC = 0.881), short ICU stays (AUC = 0.889)
Alghatani et al. [61], 2021	ICU LOS*	53,423 patients	Random forest	AUC = 0.65 (binary classification as less than 2.64 days or more)
Peres et al. [62], 2022	ICU LOS*	99,492 admissions	Stacking model combining random forests and linear regression	AUC = 0.87 for short and long stays
Weissman et al. [63], 2018	ICU LOS*	25,947 admissions	Gradient boosting including unstructured clinical text data	AUC = 0.89 (for stays > 7 days)

ICU: intensive care unit, SVM: support vector machine, MAE: mean absolute error, RMSE: root mean square error, LSTM: long short-term memory, MIMIC: Medical Information Mart for Intensive Care, AUC: area under the curve.

"*" indicates classification and "**" regression.

into a classification task to identify patients with extended stays (beyond 10 days). Subsequently, they tackled the stays shorter than 10 days as a regression problem, achieving a MAE of 1.79 days for this subgroup.

3. Mechanical Ventilation Management Benchmarking

Although less investigated than either mortality or LOS, the last few years have seen several new approaches for the prediction of both probability and duration of mechanical ventilation (Table 3). Note that unlike mortality and LOS, where our focus was on models for generic ICU patients, few papers predicted ventilation for all groups, so we present work that focused on certain cohorts (such as acute respiratory distress syndrome and COVID-19 patients). Other important areas where ML is used are ventilation weaning and extubation outcomes. However, they are out of the scope of this review.

Much like the case of predicting mortality, drawing direct comparisons between the methods outlined in Table 3 proves challenging when considering the results alone [11,12,66-72]. The variability in cohorts and target variables, such as distinguishing between ventilation duration and

Table 3. Summary of studies using machine learning to predict ventilation (probability and duration)

Study, year	Outcome	Number of patients/stays	Method	Main results
Sayed et al. [66], 2021	MV duration after ARDS onset	Two cohorts from different databases: Set 1: 2,466 (MIMIC-III), Set 2: 5,153 (eICU database)	Light-gradient boosting machine	RMSE: Set 1: 6.10 days, Set 2: 5.87 days
Seneff et al. [11], 1996	MV duration	42 ICU, 40 hospitals, 17,400 ICU admission, 6,000 patients with MV	Multivariate regression analysis	RMSE: 8.01 days
Kramer et al. [67], 2016	MV duration	56,336 patients	Multivariable logistic regression model	For individual patients, the difference between observed and predicted mean duration of MV: 3.3 hours (95% CI, 2.8–3.9) with R-squared equal to 21.6%
Kulkarni et al. [68], 2021	Probability of MV for COVID-19 patients	528 patients (X-ray images)	Deep learning	90% accuracy
Yu et al. [69], 2021	Probability of MV for COVID-19 patients based on ER data	1,980 patients	Boosting (XGBoost)	85% accuracy
Shashikumar et al. [12], 2021	Probability of MV (including COVID-19 patients)	30,000 ICU patients	Deep learning	AUC = 0.895 vs. 0.882, development and validation sites
Douville et al. [70], 2021	Probability of MV for COVID-19 patients	398 patients	Random forest model	AUC = 0.858
Karri et al. [71], 2022	Probability of MV for COVID-19 patients	300 admissions	Random forest model/Gradient boosting	AUC = 0.69 (Random forest); AUC = 0.68 (Gradient boosting)
Parreco et al. [72], 2018	Predicting prolonged mechanical ventilation (over 7 days) for ICU patients	20,262 ICU stays	Gradient boosting algorithms	AUC = 0.852

MV: mechanical ventilation, ICU: intensive care unit, MIMIC: Medical Information Mart for Intensive Care, RMSE: root mean square error, COVID-19: coronavirus disease 2019, ER: emergency room, AUC: area under the curve, CI: confidence interval.

 https://doi.org/10.4258/hir.2023.29.4.301

daily/entire-stay probabilities, contributes to this complexity. Some observations are as follows:

Features: The use of imaging (X-rays), especially for COVID-19 patients [68], provides a new source of data that can be leveraged for increased precision, especially when combined with deep learning approaches. Other studies depended on more standard features, such as patient characteristics, baseline comorbidities, vital signs, laboratory values, medication administration records, and processes of care.

Models: For ventilation duration, gradient boosting showed promising results [66,72], as did simpler methods like multivariable log regression models [67]. For ventilation probability, the choice of methods depended on the targets and features. Deep learning was used for X-ray imaging datasets as well as for clinical features, with promising results in both cases.

IV. Discussion

ML has provided a novel means of benchmarking critical care through utilizing the power of large datasets and improved algorithms for outcome prediction. However, despite the plethora of articles appearing in the last two decades, the comparison of results and performance remains challenging. Despite some attempts to offer unified datasets for comparison [21], many of the models are developed on different databases, which may be country-specific, be disease/cohort-specific, or even target different outcomes (such as mortality in the ICU, hospital, or after release). Several ICU databases have recently been shared publicly, which can facilitate the comparison of modeling approaches [73].

The studies reviewed showed a variety of inputs or predictors used. Traditionally, features were hand-crafted and included demographics, characteristics, input diagnoses, labs, and vitals. However, we have recently seen more studies that devote less effort to fine-tuning, features yet achieve good results based on learning from raw data [32,33]. New predictors have also been added, such as imaging [68], clinical notes, and premorbid functional status [34], which show improvements in outcome prediction.

In terms of the models selected, the studies show a large variety, including support vector machines, gradient boosting, hidden Markov models, and deep learning. Model selection is affected by performance, data size, the handling of missing/erroneous data, and interpretability. Multi-task learning is an interesting direction because it improves generalization by leveraging the domain-specific information contained in the training signals of related tasks. Harutyunyan et al. [53] applied a deep learning multi-task learning framework to cover a range of clinical problems, including modeling risk of mortality, forecasting LOS, detecting physiologic decline, and classifying phenotype. In terms of interpretability, methods such as Shapley values can be used to convey the importance that an ML model assigns to input features [23,24].

Second, a significant concern during the training and evaluation of benchmarking models is class imbalance, a phenomenon evident across all the clinical outcomes examined for the current study. This imbalance is particularly pronounced in cases of mortality, as a relatively small subset of critical care patients experience death. Furthermore, this issue extends to recent studies that assess the efficacy of established LOS models. Interestingly, these models do not distinguish between patients who have survived and those who have not, leading to the overrepresentation of surviving and lower-risk patients [74]. We presented some approaches that addressed class imbalance, such as skewness-based transformations [29] and balanced random forest algorithms [48]. We point the reader to several reviews on this active research area [75,76].

Third, one factor contributing to differences among studies relates to how stays are defined and consolidated. While in a hospital setting, patients could experience multiple instances of being discharged and readmitted to the ICU. To maintain uniformity in model development, it becomes essential to define standardized criteria for classifying these occurrences as either one continuous stay or several separate stays. This effort aims to reduce variation in the resulting models. An associated subject pertains to ICU type, encompassing different patient groups, treatments, and results, such as cardiac care versus neurological cases. Despite its significance warranting deeper exploration, the published literature shows limited emphasis in this domain.

A further issue that could affect model performance is that some sub-populations, such as ethnic minorities, may be underrepresented even in large datasets. Other sources of bias that could influence performance are related to variations in documentation across sites and geographic regions, due mostly to subjective evaluation. Both the reason for admission to the ICU and the Glasgow Coma Score, for example, may incorporate subjective evaluation from clinicians [77].

Most existing benchmarking models were developed on country-specific databases. The APACHE scores, for example, were United States-trained and tested. However, clinical practice, documentation, and patient diversity differ across geographic regions, requiring model recalibration and training.

As in other fields, ML in benchmarking ICU outcomes has focused on developing models with improved performance on retrospective data. However, little work has occurred on long-term validation post-deployment, which would observe data drift, model drift, and performance over time. Existing models merit recalibration every few years due to data drift. Some reasons for data drift in critical care include changes in data due to seasonality, changes in documentation practices, the addition of new devices, missing data, and changes in clinical practices over time. The same applies to bias, model generalizability, and fairness [78]. Federated learning, a distributed technique for training ML models without exchanging data, presents an intriguing paradigm for locations where data sharing is not feasible or for refining models using local datasets for updates [79].

Although the models highlighted in the current review attempt to adjust for measured risk factors, unobserved patient attributes mean that risk adjustment is never perfect [80]. Areas such as medication adherence, social support, or mobility before admission can be considered as unmeasured factors. Even when models are accurately calibrated to the collected data, the influence of these factors continues to impact the results. Finding ways to incorporate some of these factors, possibly through clinical notes and patient interactions, remains crucial. This could emerge as a thriving research domain for large language models or generative artificial intelligence (AI) methods to offer a potential solution that would bridge this gap.

In conclusion, ML has provided novel tools for benchmarking critical care outcomes, leading to improved results as well as addressing important drawbacks of previous methods, such as reducing biases due to documentation, missing data, and class imbalance, as well as modeling non-linear relationships between variables and outcomes. Prospects exist for using ML to encompass a broader array of data types, including imaging, medical notes, and diagnoses. The utilization of multi-national datasets via techniques like federated learning could also prove advantageous in developing models that find broader relevance across diverse patient groups and geographic regions where data sharing is not possible. Generative AI and large language models present a fresh approach for scrutinizing extensive datasets, including medical notes, thereby enhancing the efficacy of future ML models within this domain. In clinical contexts, we suggest that healthcare practitioners opt for well-validated models tailored to their specific geographic and patient-demographic considerations.

Conflict of Interest

This work was funded by Philips Healthcare and all authors are fully employed by Philips. The authors have no competing interests.

Acknowledgments

The authors would like to thank Dr Omar Badawi and Robin French for their ideas on bottle necks in benchmarking critical care outcomes based on their extensive experience.

ORCID

Louis Atallah (https://orcid.org/0000-0002-6657-319X)
Mohsen Nabian (https://orcid.org/0000-0001-6787-4827)
Ludmila Brochini (https://orcid.org/0000-0002-4017-0157)
Pamela J. Amelung (https://orcid.org/0000-0002-4023-128X)

References

1. Rhodes A, Moreno RP, Azoulay E, Capuzzo M, Chiche JD, Eddleston J, et al. Prospectively defined indicators to improve the safety and quality of care for critically ill patients: a report from the Task Force on Safety and Quality of the European Society of Intensive Care Medicine (ESICM). Intensive Care Med 2012;38(4):598-605. https://doi.org/10.1007/s00134-011-2462-3
2. Vincent JL, Marshall JC, Namendys-Silva SA, Francois B, Martin-Loeches I, Lipman J, et al. Assessment of the worldwide burden of critical illness: the intensive care over nations (ICON) audit. Lancet Respir Med 2014;2(5):380-6. https://doi.org/10.1016/S2213-2600(14)70061-X
3. Higgins TL. Quantifying risk and benchmarking performance in the adult intensive care unit. J Intensive Care Med 2007;22(3):141-56. https://doi.org/10.1177/0885066607299520
4. Braun JP, Kumpf O, Deja M, Brinkmann A, Marx G, Bloos F, et al. The German quality indicators in intensive care medicine 2013: second edition. Ger Med Sci 2013;11:Doc09. https://doi.org/10.3205/000177
5. Kumpf O, Braun JP, Brinkmann A, Bause H, Bellgardt M, Bloos F, et al. Quality indicators in intensive care medicine for Germany: third edition 2017. Ger Med Sci 2017;15:Doc10. https://doi.org/10.3205/000251
6. Brown SE, Ratcliffe SJ, Halpern SD. An empirical comparison of key statistical attributes among potential ICU quality indicators. Crit Care Med 2014;42(8):1821-31.

https://doi.org/10.4258/hir.2023.29.4.301

https://doi.org/10.1097/CCM.0000000000000334
7. Salluh JIF, Soares M, Keegan MT. Understanding intensive care unit benchmarking. Intensive Care Med 2017;43(11):1703-7. https://doi.org/10.1007/s00134-017-4760-x
8. Higgins TL, Stark MM, Henson KN, Freeseman-Freeman L. Coronavirus disease 2019 ICU patients have higher-than-expected Acute Physiology and Chronic Health Evaluation-adjusted mortality and length of stay than viral pneumonia ICU patients. Crit Care Med 2021;49(7):e701-6. https://doi.org/10.1097/ccm.0000000000005012
9. Verburg IW, Atashi A, Eslami S, Holman R, Abu-Hanna A, de Jonge E, et al. Which models can I use to predict adult ICU length of stay? A systematic review. Crit Care Med 2017;45(2):e222-31. https://doi.org/10.1097/ccm.0000000000002054
10. van Sluisveld N, Bakhshi-Raiez F, de Keizer N, Holman R, Wester G, Wollersheim H, et al. Variation in rates of ICU readmissions and post-ICU in-hospital mortality and their association with ICU discharge practices. BMC Health Serv Res 2017;17(1):281. https://doi.org/10.1186/s12913-017-2234-z
11. Seneff MG, Zimmerman JE, Knaus WA, Wagner DP, Draper EA. Predicting the duration of mechanical ventilation. The importance of disease and patient characteristics. Chest 1996;110(2):469-79. https://doi.org/10.1378/chest.110.2.469
12. Shashikumar SP, Wardi G, Paul P, Carlile M, Brenner LN, Hibbert KA, et al. Development and prospective validation of a deep learning algorithm for predicting need for mechanical ventilation. Chest 2021;159(6):2264-73. https://doi.org/10.1016/j.chest.2020.12.009
13. Malmgren J, Waldenstrom AC, Rylander C, Johannesson E, Lundin S. Long-term health-related quality of life and burden of disease after intensive care: development of a patient-reported outcome measure. Crit Care 2021;25(1):82. https://doi.org/10.1186/s13054-021-03496-7
14. Higgins TL, Teres D, Nathanson B. Outcome prediction in critical care: the Mortality Probability Models. Curr Opin Crit Care 2008;14(5):498-505. https://doi.org/10.1097/mcc.0b013e3283101643
15. Ferrando-Vivas P, Jones A, Rowan KM, Harrison DA. Development and validation of the new ICNARC model for prediction of acute hospital mortality in adult critical care. J Crit Care 2017;38:335-9. https://doi.org/10.1016/j.jcrc.2016.11.031
16. Keuning BE, Kaufmann T, Wiersema R, Granholm A, Pettila V, Moller MH, et al. Mortality prediction models in the adult critically ill: a scoping review. Acta Anaesthesiol Scand 2020;64(4):424-42. https://doi.org/10.1111/aas.13527
17. Strand K, Flaatten H. Severity scoring in the ICU: a review. Acta Anaesthesiol Scand 2008;52(4):467-78. https://doi.org/10.1111/j.1399-6576.2008.01586.x
18. Siontis GC, Tzoulaki I, Ioannidis JP. Predicting death: an empirical evaluation of predictive tools for mortality. Arch Intern Med 2011;171(19):1721-6. https://doi.org/10.1001/archinternmed.2011.334
19. Raj R, Luostarinen T, Pursiainen E, Posti JP, Takala RS, Bendel S, et al. Machine learning-based dynamic mortality prediction after traumatic brain injury. Sci Rep 2019;9(1):17672. https://doi.org/10.1038/s41598-019-53889-6
20. Subudhi S, Verma A, Patel AB, Hardin CC, Khandekar MJ, Lee H, et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. NPJ Digit Med 2021;4(1):87. https://doi.org/10.1038/s41746-021-00456-x
21. 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 (2010) 2012;39:245-8.
22. Barboi C, Tzavelis A, Muhammad LN. Comparison of severity of illness scores and artificial intelligence models that are predictive of intensive care unit mortality: meta-analysis and review of the literature. JMIR Med Inform 2022;10(5):e35293. https://doi.org/10.2196/35293
23. 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(4):e179-91. https://doi.org/10.1016/s2589-7500(20)30018-2
24. Caicedo-Torres W, Gutierrez J. ISeeU: visually interpretable deep learning for mortality prediction inside the ICU. J Biomed Inform 2019;98:103269. https://doi.org/10.1016/j.jbi.2019.103269
25. Aczon MD, Ledbetter DR, Laksana E, Ho LV, Wetzel RC. Continuous prediction of mortality in the PICU: a recurrent neural network model in a single-center dataset. Pediatr Crit Care Med 2021;22(6):519-29. https://doi.org/10.1097/pcc.0000000000002682
26. Grnarova P, Schmidt F, Hyland SL, Eickhoff C. Neural document embeddings for intensive care patient mor-

tality prediction [Internet]. Ithaca (NY): arXiv.org; 2016 [cited at 2023 Sep 30]. Available from: http://arxiv.org/abs/1612.00467.
27. Ghassemi M, Pimentel MA, Naumann T, Brennan T, Clifton DA, Szolovits P, et al. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. Proc AAAI Conf Artif Intell 2015;2015:446-53.
28. Badawi O, Breslow MJ. Readmissions and death after ICU discharge: development and validation of two predictive models. PLoS One 2012;7(11):e48758. https://doi.org/10.1371/journal.pone.0048758
29. Bhattacharya S, Rajan V, Shrivastava H. ICU mortality prediction: a classification algorithm for imbalanced datasets. Proc AAAI Conf Artif Intell 2017;31(1):1288-94. https://doi.org/10.1609/aaai.v31i1.10721
30. Guo C, Liu M, Lu M. A dynamic ensemble learning algorithm based on K-means for ICU mortality prediction. Appl Soft Comput 2021;103:107166. https://doi.org/10.1016/j.asoc.2021.107166
31. Kim S, Kim W, Park RW. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Healthc Inform Res 2011;17(4):232-43. https://doi.org/10.4258/hir.2011.17.4.232
32. Awad A, Bader-El-Den M, McNicholas J, Briggs J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. Int J Med Inform 2017;108:185-95. https://doi.org/10.1016/j.ijmedinf.2017.10.002
33. 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(1):42-52. https://doi.org/10.1016/s2213-2600(14)70239-5
34. Moser A, Reinikainen M, Jakob SM, Selander T, Pettila V, Kiiski O, et al. Mortality prediction in intensive care units including premorbid functional status improved performance and internal validity. J Clin Epidemiol 2022;142:230-41. https://doi.org/10.1016/j.jclinepi.2021.11.028
35. El-Rashidy N, El-Sappagh S, Abuhmed T, Abdelrazek S, El-Bakry HM. Intensive care unit mortality prediction: an improved patient-specific stacking ensemble model. IEEE Access 2020;8:133541-64. https://doi.org/10.1109/ACCESS.2020.3010556
36. Badawi O, Liu X, Hassan E, Amelung PJ, Swami S. Evaluation of ICU risk models adapted for use as continuous markers of severity of illness throughout the ICU stay. Crit Care Med 2018;46(3):361-7. https://doi.org/10.1097/ccm.0000000000002904
37. Chiu CC, Wu CM, Chien TN, Kao LJ, Qiu JT. Predicting the mortality of ICU patients by topic model with machine-learning techniques. Healthcare (Basel) 2022;10(6):1087. https://doi.org/10.3390/healthcare10061087
38. Iwase S, Nakada TA, Shimada T, Oami T, Shimazui T, Takahashi N, et al. Prediction algorithm for ICU mortality and length of stay using machine learning. Sci Rep 2022;12(1):12912. https://doi.org/10.1038/s41598-022-17091-5
39. Pang K, Li L, Ouyang W, Liu X, Tang Y. Establishment of ICU mortality risk prediction models with machine learning algorithm using MIMIC-IV database. Diagnostics (Basel) 2022;12(5):1068. https://doi.org/10.3390/diagnostics12051068
40. Safaei N, Safaei B, Seyedekrami S, Talafidaryani M, Masoud A, Wang S, et al. E-CatBoost: an efficient machine learning framework for predicting ICU mortality using the eICU Collaborative Research Database. PLoS One 2022;17(5):e0262895. https://doi.org/10.1371/journal.pone.0262895
41. Stenwig E, Salvi G, Rossi PS, Skjaervold NK. Comparative analysis of explainable machine learning prediction models for hospital mortality. BMC Med Res Methodol 2022;22(1):53. https://doi.org/10.1186/s12874-022-01540-w
42. Zhao S, Tang G, Liu P, Wang Q, Li G, Ding Z. Improving mortality risk prediction with routine clinical data: a practical machine learning model based on eICU patients. Int J Gen Med 2023;16:3151-61. https://doi.org/10.2147/ijgm.s391423
43. Meiring C, Dixit A, Harris S, MacCallum NS, Brealey DA, Watkinson PJ, et al. Optimal intensive care outcome prediction over time using machine learning. PLoS One 2018;13(11):e0206862. https://doi.org/10.1371/journal.pone.0206862
44. Davoodi R, Moradi MH. Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier. J Biomed Inform 2018;79:48-59. https://doi.org/10.1016/j.jbi.2018.02.008
45. Marafino BJ, Park M, Davies JM, Thombley R, Luft HS, Sing DC, et al. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. JAMA Netw Open 2018;1(8):e185097. https://doi.org/10.1001/jamanet-

workopen.2018.5097
46. Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. J Biomed Inform 2018;83:112-34. https://doi.org/10.1016/j.jbi.2018.04.007
47. Shapley LS. A value for n-person games. In: Kuhn AW, Tucker HW, editors. Contributions to the theory of games (Volume II). Princeton (NJ): Princeton University Press; 1953. p. 307-18.
48. Li L, Liu G. In-hospital mortality prediction for ICU patients on large healthcare MIMIC datasets using class imbalance learning. Proceedings of 2020 5th IEEE International Conference on Big Data Analytics (ICBDA); 2020 May 8-11; Xiamen, China. p. 90-3. https://doi.org/10.1109/ICBDA49040.2020.9101272
49. Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL. Intensive care unit length of stay: benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. Crit Care Med 2006;34(10):2517-29. https://doi.org/10.1097/01.ccm.0000240233.01711.d9
50. Vasilevskis EE, Kuzniewicz MW, Cason BA, Lane RK, Dean ML, Clay T, et al. Mortality probability model III and simplified acute physiology score II: assessing their value in predicting length of stay and comparison to APACHE IV. Chest 2009;136(1):89-101. https://doi.org/10.1378/chest.08-2591
51. Peres IT, Hamacher S, Oliveira FLC, Bozza FA, Salluh JIF. Prediction of intensive care units length of stay: a concise review. Rev Bras Ter Intensiva 2021;33(2):183-7. https://doi.org/10.5935/0103-507x.20210025
52. Bacchi S, Tan Y, Oakden-Rayner L, Jannes J, Kleinig T, Koblar S. Machine learning in the prediction of medical inpatient length of stay. Intern Med J 2022;52(2):176-85. https://doi.org/10.1111/imj.14962
53. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data 2019;6(1):96. https://doi.org/10.1038/s41597-019-0103-9
54. Houthooft R, Ruyssinck J, van der Herten J, Stijven S, Couckuyt I, Gadeyne B, et al. Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores. Artif Intell Med 2015;63(3):191-207. https://doi.org/10.1016/j.artmed.2014.12.009
55. Li C, Chen L, Feng J, Wu D, Zimeng W, Liu J, et al. Prediction of length of stay on the intensive care unit based on least absolute shrinkage and selection operator. IEEE Access 2019;7:110710-21. https://doi.org/10.1109/ACCESS.2019.2934166
56. Sotoodeh M, Ho JC. Improving length of stay prediction using a hidden Markov model. AMIA Jt Summits Transl Sci Proc 2019;2019:425-34.
57. Gentimis T, Ala'J A, Durante A, Cook K, Steele R. Predicting hospital length of stay using neural networks on MIMIC III data. Proceedings of 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 15th International Conference on Pervasive Intelligence and Computing, 3rd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech); 2017 Nov 6-10; Orlando, FL. p. 1194-201. https://doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2017.191
58. Ma X, Si Y, Wang Z, Wang Y. Length of stay prediction for ICU patients using individualized single classification algorithm. Comput Methods Programs Biomed 2020;186:105224. https://doi.org/10.1016/j.cmpb.2019.105224
59. Muhlestein WE, Akagi DS, Davies JM, Chambless LB. Predicting inpatient length of stay after brain tumor surgery: developing machine learning ensembles to improve predictive performance. Neurosurgery 2019;85(3):384-93. https://doi.org/10.1093/neuros/nyy343
60. Wu J, Lin Y, Li P, Hu Y, Zhang L, Kong G. Predicting prolonged length of ICU stay through machine learning. Diagnostics (Basel) 2021;11(12):2242. https://doi.org/10.3390/diagnostics11122242
61. Alghatani K, Ammar N, Rezgui A, Shaban-Nejad A. Predicting intensive care unit length of stay and mortality using patient vital signs: machine learning model development and validation. JMIR Med Inform 2021;9(5):e21347. https://doi.org/10.2196/21347
62. Peres IT, Hamacher S, Cyrino Oliveira FL, Bozza FA, Salluh JI. Data-driven methodology to predict the ICU length of stay: a multicentre study of 99,492 admissions in 109 Brazilian units. Anaesth Crit Care Pain Med 2022;41(6):101142. https://doi.org/10.1016/j.accpm.2022.101142
63. Weissman GE, Hubbard RA, Ungar LH, Harhay MO, Greene CS, Himes BE, et al. Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay. Crit Care Med 2018;46(7):1125-32. https://doi.org/10.1097/ccm.0000000000003148
64. Williford E, Haley V, McNutt LA, Lazariu V. Dealing with highly skewed hospital length of stay distributions: the use of Gamma mixture models to study delivery hospitalizations. PLoS One 2020;15(4):e0231825. https://doi.org/10.1371/journal.pone.0231825

65. Peres IT, Hamacher S, Oliveira FLC, Thome AMT, Bozza FA. What factors predict length of stay in the intensive care unit? Systematic review and meta-analysis. J Crit Care 2020;60:183-94. https://doi.org/10.1016/j.jcrc.2020.08.003
66. Sayed M, Riano D, Villar J. Predicting duration of mechanical ventilation in acute respiratory distress syndrome using supervised machine learning. J Clin Med 2021;10(17):3824. https://doi.org/10.3390/jcm10173824
67. Kramer AA, Gershengorn HB, Wunsch H, Zimmerman JE. Variations in case-mix-adjusted duration of mechanical ventilation among ICUs. Crit Care Med 2016;44(6):1042-8. https://doi.org/10.1097/ccm.0000000000001636
68. Kulkarni AR, Athavale AM, Sahni A, Sukhal S, Saini A, Itteera M, et al. Deep learning model to predict the need for mechanical ventilation using chest X-ray images in hospitalised patients with COVID-19. BMJ Innov 2021;7(2):261-70. https://doi.org/10.1136/bmjinnov-2020-000593
69. Yu L, Halalau A, Dalal B, Abbas AE, Ivascu F, Amin M, et al. Machine learning methods to predict mechanical ventilation and mortality in patients with COVID-19. PLoS One 2021;16(4):e0249285. https://doi.org/10.1371/journal.pone.0249285
70. Douville NJ, Douville CB, Mentz G, Mathis MR, Pancaro C, Tremper KK, et al. Clinically applicable approach for predicting mechanical ventilation in patients with COVID-19. Br J Anaesth 2021;126(3):578-89. https://doi.org/10.1016/j.bja.2020.11.034
71. Karri R, Chen YP, Burrell AJC, Penny-Dimri JC, Broadley T, Trapani T, et al. Machine learning predicts the short-term requirement for invasive ventilation among Australian critically ill COVID-19 patients. Plops One 2022;17(10):e0276509. https://doi.org/10.1371/journal.pone.0276509
72. Parreco J, Hidalgo A, Parks JJ, Kozol R, Rattan R. Using artificial intelligence to predict prolonged mechanical ventilation and tracheostomy placement. J Surg Res 2018;228:179-87. https://doi.org/10.1016/j.jss.2018.03.028
73. Sauer CM, Dam TA, Celi LA, Faltys M, de la Hoz MAA, Adhikari L, et al. Systematic review and comparison of publicly available ICU data sets: a decision guide for clinicians and data scientists. Crit Care Med 2022;50(6):e581-8. https://doi.org/10.1097/ccm.0000000000005517
74. Liu X, Badawi O. 369: ICU length-of-stay models should account for the interaction between survival and patient severity. Crit Care Med 2020;48(1):166. https://doi.org/10.1097/01.ccm.0000619828.57026.19
75. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. J Big Data 2018;5(1):42. https://doi.org/10.1186/s40537-018-0151-6
76. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data 2019;6(1):27. https://doi.org/10.1186/s40537-019-0192-5
77. 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(5):1297-310. https://doi.org/10.1097/01.ccm.0000215112.84523.f0
78. Roosli E, Bozkurt S, Hernandez-Boussard T. Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model. Sci Data 2022;9(1):24. https://doi.org/10.1038/s41597-021-01110-7
79. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res 2021;5(1):1-19. https://doi.org/10.1007/s41666-020-00082-4
80. Lane-Fall MB, Neuman MD. Outcomes measures and risk adjustment. Int Anesthesiol Clin 2013;51(4):10-21. https://doi.org/10.1097/aia.0b013e3182a70a52

https://doi.org/10.4258/hir.2023.29.4.301