Health Informatics Journal

Original Article

Multicenter validation of a machine-learning algorithm for 48-h all-cause mortality prediction

Health Informatics Journal
1–14

DOI: 10.1177/1460458219894494
journals.sagepub.com/home/jhi
SAGE

Hamid Mohamadlou*
Saarang Panchavati*
Jacob Calvert
Anna Lynn-Palevsky
Sidney Le
Angier Allen
Emily Pellegrini
and Abigail Green-Saxena
Dascena, Inc., USA

Christopher Barton
University of California San Francisco, USA

Grant Fletcher
University of Washington, USA

Lisa Shieh
Stanford University, USA

Philip B Stark
University of California, Berkeley, USA

Uli Chettipally
University of California San Francisco, USA; Kaiser Permanente South San Francisco Medical Center, USA

*These authors contributed equally to this work.

Corresponding author:
Jacob Calvert, Dascena, Inc., 414 13th Street, Suite #500, Oakland, CA 94612, USA.
Email: jake@dascena.com

David Shimabukuro
Mitchell Feldman
University of California San Francisco, USA

Ritankar Das
Dascena, Inc., USA

Abstract
In order to evaluate mortality predictions based on boosted trees, this retrospective study uses electronic medical record data from three academic health centers for inpatients 18 years or older with at least one observation of each vital sign. Predictions were made 12, 24, and 48 hours before death. Models fit to training data from each institution were evaluated using hold-out test data from the same institution, and from the other institutions. Gradient-boosted trees (GBT) were compared to regularized logistic regression (LR) predictions, support vector machine (SVM) predictions, quick Sepsis-Related Organ Failure Assessment (qSOFA), and Modified Early Warning Score (MEWS) using area under the receiver operating characteristic curve (AUROC). For training and testing GBT on data from the same institution, the average AUROCs were 0.96, 0.95, and 0.94 across institutional test sets for 12-, 24-, and 48-hour predictions, respectively. When trained and tested on data from different hospitals, GBT AUROCs achieved up to 0.98, 0.96, and 0.96, for 12-, 24-, and 48-hour predictions, respectively. Average AUROC for 48-hour predictions for LR, SVM, MEWS, and qSOFA were 0.85, 0.79, 0.86 and 0.82, respectively. GBT predictions may help identify patients who would benefit from increased clinical care.

Keywords
electronic health record, machine learning, mortality, prediction

Introduction

Timely identification of patients with elevated risk of in-hospital mortality is necessary to best allocate limited and costly hospital resources, focus care to prevent patients from deteriorating, and anticipate probable patient outcomes.[1] This is particularly relevant to emergency departments (EDs), where patient assessments are made across a broad spectrum of patient conditions, with potentially vastly different and dynamic degrees of urgency. Accurate identification of patients who require ICU admission is especially important for resource allocation. While ICU beds account for less than 10 percent of beds in US hospitals, ICU bed use and associated costs continue to rise, nearly doubling between 2000 and 2010.[2,3] Accurate early prediction of deterioration and death can alert medical teams to the need for more aggressive care while also helping to minimize overtreatment of more stable patients, in turn lowering health care costs. Mortality prediction tools that are accurate at long lookahead times have particular promise for providing optimal care and allocating hospital resources.[4,5] A 24-h lookahead time is considered clinically relevant in busy hospital settings[6] and has been incorporated by several studies into standard metrics for evaluating early warning systems.[6–9]

There are several existing mortality prediction tools; they generally use rule-based approaches. Such rule-based systems include the Acute Physiology, Age, Chronic Health Evaluation (*APACHE*),[10] the Modified Early Warning Score (*MEWS*),[11] the Sepsis-Related Organ Failure Assessment (*SOFA*),[12] and the quick SOFA (*qSOFA*) score.[13] The clinical utility of these tools is limited due to inadequate specificity and sensitivity,[1,14] and many rule-based scores, such as MEWS and SOFA,

are still manually tabulated at the bedside,[15] requiring precious time and attention that could be used elsewhere in EDs. These tools weight a collection of patient characteristics at a moment in time, then sum the weighted values to create an overall score intended to reflect the risk. Because they use a single snapshot of a patient's characteristics, the scores do not incorporate trends in patient conditions, which can be useful for predicting patient decline.[16] Changes in these scores may be informative, but a tool that directly incorporates trends in patient vital signs may be more reliable.

A machine-learning mortality prediction tool integrated into an electronic health record (EHR) system can exploit such trends. Machine-learning algorithms (MLAs) can produce a score that depends not only on linear combinations of the input variables but also on nonlinear functions of the variables and temporal changes in the variables. Previous studies have demonstrated that machine-learning mortality risk scores that incorporate temporal information can more accurately predict patient outcomes.[16–18] Moreover, MLAs are readily optimized for different populations, while existing rule-based methods are "one size fits all."[17] Machine-learning-based mortality prediction methods have been tailored to specific critical health conditions, such as sepsis[19,20] and cardiac arrest,[21] as well as specific settings, such as the ICU [2,17,22,23] and ED.[20] Results from a retrospective evaluation of an auotmated risk adjustment algorithm indicated that the tool demonstrated strong discrimination (AUROC = 0.94) for ICU mortality prediction among critical care patients;[22] however, the tool has only been evaluated for the prediction of mortality using data drawn from a common EHR system, and it has not been used to risk adjust for other outcomes, such as discharge recommendation and prediction of unplanned ICU admission. In addition to APACHE,[10,24] other commonly used risk adjustment tools include the Simplified Acute Physiology Score (SAPS)[25] and the Mortality Probability Model (MPM).[26] These mature algorithms have been validated in a variety of studies and have been shown to successfully predict hospital mortality,[22,23,27–29] but have low rates of adoption due to cost-prohibitive licensing fees, as well as high costs associated with intensive data collection. For example, although APACHE-IV and MPM0-III algorithms are available for use in the public domain at no cost, use of the tools in clinical practice generally requires payment of maintenance fees which are not financially feasible for many hospitals.[22] Implementation of common risk adjustment algorithms also requires clinicians to collect and document patient data which is often not readily available in clinical information systems, creating a data collection burden.[30,31] Scoring systems, such as MPM, SAPS, and APACHE, also require constant updates to ensure accuracy and avoid deterioration in model performance.[30] Although APACHE offers high predictive accuracy with unlimited resources, constraints on cost and labor create a need for tools which can provide viable alternatives to risk prediction without substantial losses in accuracy.[31] Therefore, there is a need for mortality risk prediction tools for which implementation is less time and labor-intensive for the clinician.

To this end, we have developed an automated gradient boosted tree (GBT) machine-learning mortality risk prediction tool with 12, 24, and 48 h prediction horizons using patient age and a series of measurements of only six vital signs and Glasgow Coma Scale (GCS) values as input. These data are already routinely entered into the EHR and therefore require no additional work or manual data entry from clinicians. In this investigation, we demonstrate that GBTs (XGBoost) outperform regularized logistic regression (LR) and support vector machine (SVM), in addition to the commonly used MEWS and qSOFA mortality prediction tools. We test predictions on patient data from three academic health centers in the United States to examine the robustness of the predictions across patient populations in different hospitals without customizing the algorithm to each hospital's data. The mortality risk prediction tool we describe is novel because it is developed using only readily available patient EHR data and has been validated for generalizability across a variety of datasets and EHR systems, with limited variation in hospital performance, and with a minimal data collection burden. It demonstrates higher performance in terms of area under the receiver operating characteristic curve (AUROC) than comparable rule-based mortality prediction

Table 1. Predictor variables used in this study.

Demographics	Age
Vital signs	Heart rate
	Respiratory rate
	Peripheral oxygen saturation (SpO_2)
	Temperature
	Systolic blood pressure
	Diastolic blood pressure
Other clinical variables	GCS

GCS: Glasgow Coma Scale.

tools and can be seamlessly integrated into existing clinical workflows. These features allow the tool to circumvent barriers to implementation in clinical settings (i.e. lack of algorithm generalizability and high costs associated with laborious data collection) which are often common to machine-learning prediction tools.

Materials and methods

Data sources

Patient records were collected by Stanford Medical Center in Stanford, California; the University of California, San Francisco Medical Center in San Francisco, CA (UCSF); and University of Washington Medical Center (UW) in Seattle, Washington. Stanford data include records from 515,452 inpatients from all hospital wards from December 2008 to May 2017. UCSF data contain information on 95,869 inpatients across all hospital wards from the Mount Zion, Mission Bay, and Parnassus Heights medical campuses. We used inpatient data from June 2011 to March 2016 drawn from patient EHR charts. UW data include records from 32,936 adult patients from all hospital wards from January 2014 to March 2017. See Table 1 for details on data inputs.

Data collection was passive and had no impact on patient safety. All data were de-identified in compliance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Studies performed on the de-identified data constitute non-human subject studies, and therefore, our study did not require Institutional Review Board approval.

Data processing

For all three data sources, we included only records for patients aged 18 years or older who had at least one recorded observation of each required measurement (heart rate, respiratory rate, peripheral oxygen saturation (SpO_2), temperature, systolic blood pressure, diastolic blood pressure, and GCS). These covariates were chosen because they are routinely entered into the EHR and would be available at the time of mortality prediction. We excluded patient records for which there were no raw data or no discharge or death dates. This resulted in 31,292 patients from Stanford, 47,289 patients from UCSF, and 32,878 patients from UW.

We minimally processed raw EHR data to generate features. Separate case–control matching experiments were performed to demonstrate that even when trained using standard procedures, models maintained comparable performance when evaluated on a case–controlled test set across three sites. Following EHR data extraction and imputation of missing values, we obtained one value for each measurement, each hour, for up to 3 h preceding prediction time. We also calculated

Table 2. Inclusion criteria for patients in the Stanford, UCSF, and UW datasets.

	Stanford	UCSF	UW
Total patients	515,452	95,869	32,936
Patients with raw data	441,211	95,869	32,936
Patients with discharge or death and age data available and ⩾18 years of age	150,838	95,869	32,936
Patients with at least one observation of each required measurement[a]	31,292	47,289	32,878
Patients with fewer than two mode fractions >0.90	24,614	46,980	32,718
Patients used to train or test the classifier	24,614	46,980	32,718

UCSF: University of California, San Francisco Medical Center in San Francisco, CA; UW: University of Washington Medical Center.

All patients who met the final inclusion criteria were included in training or testing sets for one or more experiments in this study.

[a]Required measurements include heart rate, respiratory rate, peripheral oxygen saturation (SpO_2), temperature, systolic blood pressure, diastolic blood pressure, and Glasgow Coma Scale (GCS).

differences between the current hour and the prior hour and between the prior hour and the hour before that. We concatenated these values from each measurement into a feature vector.

All data were discretized into 1-h intervals, beginning at the time of the first recorded patient measurement, and hourly measurements were required for each input variable. Measurements were averaged to produce a single value in cases when multiple observations of the same patient measurement were taken within a given hour. This ensures that the measurement rate was the same across patients and across time. Missing values were imputed by carrying forward the most recent past measurement in cases where no measurement of a clinical variable was available for a given hour. For some patients with infrequent measurements of one or more vital signs (not including GCS), this simple imputation resulted in many consecutive hours with identical values. We excluded patients for whom measurements of two or more vital signs were missing 90 percent or more of the time (inferred from the fraction of identical values after imputation), leaving 24,614 patients from Stanford, 46,980 patients remaining from UCSF, and 32,718 patients from UW. Table 2 lists the number of patients sequentially meeting each inclusion criterion.

Our previous publication on the use of GBTs for sepsis detection and prediction describes the data processing in further detail.[15] Predictions were generated for all experiments using the variables described in Table 1, including patient age. These measurements were selected for use in this study because they are readily available at the patient bedside.

Gold standard

The outcome of interest was in-hospital patient mortality, which was determined retrospectively for each patient. In the Stanford and UW datasets, in-hospital mortality was indicated by a death date field for each patient. In the UCSF dataset, we used the in-hospital mortality field for each patient. In all, 1810 patients of 46,980 from UCSF (3.85%), 263 patients of 24,614 from Stanford (1.07%), and 965 patients of 32,718 from UW (2.95%) died in the hospital. However, as the end of the next section describes, the mortality rates are effectively higher during training and testing.

The MLA

The classifier was created using the XGBoost method for fitting "boosted" decision trees. We applied the XGBoost package for Python[32] to the patient age and vital sign measurements and their

temporal changes, where temporal changes included hourly differences between each measurement beginning 3 h before prediction time. Gradient boosting, which XGBoost implements, is an ensemble learning technique that combines results from multiple decision trees to create prediction scores. Each tree splits the patient population into smaller and smaller groups, successively. Each branch splits the patients who enter it into two groups, based on whether their value of some covariate is above or below some threshold—for instance, a branch might divide patients according to whether their temperature is above or below 100°F. After some number of branches, the tree ends in a set of "leaves." Each patient is in exactly one leaf, according to the values of his or her measurements. Each "leaf" of the tree is predicted to have the same risk of mortality. The covariate involved in each split and the threshold value are selected by an algorithm designed to trade off fit to the training data and accuracy on out-of-sample data by using cross-validation to avoid "overfitting." We restricted tree depth to a maximum of six branching levels, set the learning rate parameter of XGBoost to 0.05, and restricted the tree ensembles to 200 trees to limit the computational burden.

For all machine-learning methods used in this study, including gradient boosting and both comparators, hyperparameter optimization was performed using cross-validated grid search. We included a hyperparameter for the early stopping of the iterative tree-addition procedure to prevent overfit of the model on the training data and optimized across this hyperparameter using fivefold cross-validation. Due to computational and time constraints, hyperparameter optimization was performed across a sparse parameter grid, where the candidate hyperparameter values were chosen to span large ranges of viable parameter space. Cross-validated grid search was conducted to determine the optimal combination of candidate hyperparameters. While XGBoost has a large number of trainable parameters, computational and time constraints limited the set of parameters to be tuned to just those parameters with the largest impact on performance on the training data and most relevant to the prediction task.

To validate the boosted tree predictor when training and testing was performed on data from the same institution, we used fivefold cross-validation. For each model, four-fifths of the patients were randomly selected to train the model and the remaining one-fifth were used as a hold-out set to test the predictions. To account for the random selection of the training set, reported performance metrics are the average performance of the five separately trained models arising from fivefold cross-validation, each of which was trained on four-fifths of the data and tested on the remaining fifth. For AUROC, we also reported the standard deviation of the five AUROC values obtained from cross-validation.

We modeled mortality 12, 24, and 48 h before death to evaluate the performance with a variety of lead times. For negative class examples, because there was no mortality event from which a specific, applicable time-point could be computed for survivors, we used their time of discharge. Predictors were trained independently for each distinct lookahead time. In 12, 24, and 48 h long lookahead predictions following a 3-h window of measurements, patients must have data for, respectively, 15, 27, or 51 respective hours preceding the time of in-hospital mortality or the time of discharge. Accordingly, we selected patients with the appropriate stays for the training and testing of each lookahead. This resulted in mortality rates of 4.0–4.5 percent for UCSF patients, 3.4–3.6 percent for Stanford patients, and 3.0–3.8 percent for UW patients.

Comparison to other machine-learning and rule-based methods

We tested two other classification algorithms, LR and SVM, and compared their performance to that of XGBoost. The hyperparameters which control the regularization strength of the LR and SVM predictors were set to 0.5 and 0.75, respectively, to limit computational burden.

To calculate the AUROC for rule-based predictors, we calculated MEWS and qSOFA scores for patients in the Stanford database. MEWS and qSOFA scores were calculated using the entire dataset. We calculated the MEWS score using systolic blood pressure, heart rate, respiratory rate, temperature, and GCS from EHR data. Scores were calculated as described in Fullerton et al.[33] MEWS is often calculated from Alert, Voice, Pain, Unresponsive (AVPU) values; however, the data more reliably contained GCS data. We, therefore, converted GCS to AVPU as follows: 13–15 GCS points as Alert; 9–12 GCS points as Voice; 4–8 GCS points as Pain; $\leqslant$3 GCS points as Unresponsive. This conversion is similar to that described in Kelly et al.,[34] but eliminates the overlap between GCS ranges. Such conversions have been used in previous retrospective studies[35] under similar data availability conditions and do not appear to lower the predictive accuracy of MEWS. The qSOFA score values were calculated from blood pressure, respiratory rate, and GCS values. To compare qSOFA and MEWS to the boosted tree predictor, the boosted tree predictor was trained and tested on Stanford data using all six clinical vital signs and GCS.

Cross-population experiments

To test the performance of the boosted tree predictor when subjects in the training set differ demographically and clinically from those in the test set, we performed cross-population experiments. We trained the boosted ensemble for mortality prediction exclusively on one of the three datasets, using the entire dataset for training, and then tested model performance on the remaining two datasets, without any site-specific retraining. Testing was performed on the entire dataset for each non-training hospital. We performed these experiments using only age and five patient's vital signs. The five vital signs are systolic blood pressure, diastolic blood pressure, heart rate, temperature, and respiratory rate. Other measurements were not included due to limited data availability in the UW Medical Center dataset. For all cross-population experiments, the predictor was trained using XGBoost[32] with the parameters described above. We ran the cross-population validation tests at 12, 24, and 48 h before patient's death.

For comparison with the cross-population experiments, we performed single-population experiments (i.e. one population was used both for training and for testing), using the same five patient's vital signs. For each single-population experiment, we performed fivefold cross-validation and reported the average AUROC, as well as the standard deviation of the AUROCs obtained from each of the five replicates.

Results

The three data sources are all large academic hospitals, but their patient demographics differ (Table 3). UCSF patients were typically older than patients at Stanford and UW. Nearly 26 percent of UCSF patients were over 70 years of age. Approximately 43 percent of UCSF patients were between 50 and 70 years of age, compared with approximately 34 percent of Stanford patients and 38 percent of UW patients. Differences in patient demographics allowed us to test whether the algorithm's predictions are reliable across different patient populations.

GBTs predicted patient mortality more accurately than the LR, SVM, qSOFA, and MEWS for almost all prediction windows on all datasets. When trained and tested on Stanford data, the boosted tree predictor had an AUROC of 0.95 for 48-h mortality (Figure 1(a)). When trained and tested on data collected from UCSF, the boosted tree predictor had an AUROC of 0.96 for 48-h mortality (Figure 1(b)). For training and testing on UW data, the boosted tree predictor had an AUROC of 0.92 for 48-h mortality (Figure 1(c)). Feature importance and feature statistics are listed in Supplementary Tables 1 and 2.

Table 3. Patient demographic information for processed Stanford, University of California, San Francisco (UCSF), and University of Washington (UW) cohorts.

	Characteristic	Stanford (%)	UCSF (%)	UW (%)
Gender	Female	51.78	49.17	38.96
	Male	48.22	50.83	61.04
Age (years)	18–29	18.06	8.20	13.65
	30–39	14.91	9.08	13.45
	40–49	14.96	13.21	14.25
	50–59	17.51	19.83	19.15
	60–69	16.31	23.15	18.57
	>70	16.82	25.92	20.22
In-hospital death	Yes	1.07	3.85	3.03
	No	98.93	96.15	96.97

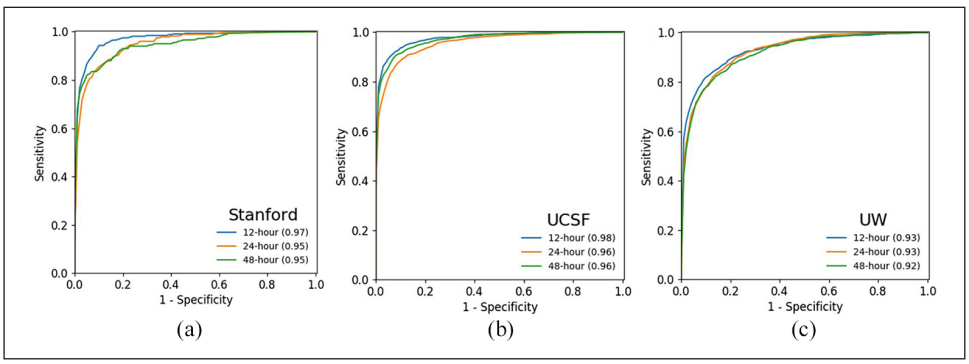

Figure 1. Comparison of receiver operating characteristic (ROC) curves for XGBoost models. ROC curves and AUROC for the boosted tree predictor are presented for 12-, 24-, and 48-h mortality prediction with training and testing performed on (a) Stanford patient data; (b) University of California, San Francisco (UCSF) patient data; and (c) University of Washington (UW) patient data. The Stanford and UCSF predictions used patient's age, systolic blood pressure, diastolic blood pressure, heart rate, temperature, respiratory rate, SpO_2, and Glasgow Coma Scale, whereas the UW predictions used only patient's age, systolic blood pressure, diastolic blood pressure, heart rate, temperature, and respiratory rate.

On the Stanford dataset, LR, SVM, MEWS, and qSOFA had AUROCs of 0.89, 0.78, 0.91, and 0.89 for 48-h mortality prediction, respectively (Table 4). On the UCSF dataset, LR, SVM, MEWS, and qSOFA had AUROCs of 0.80, 0.76, 0.93, and 0.88 for 48-h mortality prediction, respectively. On the UW dataset, LR, SVM, MEWS, and qSOFA had AUROCs of 0.88, 0.83, 0.74, and 0.70 for 48-h mortality prediction, respectively. Across the three datasets, LR, SVM, MEWS, and qSOFA had average AUROCs of 0.85, 0.79, 0.86, and 0.82 for 48-h mortality prediction, respectively. In comparison, the boosted tree predictor had an average AUROC of 0.94 for 48-h prediction.

Detailed performance metrics for the boosted tree predictor, LR predictor, SVM, MEWS, and qSOFA are presented in Table 4. All predictor training and testing was done on the Stanford data set. The diagnostic odds ratio (DOR) is a global measure for comparing diagnostic accuracy between diagnostic tools and is calculated as (True Positive/False Negative)/(False Positive/True Negative). Here, DOR represents the ratio of the odds of a true positive prediction of mortality in

Table 4. Comparison of AUROC, diagnostic odds ratio (DOR), sensitivity, specificity, and positive and negative likelihood ratios (LR+ and LR−) obtained by XGBoost, LR, SVM, qSOFA, and MEWS for mortality prediction using the Stanford dataset.

		XGBoost boosted tree	LR	SVM	MEWS score	qSOFA score
12 h before death	AUROC (SD)	0.972 (0.005)	0.876 (0.009)	0.802 (0.021)	0.949	0.924
	DOR	170.073	16.414	7.471	120.624	86.556
	Sensitivity	0.797	0.795	0.800	0.811	0.960
	Specificity	0.973	0.806	0.634	0.954	0.784
	LR+	35.466	4.162	2.294	17.471	4.436
	LR−	0.209	0.255	0.320	0.198	0.051
24 h before death	AUROC (SD)	0.951 (0.020)	0.880 (0.003)	0.759 (0.015)	0.933	0.899
	DOR	76.614	21.272	3.834	61.006	40.556
	Sensitivity	0.800	0.800	0.792	0.887	0.920
	Specificity	0.935	0.841	0.501	0.886	0.779
	LR+	16.123	5.054	1.590	7.807	4.158
	LR−	0.214	0.238	0.417	0.128	0.103
48 h before death	AUROC (SD)	0.948 (0.013)	0.885 (0.002)	0.778 (0.005)	0.911	0.887
	DOR	137.540	21.056	5.649	41.504	33.585
	Sensitivity	0.800	0.800	0.810	0.838	0.905
	Specificity	0.967	0.840	0.565	0.889	0.780
	LR+	28.308	5.011	1.885	7.558	4.103
	LR−	0.207	0.238	0.340	0.182	0.122

Predictions were performed 12, 24, and 48 h in advance of patient death. All predictor training and testing was done on the Stanford data set using patient measurements for heart rate, respiratory rate, temperature, SpO_2, diastolic blood pressure, and systolic blood pressure and Glasgow Coma Scale (GCS). Each boosted tree predictor value is the average of that value over fivefold cross-validation. Consequently, the metrics have not been calculated directly from one another (e.g. boosted tree predictor DOR does not agree with the ratio of LR+ to LR−). For each boosted tree, LR, and SVM predictor AUROC, the standard deviation (SD) of the five cross-validation AUROCs is also reported. SVM: support vector machine; MEWS: Modified Early Warning Score; AUROC: area under the receiver operating characteristic curve; SD: standard deviation; DOR: diagnostic odds ratio; LR: logistic regression.

patients who died within a given prediction window to the odds of a false positive prediction of mortality in patients who did not die within a given prediction window. For all prediction windows, the boosted tree predictor had a significantly higher DOR than LR, SVM, MEWS, and qSOFA.

In the cross-population experiments, the GBT algorithm trained and tested on various data sets achieved AUROC values up to 0.98 at 12 h before death, 0.96 at 24 h before death, and 0.96 at 48 h preceding death (Table 5). Averaged across pairs of training and test sets from different data sources, the GBT algorithm produced AUROCs of 0.88, 0.86, and 0.80 for 12-, 24-, and 48-h prediction, respectively. Across training sets, average 12-h cross-population performance was best when the predictor was trained on the UW data set.

Discussion

When trained and tested on retrospective data, boosted trees predicted patient mortality more accurately than did LR and SVM comparators and the MEWS and qSOFA risk scoring systems, as evidenced by a battery of metrics at a particular operating point (Table 4) and by ROC curves, summarizing performance across operating points (Figure 1). MEWS is commonly used in studies

Table 5. Average AUROC values for single-population and cross-population experiments using systolic blood pressure, diastolic blood pressure, heart rate, temperature, and respiratory rate.

	UCSF 12 h (SD)	UCSF 24 h (SD)	UCSF 48 h (SD)	Stanford 12 h (SD)	Stanford 24 h (SD)	Stanford 48 h (SD)	UW 12 h (SD)	UW 24 h (SD)	UW 48 h (SD)
Trained on UCSF	0.976 (0.005)	0.962 (0.002)	0.962 (0.003)	0.843	0.802	0.825	0.893	0.879	0.857
Trained on Stanford	0.905	0.873	0.734	0.972 (0.005)	0.951 (0.020)	0.948 (0.013)	0.842	0.859	0.663
Trained on UW	0.926	0.920	0.908	0.880	0.846	0.829	0.933 (0.009)	0.928 (0.019)	0.920 (0.007)

SD: standard deviation; UCSF: University of California, San Francisco Medical Center; UW: University of Washington Medical Center.

For single-population experiments, the standard deviation (SD) of the AUROCs obtained from fivefold cross-validation is reported. Testing was performed 12, 24, and 48 h in advance of patient death.

evaluating early warning systems to provide context for results.[6–9] We chose to compare our algorithm to the MEWS and qSOFA scoring systems because they are well-validated, easily measured scores commonly used to predict all-cause mortality in clinical settings in the United States. While a 24-h lead time has been previously considered clinically relevant for busy hospital settings[6] because it could be argued that this timescale is too short or too long, we also present results for 12- and 48-h lead times.

In this context, the AUROC can be somewhat misleading, as it considers performance at operating points that are not clinically relevant (e.g. sensitivity of 0.99 and specificity of 0.05) and because the low prevalence of in-hospital mortality in the three datasets allows even trivial predictors to obtain respectable accuracy (e.g. a predictor that predicts that every patient will survive). For this reason, the comparison between the boosted tree predictor, MEWS, and qSOFA in Table 4 is perhaps most informative. For 12-h prediction, the differences in DOR and specificity are staggering; a positive-class prediction appears to be strong evidence that the patient's risk of in-hospital mortality is substantially elevated.

Even when trained and tested on data from different hospitals with different patient demographics, the boosted tree predictor maintained high levels of accuracy as demonstrated by AUROC values up to 2 days in advance of patient death (Table 5). This comparison quantifies the reliability of the boosted tree predictor across patient populations (Table 3) and hospitals with different rates of in-hospital mortality. The UCSF and UW data are more similar to each other than to the Stanford data; both have older patients and higher mortality rates than Stanford (3.85% and 2.95%, respectively, vs 1.07%). This is reflected in the cross-population results (Table 5), for which 12-h prediction on the UW data set is better when the boosted tree predictor is trained on the UCSF set than when it is trained on the Stanford set (respective AUROCs of 0.893 and 0.842).

Other machine-learning methods have been developed for in-hospital mortality prediction for specific acute conditions, such as sepsis[19] and cardiac arrest,[21] or for specific settings, such as the ICU;[17] however, relatively little work has been done using machine-learning methods to predict all-cause mortality across all hospital wards. A random forest model developed by Churpek et al.[9] achieved an AUROC of 0.80 on hospital floor patients. An ensemble learning approach by Pirracchio et al.[2] reported an AUROC of 0.88 for in-hospital mortality in the ICU. Taylor et al.[20] describe a random forest model achieving up to 0.86 AUROC in the ED for patients with sepsis. Escobar et al.[36] have reported an in-hospital mortality AUROC up to 0.883 across all hospital wards using LR. However, each of these approaches requires extensive patient data, including

laboratory results, patient histories, and patient demographics. In contrast, the boosted tree predictor described here makes accurate mortality predictions using as few as five routinely collected vital signs.

While MEWS and qSOFA always use the same variables, the boosted tree predictor may use other inputs to improve accuracy. The boosted tree predictor can also be tailored to specific patient populations and does not require manual data entry or calculation, which most comparable prediction tools currently require. This flexibility provides important clinical advantages over MEWS and qSOFA for predicting all-cause mortality.

The boosted tree predictor's combination of high sensitivity and specificity means that it can identify more at-risk patients than LR, SVM, MEWS, and qSOFA while also reducing the number of false alarms. The low specificity of many rule-based systems can lead to alarm fatigue; this desensitization to alarms is a leading patient safety concern in the United States.[37] As demonstrated above, the boosted tree predictor we have developed reduces this risk without sacrificing high sensitivity.

By providing more accurate predictions of mortality risk up to 48 h in advance, the boosted tree predictor could provide clinical teams more time to intervene and potentially improve patient outcomes. Many conditions can be readily treated in their early stages but have high costs and mortality once they progress. For instance, survival rates for patients with septic shock have been shown to decrease by 7.6 percent each hour before antibiotics are administered after onset, and delays in acute kidney injury treatment can lead to total renal failure requiring kidney replacement or to increased patient mortality.[38,39] For conditions without treatment options, early warning of mortality may provide patients, families, and caregivers with the means to reduce unnecessary suffering and to prepare for the possibility that a patient or family member may not survive.[40]

The cross-population results also have promising clinical implications. Machine-learning prediction systems generally must be trained on large amounts of retrospective data from a given hospital, a process that is burdensome for the hospitals and can delay the implementation of life-saving systems. Because the predictor was accurate even when trained and tested on data from different hospitals, it might outperform existing mortality predictors even without site-specific training. This could allow hospitals to adopt the mortality predictor more rapidly and still improve patient outcomes.

There are several limitations to our study. Because the data are retrospective, we cannot draw strong conclusions about performance in a prospective clinical setting. Although there are important demographic differences across the datasets used in this study, all data came from large, urban research universities. Performance on patient populations that differ substantially from these, such as that of a rural community hospital, may differ. Because of the retrospective nature of this work, we do not know how clinicians might adjust their actions based on risk predictions. Nor can we know whether earlier or more aggressive treatment would have prevented or postponed the deaths of those patients who died.

Reported performance metrics are the average performance of five separately trained models. This process was necessary due to the low incidence of mortality in our datasets; dividing the data into a larger number of folds for cross-validation would provide too few examples of patient mortality for accurate training.

Conclusion

The boosted tree predictor predicted patient mortality 48 h in advance of death using only patient's vital signs substantially more accurately than two other machine-learning methods and two commonly used mortality risk stratification tools. In future studies, we intend to test the algorithm

prospectively using real-time clinical data. In a clinical setting, this algorithm may help clinicians identify patients for whom more intensive care would prevent deterioration.

Acknowledgements

The authors thank Jana Hoffman and Emily Huynh for their suggestions and assistance in editing this article. They also thank Thomas Desautels for helpful discussions during this study.

Author contributions

H.M. and R.D. conceived and designed this study; S.P. performed the modeling; H.M., S.P., and P.B.S. performed statistical analysis; all authors contributed to acquisition, analysis, or interpretation of data; H.M., J.C., and A.L-P. drafted the article; all authors revised the article for important intellectual content; and R.D. obtained funding.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: H.M., S.P., J.C., A.L-P., S.L., A.A., A.G-S., E.P., and R.D. are employees of Dascena. C.B. reports receiving consulting fees and grant funding from Dascena. G.F., L.S., D.S., and P.B.S. report receiving grant funding from Dascena. M.F. and U.C. report no conflicts of interest.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Institute of Nursing Research of the National Institutes of Health (Grant No. R43NR015945).

ORCID iD

Jacob Calvert https://orcid.org/0000-0001-7301-8090

Supplemental material

Supplemental material for this article is available online.

References

1. Reini K, Fredrikson M and Oscarsson A. The prognostic value of the Modified Early Warning Score in critically ill patients: a prospective, observational study. *Eur J Anaesthesiol* 2012; 29(3): 152–157.
2. Pirracchio R, Petersen ML, Carone M, et al. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. *Lancet Resp Med* 2015; 3(1): 42–52.
3. Halpern NA, Goldman DA, Tan KS, et al. Trends in critical care beds and use among population groups and Medicare and Medicaid beneficiaries in the United States: 2000-2010. *Crit Care Med* 2016; 44(8): 1490–1499.
4. Beckmann U, Gillies D, Berenholtz S, et al. Incidents relating to the intra-hospital transfer of critically ill patients: an analysis of the reports submitted to the Australian Incident Monitoring Study in Intensive Care. *Intensive Care Med* 2004; 30(8): 1579–1585.
5. Higgins TL, McGee WT, Steingrub JS, et al. Early indicators of prolonged intensive care unit stay: impact of illness severity, physician staffing, and pre-intensive care unit length of stay. *Crit Care Med* 2003; 31(1): 45–51.
6. Prytherch DR, Smith GB, Schmidt PE, et al. ViEWS: towards a National Early Warning Score for detecting adult inpatient deterioration. *Resuscitation* 2010; 81(8): 932–937.
7. Smith GB, Prytherch DR, Meredith P, et al. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* 2013; 84(4): 465–470.

8. Churpek MM, Yuen TC, Winslow C, et al. Multicenter development and validation of a risk stratification tool for ward patients. *Am J Respir Crit Care Med* 2014; 190(6): 649–655.
9. Churpek MM, Yuen TC, Winslow C, et al. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med* 2016; 44(2): 368–374.
10. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system. *Chest* 1991; 100(6): 1619–1636.
11. Subbe CP, Slater A, Menon D, et al. Validation of physiological scoring systems in the accident and emergency department. *Emerg Med J* 2006; 23(11): 841–845.
12. Ferreira FL. Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA* 2001; 286(14): 1754–1758.
13. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA* 2016; 315(8): 801–810.
14. Siontis GCM. Predicting death: an empirical evaluation of predictive tools for mortality. *Arch Intern Med* 2011; 171(19): 1721–1726.
15. Mao Q, Jay M, Hoffman JL, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open* 2018; 8(1): e017833.
16. Lehman L-WH, Adams RP, Mayaud L, et al. A physiological time series dynamics-based approach to patient monitoring and outcome prediction. *IEEE J Biomed Health* 2015; 19(3): 1068–1076.
17. Johnson AE, Pollard TJ, Mark RG, et al. Reproducibility in critical care: a mortality prediction case study. In: *Proceedings of the 2nd machine learning for healthcare conference* (vol. 68), Boston, MA, 18–19 August 2017.
18. Ghose S, Mitra J, Khanna S, et al. An improved patient-specific mortality risk prediction in ICU in a random forest classification framework. *Stud Health Technol Inform* 2015; 214: 56–61.
19. Vieira SM, Mendonça LF, Farinha GJ, et al. Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. *Appl Soft Comput* 2013; 13(8): 3494–3504.
20. Taylor RA, Pare JR, Venkatesh AK, et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med* 2016; 23(3): 269–278.
21. Acharya UR, Fujita H, Sudarshan VK, et al. An integrated index for detection of sudden cardiac death using discrete wavelet transform and nonlinear features. *Knowl-Based Syst* 2015; 83: 149–158.
22. Delahanty RJ, Kaufman D and 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(6): e481–e488.
23. Marafino BJ, Park M, Davies JM, 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.
24. Kramer AA, Higgins TL and Zimmerman JE. Comparison of the mortality probability admission model III, national quality forum, and acute physiology and chronic health evaluation IV hospital mortality models: implications for national benchmarking. *Crit Care Med* 2014; 42(3): 544–553.
25. Le JG, Loirat P, Alperovitch A, et al. A simplified acute physiology score for ICU patients. *Crit Care Med* 1984; 12(11): 975–977.
26. Higgins TL, Teres D, Copes WS, et al. Assessing contemporary intensive care unit outcome: an updated Mortality Probability Admission Model (MPM0-III). *Crit Care Med* 2007; 35(3): 827–835.
27. Fortis S, O'Shea AMJ, Beck BF, et al. An automated computerized critical illness severity scoring system derived from APACHE III: modified APACHE. *J Crit Care* 2018; 48: 237–242.
28. Ma J, Lee DKK, Perkins ME, et al. Using the shapes of clinical data trajectories to predict mortality in ICUs. *Crit Care Expl* 2019; 1: e0010.
29. Weissman GE, Hubbard RA, Ungar LH, et al. Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay. *Crit Care Med* 2018; 46(7): 1125–1132.
30. Salluh JI and Soares M. ICU severity of illness scores: APACHE, SAPS and MPM. *Curr Opin Crit Care* 2014; 20(5): 557–565.

31. Kuzniewicz MW, Vasilevskis EE, Lane R, et al. Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. *Chest* 2008; 133(6): 1319–1327.
32. Chen T and Guestrin C. XGBoost: a scalable tree boosting system. In: *KDD '16 Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, San Francisco, CA, 13–17 August 2016, pp. 785–794. New York: ACM.
33. Fullerton JN, Price CL, Silvey NE, et al. Is the Modified Early Warning Score (MEWS) superior to clinician judgement in detecting critical illness in the pre-hospital environment. *Resuscitation* 2012; 83(5): 557–562.
34. Kelly CA, Upex A and Bateman DN. Comparison of consciousness level assessment in the poisoned patient using the alert/verbal/painful/unresponsive scale and the Glasgow Coma Scale. *Ann Emerg Med* 2004; 44(2): 108–113.
35. Finlay GD, Rothman MJ and Smith RA. Measuring the Modified Early Warning Score and the Rothman Index: advantages of utilizing the electronic medical record in an early warning system—measuring the MEWS and the Rothman Index. *J Hosp Med* 2014; 9(2): 116–119.
36. Escobar GJ, Gardner MN, Greene JD, et al. Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system. *Med Care* 2013; 51: 446–453.
37. Cvach M. Monitor alarm fatigue: an integrative review. *Biomed Instrum Technol* 2012; 46(4): 268–277.
38. Kumar A, Roberts D, Wood KE, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit Care Med* 2006; 34(6): 1589–1596.
39. De Corte W, Dhondt A, Vanholder R, et al. Long-term outcome in ICU patients with acute kidney injury treated with renal replacement therapy: a prospective cohort study. *Crit Care* 2016; 20(1): 256.
40. Searl MF. A case for the use of validated physiological mortality metrics to guide early family intervention in intensive care unit patients. *AACN Adv Crit Care* 2015; 26(1): 13–22.