Adding Continuous Vital Sign Information to Static Clinical Data Improves the Prediction of Length of Stay After Intubation: A Data-Driven Machine Learning Approach

David Castiñeira, Katherine R Schlosser, Alon Geva, Amir R Rahmani, Gaston Fiore, Brian K Walsh, Craig D Smallwood,[†] John H Arnold, and Mauricio Santillana

BACKGROUND: Bedside monitors in the ICU routinely measure and collect patients' physiologic data in real time to continuously assess the health status of patients who are critically ill. With the advent of increased computational power and the ability to store and rapidly process big data sets in recent years, these physiologic data show promise in identifying specific outcomes and/or events during patients' ICU hospitalization. METHODS: We introduced a methodology designed to automatically extract information from continuous-in-time vital sign data collected from bedside monitors to predict if a patient will experience a prolonged stay (length of stay) on mechanical ventilation, defined as >4 d, in a pediatric ICU. RESULTS: Continuous-in-time vital signs information and clinical history data were retrospectively collected for 284 ICU subjects from their first 24 h on mechanical ventilation from a medical-surgical pediatric ICU at Boston Children's Hospital. Multiple machine learning models were trained on multiple subsets of these subjects to predict the likelihood that each of these subjects would experience a long stay. We evaluated the predictive power of our models strictly on unseen hold-out validation sets of subjects. Our methodology achieved model performance of >83% (area under the curve) by using only vital sign information as input, and performances of 90% (area under the curve) by combining vital sign information with subjects' static clinical data readily available in electronic health records. We implemented this approach on 300 independently trained experiments with different choices of training and hold-out validation sets to ensure the consistency and robustness of our results in our study sample. The predictive power of our approach outperformed recent efforts that used deep learning to predict a similar task. CONCLUSIONS: Our proposed workflow may prove useful in the design of scalable approaches for real-time predictive systems in ICU environments, exploiting real-time vital sign information from bedside monitors. (ClinicalTrials.gov registration NCT02184208.) *Key words: mechanical ventilation; pediatrics; machine learning; length of stay; prediction; predictive analytics; intensive care; critical care; biomedical and health data science; data driven machine learning; decision support systems; length of stay estimation; clinical decision making; precision medicine; big data in medicine.* [Respir Care 2020;65(9):1367–1377. © 2020 Daedalus Enterprises]

Introduction

Bedside monitors in ICUs measure patients' physiologic data in real time to continuously assess the health status of patients who are critically ill. These continuous-in-time monitored values often include vital signs (heart rate, breathing frequency, and oxygenation levels), electrocardiogram tracings, and mechanical ventilation parameters

Dr Castiñeira affiliated with Massachusetts Institute of Technology, Cambridge, Massachusetts. Dr Castiñeira, Dr Geva, Mr Fiore, and Dr Santillana are affiliated with Computational Health Informatics Program, Boston Children's Hospital, Boston, Massachusetts. Drs Schlosser, Geva, Walsh, Smallwood, Arnold are affiliated with the Department of Anesthesiology, Critical Care and Pain Medicine, Boston Children's Hospital, Boston, Massachusetts and also with the Department of Anaesthesia, Harvard Medical School, Boston, Massachusetts. Dr

(F_{IO_2}, PEEP, peak inspiratory pressure) during a patient's visit. The continuous monitoring of data is designed to help clinicians intervene in a timely manner if a patient experiences deterioration in his or her health status; however, the massive amount of data displayed in a large ICU taxes human cognition.[1] With the recent advent of increased computational power and the ability to store and rapidly process large data sets, these continuous-in-time data show promise to be used not only as brief snapshots of information routinely absorbed by clinicians during their rounds but also to identify subjects' health trends and predict events and specific outcomes during their ICU stay. This additional information may then be translated into early warning systems that may help improve the care of future patients.[2-6]

Previous efforts to predict health outcomes or the need to change a patient's care strategy have mainly focused on using static or slowly evolving information contained in electronic health records or medical notes.[7,8] These include using early assessments on admission to the emergency department to predict the need of hospitalization[9] and the use of laboratory test information and medication history to predict treatment options. Few studies, however, have focused on examining automated methods capable of extracting meaningful information from multiple continuous-in-time vital sign data sets for event detection or to predict patient outcomes.[10,11] This may be a consequence of the well-known challenges to the development of predictive models that use vital sign data, which include the presence of recording errors, omissions, and outliers in measurements during intensive care.[12]

Our Contribution

We introduced a methodology designed to automatically extract information (features) from continuous-in-time vital sign data collected from bedside monitors to predict patient outcomes via unsupervised and supervised machine learning techniques. We applied this approach in a pilot study aimed at predicting the likelihood that a patient will experience a prolonged stay. Specifically, we showed (1) that continuous-in-time monitor data from the first 24 hours of a patient's ICU stay while on mechanical ventilation have meaningful predictive power to identify stays > 4 d, (2) how model performance was improved when combining

QUICK LOOK

Current knowledge

Previous efforts to predict ICU stay of patients on mechanical ventilation or the need to change a patient's care strategy have mainly focused on using static or slowly evolving information contained in electronic health records or medical notes. With integration of biomedical devices within the ICU, few studies have focused on examining automated methods capable of extracting meaningful information from multiple continuous-in-time vital sign data sets for event detection or to predict patient outcomes.

What this paper contributes to our knowledge

We introduce a methodology designed to automatically identify patterns in subjects' vital sign trends not previously identified in the literature to provide predictions of ICU stay. Based on unsupervised and supervised machine learning approaches, we showed that continuous-in-time monitor data from the first 24 h of a patient's ICU stay on mechanical ventilation had meaningful predictive power to identify prolonged LOS. We further demonstrated that prolonged stay predictions improve when we combine static clinical data, such as patient demographics, the pediatric index of mortality, and clinical history data, with continuous-in-time vital sign information.

Schlosser is affiliated with the Department of Pediatrics, Division of Pediatric Critical Care, Columbia University Irving Medical Center, New York, New York. Dr Santillana is affiliated with the Department of Pediatrics, Harvard Medical School, Boston, Massachusetts. Dr Walsh is affiliated with the Department of Allied Health Professions, School of Health Sciences, Liberty University, Lynchburg, Virginia. Dr Rahmani was affiliated to the Data Science Institute, Columbia University at the time the research was conducted.

Arnold and Santillana jointly supervised this work.

† Deceased.

Supplementary material related to this paper is available at http://www.rcjournal.com.

The data sets generated and analyzed during the current study are not publicly available to protect the privacy of individually identifiable health information, but the de-identified features used for the predictive algorithms are available from the corresponding authors on reasonable request.

The authors have disclosed no conflicts of interest.

The study was performed at Boston Children's Hospital, Boston, Massachusetts.

Dr Geva was funded by NICHD T32 HD040128 and NICHD K12 HD047349.

Correspondence: David Castiñeira PhD, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Building 48 Cambridge, MA 02139. E-mail: davidcastineira@outlook.com; Mauricio Santillana PhD, Computational Health Informatics Program, Boston Children's Hospital, 401 Park Dr, Boston, MA 02215. E-mail: msantill@g.harvard.edu.

DOI: 10.4187/respcare.07561

subjects' static clinical data and continuous-in-time data from vital signs, (3) a parsimonious and scalable machine learning workflow that could be implemented in a real-time setting, and (4) a performance evaluation framework capable of showing that our predictive modeling approach led to robust findings and ensured the validity of our results in the face of a relatively small sample size. The predictive power of our approach outperforms recent efforts by Google (Google, Mountain View, California) (Google's area under the curve of 0.85–0.86 compared with our area under the curve of 0.9, 95% CI [0.80–0.96]) to predict a similar task.[7]

Our approach was novel in the sense that it could identify patterns in the subjects' vital sign trends, not previously identified in the literature, to provide an early signal or indicator of an event or outcome. Based on unsupervised and supervised machine learning approaches, the workflow behind our strategy is scalable because it can produce a prediction with a reduced set of important features from the vital sign information of a given unseen patient, identified a priori, during the training model design.

Motivation and Relevance of Case Study

Patient flow through the pediatric ICU affects resource utilization, surgical scheduling, and staffing. Limited pediatric ICU bed availability and the subsequent inability to admit a patient to the appropriate level of care may result in increased mortality and morbidity.[13] Accurate stay prediction is essential to managing resources effectively and reducing unnecessary wait periods for patients who are vulnerable.

We hypothesized that extracting information from continuous-in-time data routinely collected by bedside monitors would improve the length of stay (LOS) prediction compared with previous efforts that only used fixed-point-in-time information, such as demographic and patients' static clinical data. Continuous-in-time data captures not only a patient's static presentation at ICU admission or another point in time, nor only their worst or best status within a pre-specified time frame but also the trajectory over a time horizon of interest. In particular, the first 24 h of vital signs reflect not only the patient's acuity of illness but also his or her stabilization or failure to respond to treatments. Although a patient's ICU stay clearly depends on the underlying medical condition at the time of admission, correctly identifying such health conditions in a timely fashion may prove challenging in actual clinical practice, especially with 12–24-h shift workers. In addition, a patient's trajectory during his or her ICU stay, as well as organizational, social, and psychological factors, may play important roles in determining the ICU LOS and outcomes of care of a given patient.[14]

Methods

Setting and Subjects

The Medical-Surgical ICU at Boston Children's Hospital is a 30-bed unit that provides care to the full range of pediatric patients, including extracorporeal life support, not admitted primarily due to congenital or acquired pediatric heart disease. The unit has an average of 2,000 annual admissions. For this study, we included all patients on mechanical ventilation for whom vital sign data were available within the first 24 h of mechanical ventilation. Patients were excluded if (a) they were discharged from the ICU $<$ 24 h after the initiation of mechanical ventilation, (b) if they were intubated before admission to the pediatric ICU, (c) if inconsistencies in the recorded ICU admission and discharge times were identified (ie, discharge time preceded admission time), or (d) if there was a gap of $>$20 min in any of the patient's vital signs time series. The length of this time window was chosen with 2 considerations in mind. First, 20 min is a small time interval, which represents $\sim$1.3% of the 24-h period considered as input for this predictive study, and, so, we considered that it would not significantly change the trends (and thus, the extracted features) in the vital signs time series. Second, allowing these time gaps enabled us to build a "realistic and generalizable" workflow capable of dealing with incomplete data (such as those due to routine patient care, eg, bathing) via a suitable data imputation approach. The Boston Children's Hospital Institutional Review Board approved the study with a waiver of informed consent.

Data Sources

All subjects' vital signs used for this study were collected by using routine bedside monitors (Philips IntelliVue MP90, Philips, Andover, Massachusetts) and recorded at a 5-s frequency by using the T3 system (Etiometry, Allston, Massachusetts). We considered only vital signs consistently available for most subjects, that is, heart rate, breathing frequency, and pulse and S_{pO_2}. Note that, although heart rate and pulse both measure the rate at which the heart beats, they are recorded with different devices, and, thus, for robustness in our analysis, we included both signals. Subjects' static clinical data (including sex, age, pre-ICU admission location) and the components of the Pediatric Index of Mortality-2 or Pediatric Index of Mortality-3 score (elective admission, recovery after the procedure, cardiac bypass, diagnosis risk, lack of pupillary response, mechanical ventilation, first systolic blood pressure, base excess, F_{IO_2}, and P_{aO_2}) were collected from the electronic health record for the second phase of the study.

We designed the present study to address the practical problem of forecasting bed availability. With knowing that

intubation is an extremely important factor that affects the stay in the ICU, we selected the total ICU days after intubation as a clinically relevant and practically important outcome for this study. Specifically, the outcome of interest was in determining whether a pediatric patient will experience a prolonged LOS (>4 d) in the ICU from the time of initiation of mechanical ventilation. We chose 4 d because the average duration of mechanical ventilation in our population was 4–5 d and because this time frame allows for modification to the surgical and staffing schedule based on accurate predictions. Discharge times were obtained from the hospital's admission-discharge-transfer system, and ICU LOS was calculated as the difference from the time of intubation, recorded in the electronic health record, to the time of ICU discharge. Deceased patients were excluded from the study because our clinical team believed that most of these individuals who were high risk would have been identified on admission and, thus, an algorithmic approach to identify them would not be of much value.

Data Preprocessing

Time series with no missing values are required for most automatic computational feature extraction techniques, including those used for this study (described in the supplementary materials [see the supplementary materials at http://www.rcjournal.com]). Thus, we used a data imputation approach based on gaussian processes[15-17] to close gaps ≤ 20 min in the vital signs data. When using data imputation techniques based on gaussian processes, the imputed values, at a given point in time, are systematically calculated without any structural assumption on the functional form of the signal, and the imputed values are part of a distribution that allows potential uncertainty quantification to be conducted. As such, they are data driven and easy to implement. We imputed the values of different vital signs in time windows with a width of less than or equal to 20 min, and data imputation for at least one the physiological time series was required for most subjects. Details of this approach are presented in the supplementary materials (see the supplementary materials at http://www.rcjournal.com).

Feature Extraction

Feature-based methods convert raw time series (monitor data) into vectors of statistical features. These features are then used as the input variables in predictive methodologies. In our study, we explored many feature extraction approaches to characterize all 4 vital signs (heart rate, breathing frequency, oxygenation levels, and pulse obtained from oximetry sensors) associated with each pediatric patient. These methodologies extract collections of features (frequently collected in multiple scientific disciplines) from multiple time series, maximizing efficiency and minimizing time,[18-21] and include summaries of the time series, for example, correlation structure, distribution, entropy, stationarity, scaling properties. Projecting vital signs time series into a feature space allowed us to capture the underlying statistical and temporal behavior of the vital signs across multiple subjects and enabled us to mitigate the impact of errors and outliers in the measurements. In addition, because "normal" vital sign values depend on age and, given the heterogeneity of ages in our study cohort, analyzing features may help us identify common properties that may not be apparent from the "raw" vital signs time series.

Overall Machine Learning Methodological Strategy

From a general machine learning perspective, solving the classification problem of identifying patients with prolonged LOS by using vital sign information (and perhaps other relevant patient information) requires 3 steps. The first step is to identify a subset of useful variables, or features, that characterize each individual patient and that can then be used as predictors to build a classifier with the desired outcome. The second step is the model formulation. This step is typically conducted on a subset of subjects (80%), often referred to as a training set. The third step consists of assessing the predictive performance of the "formulated" model on a subset of unseen subjects (20%), that is, the evaluation is performed in a strictly out-of-sample fashion. We refer to these 3 steps sequentially as the following: feature selection, model formulation, and model prediction assessment.

We used 2 different strategies to solve this classification problem. In the first strategy, we experimented with multiple modeling approaches and imposed no constraints on the computational power involved in the solution of the problem. This strategy, denoted as the "baseline approach" in the rest of this article, was used to identify the classification performance of multiple methodologies to solve the task in a research environment, with no constraints on computational time. The second strategy involved the construction of a workflow that could be deployable in a real-life scenario and capable of achieving comparable performance as the best methodologies found in the first strategy but with significantly less computational effort and more parsimony. Both strategies are briefly described below. Detailed information that characterizes these approaches is shown in the supplementary materials (see the supplementary materials at http://www.rcjournal.com).

Baseline Approach

In this strategy, we experimented with feature selection approaches that included, for example, removing low

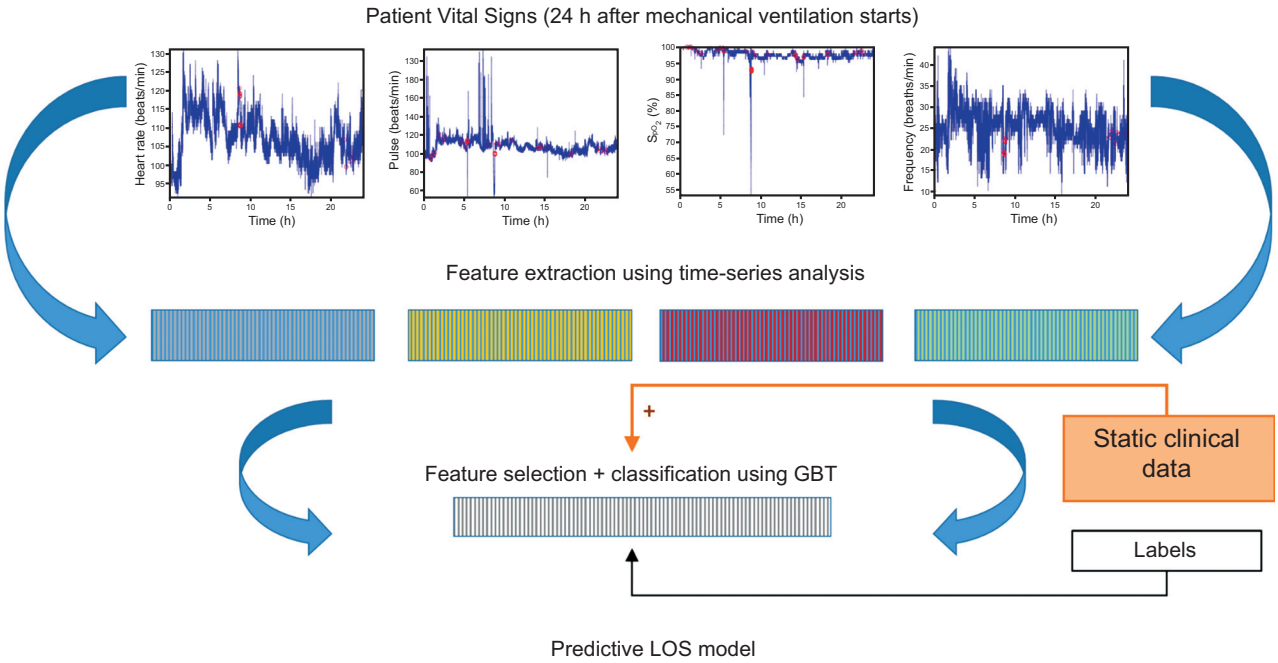

Fig. 1. Proposed methodology for feature engineering. GBT = gradient boosting tree; LOS = length of stay.

variance features, dimensionality reduction by using principal component analysis[22] and removing highly linearly correlated features to remove redundant information. For the model formulation or classification task on the training set, multiple supervised learning approaches were implemented, including logistic regressions, random forests, support vector machine approaches, and gradient boosted trees.[22,23] Furthermore, ensemble methods, such as voting[24] and stacking classifiers,[25,26] were also used to combine individual classifiers to further improve the accuracy of the model.

Scalable Robust Approach

Many of the models studied in the baseline approach use a large number of features as input to complete the classification task. As a result, they lead to accurate results but may not be suitable for practical applications in real-time medical settings because solving the classification problem for a new patient in real-time might demand an overly lengthy time period because the process may require extracting and processing all the time-series features. For reference, the total number of data points from 24 h worth of raw data for 4 different time series collected at 5-s intervals totals 69,120 (ie, 17,280 × 4); whereas all extracted features from the raw data led to a total of 13,332 data points per subject. This likely rendered the baseline methodology impractical for real-time decision making. Thus, a methodology capable of effectively decreasing the number of input variables (dimensionality reduction) and maintaining the predictive performance is desirable.

Furthermore, robustness and interpretability are important in medical applications. For robustness, any practical machine-learning solution should be presented with proper uncertainty quantification envelopes to understand the accuracy and expected variability of any given solution. Avoiding overfitting issues is essential, and uncertainty quantification analysis must assure that this is avoided. For interpretability, at least a general understanding of the families of features that drive those solutions is important. Thus, the proposed deployable workflow relies on an unsupervised learning approach for feature selection that leads to a small and potentially interpretable set of features, and on a scalable and parsimonious (and thus computationally efficient) supervised learning approach to maximize the predictive power via a robust evaluation design.

Feature Selection

We implemented an *unsupervised* learning approach, based on clustering strategies, to identify a reduced number of features capable of characterizing most of the input parameter space. Our approach applies the affinity propagation method[27] by using the maximal information coefficient[28] as a similarity matrix. This methodology outputs representative features that can be interpreted, as opposed to projected features obtained from principal component

analysis like approaches. Details on this approach are presented in the supplementary materials (see the supplementary materials at http://www.rcjournal.com).

Supervised Learning

The supervised learning solution for the problem of predicting LOS is shown in context in Figure 1. Different supervised learning approaches were initially considered to be applicable to our classification problem, such as gradient boosting trees,[29] artificial neural networks, and support vector machine.[23] A gradient boosting tree is typically used with decision trees, which have become a popular machine learning technique because of its simplicity and ease of use. A gradient boosting tree, which combines weak "learners" into a single strong learner in an iterative fashion, provides an extra degree of freedom in the classic bias-variance trade-off. Ultimately, a gradient boosting tree provides a good level of robustness toward overfitting (because boosting builds models intelligently by giving more and more weight to observations that are hard to classify), rendering this solution as appropriate for the problem at hand. To quantitatively evaluate the contribution of each type of data used for prediction, static clinical data versus information contained in the vital signs during the first 24 h of their stay, we designed 3 distinct classification studies:

(a) classification by using vital sign information (time-series features) only,

(b) classification by using static clinical data only, and

(c) classification by using both time series and static clinical data.

Evaluation Approach

To ensure the validity and robustness of our approach, we evaluated the performance of gradient boosting trees under 3 different classification scenarios: (1) classification by using vital signs time-series features only, (2) classification by using readily available electronic health record static clinical data, and (3) classification by using a combination of both vital signs time-series features and static clinical data. For each scenario, that subjects' data were divided into training, validation, and test sets according to the following rule: 80% training, 10% validation, and 10% evaluation. To obtain a proper distribution of the evaluation of model accuracy and to mitigate overfitting due to specific choices of subjects sets, a total of 300 independent experiments, by using random partitions of subjects for training, validation, and evaluation sets, were conducted for each scenario.

Moreover, to reduce the risk of overfitting a problem of 186 predictors and 284 observations/subjects, we used a shrinking factor within the gradient boosting tree algorithm to better control the learning rate and favor more parsimonious models. Note that, similar to a learning rate in stochastic optimization, shrinkage reduces the influence of each individual tree and allows space for future trees to improve the model; ultimately, this provides a configuration trade-off between the number of trees and learning rate. We kept the same proportion of each class as in the original cohort in the training across the different sets to avoid potentially unbalanced distribution of classes and to assess the quality of predictions on both classes in all 300 experiments. In all cases, the training set was used to construct the classification model, the validation set was used to optimize the hyperparameters of the classification algorithm, and the (holdout) evaluation set was used exclusively to quantify the accuracy of the proposed model.

Results

Subjects

A total of 284 subjects were considered for analysis based on the availability of data for the 4 vital signs (heart rate, pulse, S_{pO_2}, and breathing frequency) recorded during the first 24-h window after the onset of mechanical ventilation. A patient inclusion-exclusion flow chart is included in supplementary Figure A9 (see the supplementary materials at http://www.rcjournal.com) to summarize this process. As an illustrative example, vital signs for 3 subjects (referred to here as generic subject nos. 1, 2, and 3) are presented in the bottom panels (C) of Figure 2. Data imputation is shown for subject nos. 1 and 2, whereas, subject no. 3 (as an illustrative example) was not included in our analysis. More detail of the data imputation for another subject in our study is shown in supplementary Figure A1 (see the supplementary materials at http://www.rcjournal.com).

The subjects' age distribution for this study is shown in supplementary Figure A5 (see the supplementary materials at http://www.rcjournal.com). The distribution of LOS across the 284 subjects considered in this work is shown in Figure 2, along with the distribution of pre-ICU location. In total, we had 136 subjects with LOS that were <4 d (48% of all subjects), whereas 148 subjects had LOS that were ≥ 4 d (52% of all subjects). Note that patients with LOS values of <1 d were not represented in this set because we imposed a minimum 24-h window of analysis for each subject.

Feature Extraction

A total of 3,333 features were extracted for each vital sign and each subject, as described in the methodology section, which rendered a total of 13,332 raw features per subject (compared with 69,120 data points contained in the

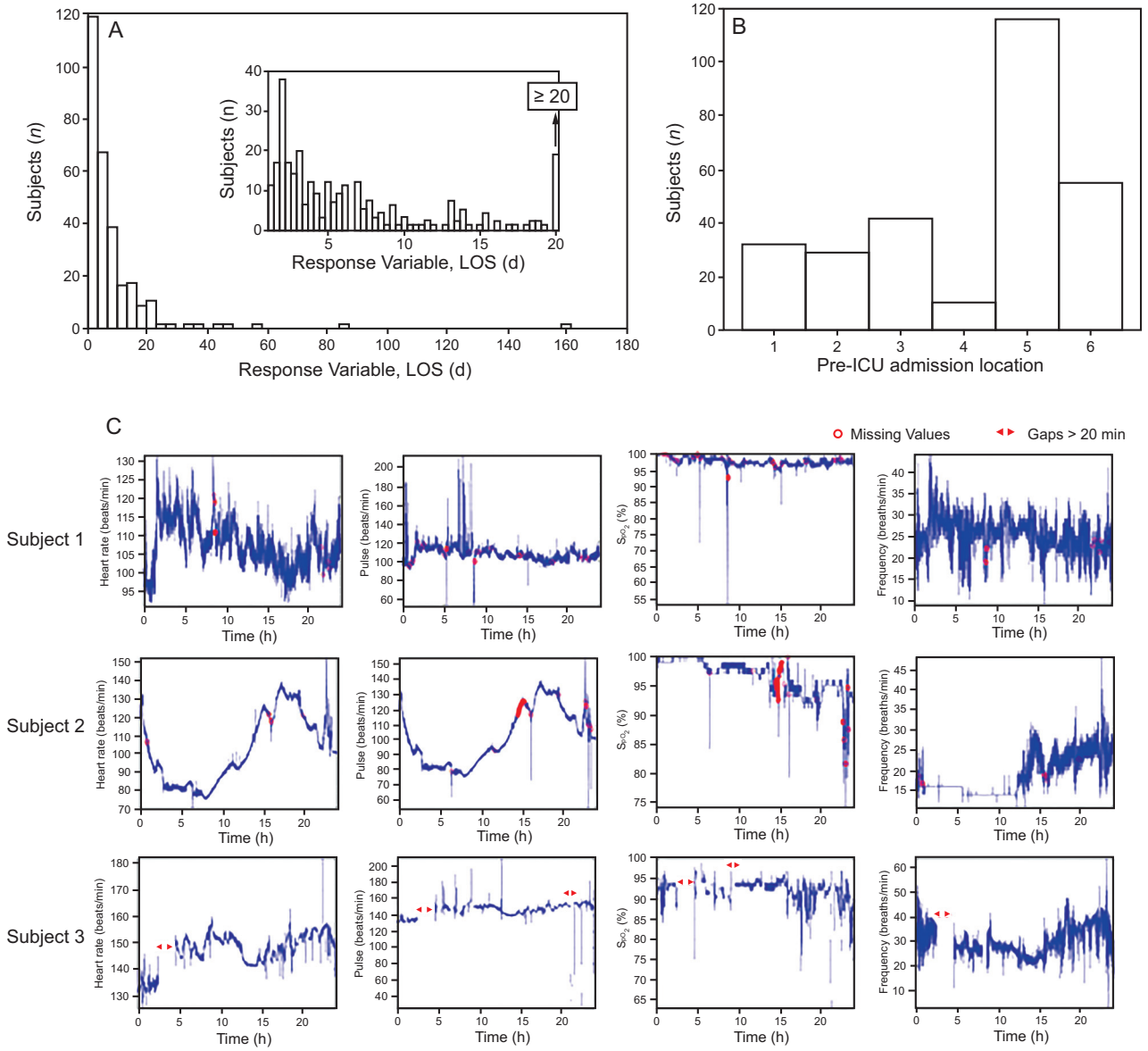

Fig. 2. A: Distribution of length of stay (LOS) across all subjects in this study; the inset in this figure shows a more detailed view of this distribution, with a focus on LOS < 20 d (while grouping all LOS > 20 d into one single bin). B: Distribution of pre-ICU admission location where 1 = in-patient surgical floors, 2 = in-patient medical floors, 3 = emergency room, 4 = other ICUs, 5 = operating room (OR)/procedures, 6 = other/unknown. C: Vital signs.

raw data of the 4 time series). Features that did not yield meaningful values (divisions by zero, for example) were automatically removed.

Baseline Approach

All the methodologies implemented in this classifying approach used only time-series vital signs as input and did not include any static clinical data. A detailed description of the results of the multiple methodologies used in this step is shown in the supplementary materials (see the supplementary materials at http://www.rcjournal.com). In summary, among the classifiers that were studied, the most accurate classifier that used time-series data was a stacking classifier (a hierarchical classifier that combines multiple classification models via a meta-classifier), with an accuracy of 90% on the unseen hold-out set of subjects. The second most accurate classifier (with an accuracy of 89%) was a voting classifier (another kind of meta-classifier for combining different classifiers via majority or plurality voting), which used the same individual classifiers as the stacking classifier. These results helped

us establish a goal for the computationally scalable approach described below.

Scalable Robust Approach

Feature Selection

Given that extracting thousands of features for an unseen patient for which prediction is needed would be too time consuming (this is, in fact, the most time-consuming task in our workflow), we used the affinity propagation unsupervised learning approach to reduce the number of features to be used for prediction from 13,332 to 186. This allowed the agile dynamic training and validation of the supervised training step. The similarity matrix of our study, in which the similarity "within" each cluster is displayed on the diagonal line and the reminder points show similarities "across" clusters, is illustrated in supplementary Figure A3 (see the supplementary materials at http://www.rcjournal.com). In general, as seen in this figure, similarities are larger within clusters than across clusters, which establishes confidence in our clustering results.

Relevance and Interpretability

Of the 186 selected features, 48% of them were extracted either from the heart rate signal (27%) or the pulse signal (21%). This means that 48% of the selected features represent some statistical features of the heart rate. Features extracted from S_{pO_2} and breathing frequency represented 22% and 30%, respectively. We provided a detailed list of these features in the supplementary materials (see the supplementary materials at http://www.rcjournal.com). The features selected from the unsupervised learning (clustering) exercise were then used for the supervised learning study, as discussed in the next section.

Supervised Learning

The accuracies obtained with the gradient boosting tree predictive models as assessed on the evaluation sets exclusively are shown in Figure 4. Note that our results were presented as a distribution of accuracies observed in the 300 independent experiments as opposed to one accuracy measurement obtained from a single model. Our results for the 3 scenarios (static clinical data only, time series only, and combined) show a median model total accuracy (the number of correct predictions/number of observations) of 64, 72, and 80% for each scenario, respectively. These findings indicated that the predictive power of the vital sign information approach (time series) alone could outperform the approach based on electronic health record static clinical data. It is also clear from Figure 3 that the model that combines the vital signs time series and the electronic health record static clinical data provides the best results. This indicates that both data types contain complementary information that lead to better model performance.

The distributions of receiver operating characteristic curves for the 300 experiments, for the 3 data type choices, on the evaluation (or test set) are presented in Figure 3 (bottom). For clarity, for each scenario we provided the 10th, 50th, and 90th percentile receiver operating characteristic curves, meaning the curves associated with the worse model performances, the median model performance, and the best model performance, respectively, and their corresponding area under the curve values are labeled as P10, P50 and P90, respectively. As the values in Figure 3 (bottom) suggest, the median (P50) area under the curve values are 73, 83, and 90% for the model when using static clinical data (green), vital signs (red), and both variable sets (blue), respectively. This is also summarized in supplementary Figure A2 (see the supplementary materials at http://www.rcjournal.com).

For clarity, only the median curves (P50) for each data type choice are displayed in supplementary Figure A2 (see the supplementary materials at http://www.rcjournal.com). This plot summarizes the main findings of this work, which can be stated as follows: continuous-in-time information from vital signs (time series) combined with static clinical data readily available in electronic health record systems via a gradient boosting tree machine learning approach can accurately predict prolonged LOS in subjects on mechanical ventilation. This modeling approach is scalable and robust.

To assess how well our gradient boosting tree approach dealt with overfitting issues, we included in supplementary Figure A8 (see the supplementary materials at http://www.rcjournal.com) a plot of the performance of our methodology via a non-informative normal distribution, with a mean at 50% and SD similar to the scenario that combines both the time series and static clinical data. The observed accuracies on the training set demonstrated the ability of the gradient boosting tree model in fitting the data (possibly signaling overfitting). However, the prediction ability of the resulting model on both validation and evaluation and/or on test sets still showed accurate mean values ($\sim$81% and 79%, respectively), which suggested that the slight potential overfitting on the training set did not affect the quality of predictions in unseen patient data. Moreover, both validation and test sets showed similar distributions for model accuracy, which can be interpreted as a sign of robustness of our predictive models.

Also, the confusion matrix that we computed for the scenario that combines both time-series and static clinical data for the 300 experiments are provided in Figure 4. Our models were better at identifying one of the classes (ie, the average 84.0% accuracy observed for predicting LOS $\geq$ 4 d; average 68.4% accuracy for predicting LOS $<$ 4 d). Although this shows that there is still room for improvements to our methodology, bed availability or other hospital

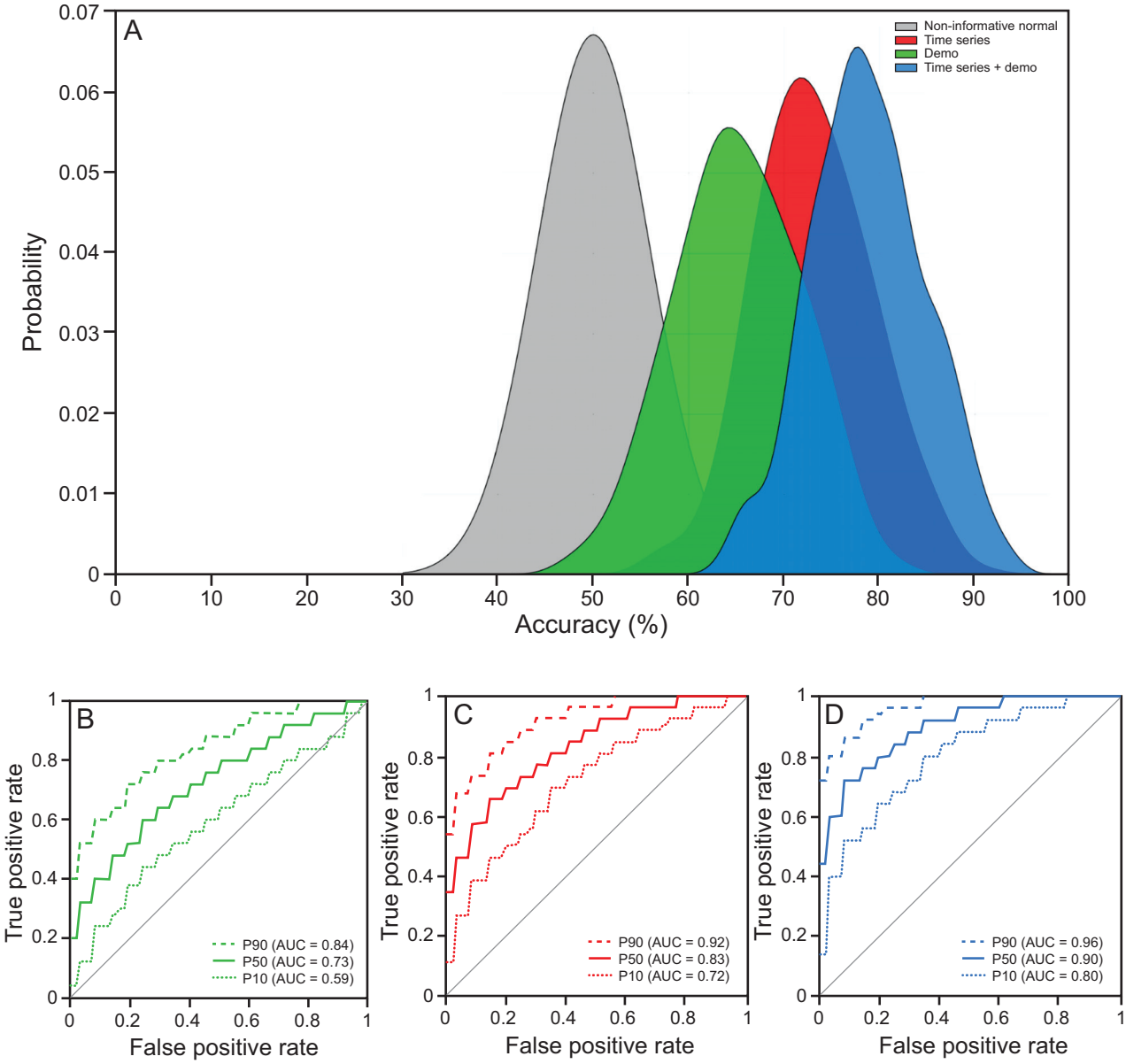

Fig. 3. A: All scenarios with total accuracy. Receiver operating characteristic curves for the 3 data types. B: Static clinical data. C: Time series data. D: Static clinical data plus time series data. The 10th, 50th, and 90th percentile curves are shown (P10, P50, P90).

flow constraints (not associated with a patient's characteristics) may affect LOS < 4 d.

Discussion

We introduced a scalable methodology designed to automatically extract information (features) from continuous-in-time vital sign data collected from bedside monitors to predict patient outcomes via unsupervised and supervised machine learning techniques. Our approach was capable of identifying patterns in subjects' vital sign trends to provide an indicator that predicted whether a patient would experience a prolonged stay (>4 d). Based on both unsupervised and supervised machine learning approaches, the workflow behind our strategy (a) was scalable because it is capable of producing a prediction with a reduced set of important features from the vital sign information of a given unseen patient, identified a priori, during the training model design; (b) achieved prediction accuracies comparable with baseline approaches that may not be practical to be implemented in a real-time medical environment; and (c) was robust, in the sense that favorable performances were observed when applying our approach to different samples of subjects' test, validation, and evaluations sets.

One of the primary reasons for admission to an ICU is to provide close monitoring of vital signs and other

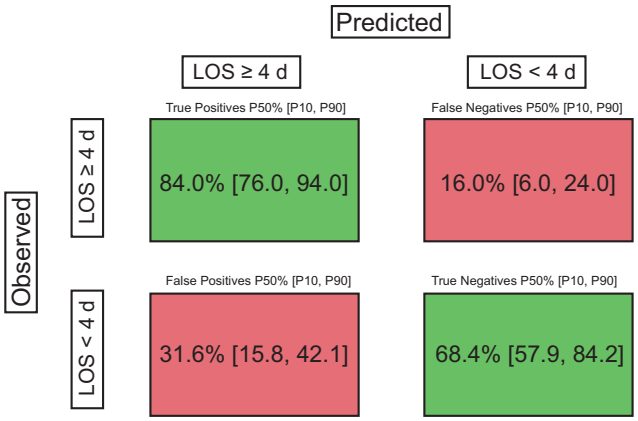

Fig. 4. Accuracies obtained with gradient boosting trees predictive models. The 10th and 90th percentiles are provided (P10, P90). LOS = Length of stay.

physiologic parameters for patients whose clinical course is rapidly changing,[12] yet few examples exist that explore the value of analyzing continuous-in-time information from bedside monitors as a way to improve care strategies. Previous studies used various data types to inform LOS prediction algorithms, including demographic information,[30] derivations of mortality prediction scores,[31] and provider order entry.[32-34] Although such studies typically use data from fixed points in time, analysis of both clinical experience and observational data suggests the importance of clinical change over time in predicting patient outcomes.[32,35]

Our study illustrated the utility of continuous-in-time information from bedside monitors when analyzed through the lens of a machine learning framework and its potential application to resource allocation decision-making. The accuracy of our predictive modeling approach demonstrates how we may design real-time computer-generated decision-support systems. Moreover, our evaluation approach ensured the validity and robustness of our findings. Specifically, and from a clinical standpoint, our modeling approach has an interesting property: it minimizes the number of false negatives, so that patients who are likely to experience prolonged ICU stays after mechanical ventilation are less frequently misclassified. From a resource allocation perspective, this model accurately predicted ICU beds that would not be available in the following half-week, which, for example, may help improve surgical scheduling or respiratory therapy or nurse staffing.

Further research is needed to determine if our model will be able to predict successful stay in a larger and more diverse cohort across hospital facilities. We also believe that higher accuracies could probably be achieved with this methodology by incorporating more human expert engineered features, more static clinical information, and perhaps more information from monitors (eg, ventilator settings) that could be modeled as time series.

Our predictive platform provides the potential to scale up to allow a continuous learning experience from every new patient treated. Moreover, this methodology could easily be extended to predict other outcomes by using continuous-in-time information from bedside monitors. Future steps also include assessing the challenges of integrating complex mathematical modeling into a physiology-driven specialty by studying how the output from these models affects clinician behavior and decisions about resource allocation or treatment options. In turn, treatment decisions made based on information displayed from such predictive modeling may alter patients' clinical course. Thus, evaluation of machine learning-based clinical decision support in a prospective manner is essential.

Limitations

Although the evaluation methodology to ensure the consistency and robustness of our findings was designed carefully and appropriately, the relative size of our study sample and the facts that this pilot study was conducted retrospectively in a single center and only included a study population of subjects on mechanical ventilation suggest that future efforts should aim at understanding how generalizable our approach is in other settings and populations. Our efforts focused on predicting prolonged LOS (>4 d) in a binary way, thus, our results placed in the same category subjects who experienced a 5-d LOS and subjects with LOS > 20 d.

Future studies should focus on identifying extreme outliers as an additional task, which may even include deceased individuals who were excluded from our study. By not including patients for whom long breaks (>20 min) in continuous monitoring occurred in our study, we may have excluded a certain population of patients (perhaps those taken to the operating room for a procedure) with specific needs and characteristics relevant to ICU patient flow. However, data were also not sent from devices in the setting of technical connectivity issues between the monitor and data aggregator, which could occur in any patient population.Nonetheless, this emphasizes the need for further studies that aim at exploring the generalizability of our methodology in larger and perhaps more diverse cohorts.

Finally, our analysis was reliant on continuous-in-time physiologic parameters that must be available for the bulk of a patient's ICU course and, therefore, requires reliable technologic infrastructure. In the present investigation, a high proportion of patients had to be excluded due to gaps in required data. Overtime, as these systems improve in their performance and availability, further investigation is warranted.

Conclusions

We presented a machine learning-based approach capable of extracting meaningful information from continuous-

in-time vital sign information from bedside monitors, from the first 24 h of a subject's ICU stay while on mechanical ventilation, to predict prolonged LOS. Our findings showed that combining subjects' static clinical data and continuous-in-time data from vital signs led to improved predictions. The framework introduced in this work was efficient and scalable, and has the potential to be implemented in real-time settings as a decision-making support tool. Also, we described a comprehensive evaluation framework that ensures the accuracy and robustness of our results. Further studies should consider solving the regression counterpart of our methodology to specifically identify the actual stay (as opposed to the binary classification problem).

REFERENCES

1. Khairat SS, Dukkipati A, Lauria HA, Bice T, Travers D, Carson SS. The impact of visualization dashboards on quality of care and clinician satisfaction: integrative literature review. JMIR Hum Factors 2018;5 (2):e22.
2. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013;309(13):1351-1352.
3. Jameson JL, Longo DL. Precision medicine—personalized, problematic, and promising. N Engl J Med 2015;372(23):2229-2234.
4. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood) 2014;33(7):1163-1170.
5. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 2014;33(7):1123-1131.
6. Parikh RB, Kakad M, Bates DW. Integrating predictive analytics into high-value care: the dawn of precision delivery. JAMA 2016;315 (7):651-652.
7. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1(18).
8. Goldstein BA, Navar AM, Pencina MJ, Ioannidis J. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017;24 (1):198-208.
9. Barak-Corren Y, Israelit SH, Reis BY. Progressive prediction of hospitalisation in the emergency department: uncovering hidden patterns to improve patient flow. Emerg Med J 2017;34(5):308-314.
10. Nagaraj SB, Biswal S, Boyle EJ, Zhou DW, McClain LM, Bajwa EK, et al. Patient-specific classification of ICU sedation levels from heart rate variability. Crit Care Med 2017;45(7):e683-e690.
11. Barquero-Pérez Ó, Figuera C, Goya-Esteban R, Mora-Jiménez I, Gimeno-Blanes FJ, Laguna P, et al. On the influence of heart rate and coupling interval prematurity on heart rate turbulence. IEEE Trans Biomed Eng 2017;64(2):302-309.
12. Maslove DM, Dubin JA, Shrivats A, Lee J. Errors, omissions, and outliers in hourly vital signs measurements in intensive care. Crit Care Med 2016;44(11):e1021-e1030.
13. McManus ML, Long MC, Cooper A, Litvak E. Queuing theory accurately models the need for critical care resources. Anesthesiology 2004;100(5):1271-1276.
14. Gruenberg DA, Shelton W, Rose SL, Rutte AE, Socaris S, McGee G. Factors influencing length of stay in the intensive care unit. Am J Crit Care 2006;15(5):502-509.
15. O'Hagan A. Curve fitting and optimal design for prediction. Journal of the Royal Statistical Society: Series B (Methodological) 1978;40(1): 1-24.
16. Rasmussen CE. Evaluation of gaussian processes and other methods for non-linear regression. Ph.D. thesis, Graduate Department of Computer Science. Toronto: University of Toronto. 1996. Available at: http://mlg.eng.cam.ac.uk/pub/pdf/Ras96b.pdf. *Accessed July 15, 2020*.
17. Rasmussen CE, Williams C. Gaussian processes for machine learning. Adaptive Computation and Machine Learning. Cambridge: MIT Press, 2006. Available at: http://www.gaussianprocess.org/gpml/chapters/RW.pdf. *Accessed July 15, 2020*.
18. Mierswa I, Morik K. Automatic feature extraction for classifying audio data. Mach Learn 2005;58(2-3):127-149.
19. Fulcher BD, Jones NS. Highly comparative feature-based time-series classification. IEEE Trans Knowl Data Eng 2014;26(12):3026-3037.
20. Nun I, Protopapas P, Sim B, Zhu M, Castro DR, Pichara NK. Fats: feature analysis for time series 2015. Available at: https://arxiv.org/abs/1506.00010. *Accessed July 15, 2019*.
21. Christ M, Kempa-Liehr AW, Feindt M. Distributed and parallel time series feature extraction for industrial big data applications 2016. Available at: https://arxiv.org/abs/1610.07717. *Accessed July 15, 2019*.
22. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2009.
23. Bishop CM. Pattern recognition and machine learning. Cambridge: Springer Science+ Business Media; 2006.
24. Kuncheva LI. Combining pattern classifiers: methods and algorithms. New Jersey: John Wiley; 2014.
25. Wolpert DH. Stacked generalization. Neural Networks 1992;5(2):241-259.
26. Rokach L. Ensemble-based classifiers. Artif Intell Rev 2010;33(1-2):1-39.
27. Frey BJ, Dueck D. Clustering by passing messages between data points. Science 2007;315(5814):972-976.
28. Cover TM, Thomas JA. Entropy, relative entropy and mutual information. Elements of Information Theory 1991;2:1-55.
29. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist 2001;29:1189-1232.
30. Marcin JP, Slonim AD, Pollack MM, Ruttimann UE. Long-stay patients in the pediatric intensive care unit. Crit Care Med 2001;29 (3):652-657.
31. 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):1297-1310.
32. Levin SR, Harley ET, Fackler JC, Lehmann CU, Custer JW, France D, Zeger SL. Real-time forecasting of pediatric intensive care unit length of stay using computerized provider orders. Crit Care Med 2012;40 (11):3058-3064.
33. Verburg IWM, 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-e231.
34. Pagowska-Klimek I, Pychynska-Pokorska M, Krajewski W, Moll JJ. Predictors of long intensive care unit stay following cardiac surgery in children. Eur J Cardiothorac Surg 2011;40(1):179-184.
35. Yehya N, Thomas NJ. Disassociating lung mechanics and oxygenation in pediatric acute respiratory distress syndrome. Crit Care Med 2017;45(7):1232-1239.