Computers in Biology and Medicine 118 (2020) 103636

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Early prediction of ICU readmissions using classification algorithms

Melina Loreto [a], Thiago Lisboa [b,c], Viviane P. Moreira [a,*]

[a] *Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil*
[b] *Programa de Pós-Graduação Ciencias Pneumologicas - UFRGS, Hospital de Clínicas de Porto Alegre, Porto Alegre, RS, Brazil*
[c] *Universidade LaSalle, Canoas, RS, Brazil*

ARTICLE INFO

Keywords:
ICU readmission
Machine learning
Classification algorithms

ABSTRACT

Context: Determining which patients are ready for discharge from an Intensive Care Unit (ICU) presents a huge challenge, as ICU readmissions are associated with several negative outcomes such as increased mortality, length of stay, and cost compared to those patients who are not readmitted during their hospital stay. For these reasons, enhancing risk stratification in order to identify patients at high risk of clinical deterioration might benefit and improve the outcomes of critically ill hospitalized patients. Existing work on predicting ICU readmissions relies on information available at the time of discharge, however, in order to be more useful and to prevent complications, predictions need to be made earlier.

Goals: In this work, we investigate the hypothesis that the basal characteristics and information collected at the time of the patient's admission can enable accurate predictions of ICU readmission.

Materials and Methods: We analyzed an anonymized dataset of 11,805 adult patients from three ICUs in a Brazilian university hospital. After excluding 1879 patients who died during their first ICU admission, our final dataset contained 9,926 patients. Of these, 658 patients (6.6%) had been readmitted to the ICU. The original dataset had 185 attributes, including demographics, length of stay prior to ICU admission, comorbidities, severity indexes, interventions, organ support care during ICU stay and laboratory results. The problem of predicting ICU readmissions was modeled as a binary classification task. We tested eight classification algorithms (including Bayesian algorithms, decision trees, rule-based, and ensemble methods) over different sets of attributes and evaluated their results based on six metrics.

Results: Predictions made solely based on the attributes collected at the admission are highly accurate. Their quality in terms of prediction is no different from predictions made using the complete set of attributes for our dataset and for a subset of attributes selected by a feature selection method. Furthermore, our AUROC score of 0.91 (95% CI [0.89,0.92]) is higher than existing results published in the literature for other datasets.

Discussion and Conclusion: The results confirm our hypothesis. Our findings suggest that early markers can be used to anticipate patients at high risk of clinical deterioration after ICU discharge.

1. Introduction

In the intensive care unit (ICU), making predictions is a complex task, since critically ill patients are a heterogeneous population. Patients with the same diagnosis at admission may have distinct acute and chronic comorbidities and may be subject to many different interventions during their ICU stay, leading to disparate outcomes [1].

Premature discharge from the ICU may endanger patients due to a lack of adequate monitoring, which can lead to readmission into the ICU. Determining who is ready for discharge is a daily challenge in any hospital, as ICU readmissions are associated with several negative outcomes such as increased mortality, length of stay (LOS) and cost compared to patients who are not readmitted during their hospital stay [2–4].

Previous studies report that the typical rate of ICU readmission, within a 72-hour window is between 4% and 11% [2,4–7]. Studies performing retrospective reviews of ICU readmissions have found that over 10% of these readmissions were potentially preventable [2,8–10]. Hence, improving risk stratification to identify patients at high risk of clinical deterioration may have a positive impact on the outcomes of critically ill hospitalized patients.

Machine learning algorithms have been applied in several studies to the prediction of clinical outcomes [2,7,11–13]. Although most studies assessing the predictors of ICU discharge focused on attributes

[*] Corresponding author.
E-mail address: viviane@inf.ufrgs.br (V.P. Moreira).

https://doi.org/10.1016/j.compbiomed.2020.103636
Received 1 January 2020; Received in revised form 28 January 2020; Accepted 28 January 2020
Available online 1 February 2020

measured at the time of discharge [2,7,11,14], making predictions based solely on ICU discharge limits our understanding of the impact that the basal characteristics of the patient and interventions during the ICU stay may have on the risk of readmission. Our study therefore specifically focuses on the characteristics of patients at admission (comorbidities, health status, frailty index, organ dysfunction), early interventions in the ICU (need for vasopressors, mechanical ventilation, and other organ dysfunction support), and interventions during the ICU stay. This approach allows us to identify early predictors of failed ICU discharge with the need for readmission.

The hypothesis investigated here is that basal characteristics may impact on and help predict the risk of readmission after ICU discharge before clinical signs alterations at ICU discharge. Thus, this article focuses on the basal characteristics of the patient and the relevant interventions in the ICU and organ support that could impact on the patient's recovery and increase the risk of ICU readmission.

The contributions of this article include:

- the creation of high-quality prediction models for ICU readmission, based solely on attributes collected at the time of admission of the patient;
- a comparison of the performance of eight learning algorithms under six evaluation metrics;
- an analysis of the most useful attributes for predicting ICU readmissions; and
- an evaluation of the feature selection and dimensionality reduction method applied to this task.

Predicting ICU readmissions is not trivial for a number of reasons. Several attributes of critical care at both patient level and organizational level can affect patients' outcomes. This is due to significant differences in the severity of the patient's illness, the case mix in the ICU, and the variations in structure and staff between different units. Readmission after unexpected clinical deterioration is an obvious risk factor for a poor outcome in an individual patient, but it does not necessarily characterize the quality of the ICU's performance [15]. ICU death during the initial admission is an obvious competing risk for subsequent ICU readmission, and the factors that drive mortality (*e.g.,*severity of illness) are similar to those that drive readmission. In addition, most studies assess the risk of readmission at discharge. Our approach to assessing the risk of readmission, which is based on attributes at the time of admission, including organ dysfunction support and interventions during ICU stay, allows us to identify patients at higher risk earlier than previous strategies [2,7,11,14].

Learning algorithms tend to perform poorly in unbalanced datasets as they are likely to be biased towards the dominant class. Our dataset contains a significant imbalance since patients who were readmitted account for less than 7% of the cases. Thus, we employ two techniques to address the class imbalance problem, and measure their effects on the six metrics used to evaluate classification quality.

The remainder of this article is organized as follows: Section 2 reports on previous work on ICU readmissions and prediction algorithms in health care. Section 3 describes the dataset and pre-processing tasks. Section 4 presents our methodology for predicting ICU readmissions and the experimental setup. Section 5 discusses our results. Finally, Section 6 concludes the article.

2. Related work

Hospital readmissions are a matter of concern and have been addressed in a number of studies [13,16–19]. More specifically, some studies assessing the risk of ICU readmission using machine learning have reported interesting findings regarding this high-risk population. One score that is often used to study ICU readmissions is a numerical index called Stability and Workload Index for Transfer (SWIFT) [20]. This metric includes several attributes that can be used to estimate the probability of unplanned ICU readmissions, such as the patient's LOS in the ICU (measured in days), the source of patient admission, the Glasgow Coma Scale (GCS), the evaluation of nursing care for respiratory problems (PCO2), and the ratio between partial pressure of oxygen in arterial blood (PaO2) and the fraction of inspired oxygen (FIO2). These attributes are then scored, taking into account the information available at the time of discharge.

Veloso et al. [14] described a clustering approach for predicting ICU readmissions using SWIFT attributes in combination with lactate and white blood cell counts, and identified two global clusters separating risk of readmission. The authors created 39 scenarios that combined different attributes and used k-means, k-medoids, and x-means algorithms for clustering, obtaining the best results with k-means. Although SWIFT has an acceptable level of accuracy for estimating risk of readmission, it has limited specificity (true negative rate), and relies on information that is only available late in the patient's evolution; it may therefore provide little value beyond clinical judgment. Since most attributes used to calculate SWIFT are collected at the time of discharge, this score cannot be used as a prediction tool, as some patients may be too sick to be discharged, and their likelihood of readmission is obvious at the time of discharge. A prediction tool should be able to identify the population at higher risk of failed discharge.

Fialho et al. [11] developed an approach to predicting ICU readmissions for the MIMIC-II database [21], based on the evaluation of vital signs (*e.g.,* heart rate, temperature, arterial blood pressure) and laboratory results (*e.g.,* platelets, SpO2, lactic acid) available 24 h prior to ICU discharge. This study reported a value of 0.72 for the area under the receiver operating characteristic curve (AUROC), outperforming the APACHE III score.

More recently, Rojas et al. [2] described a machine learning approach to predicting ICU readmission that was significantly more accurate than previously published algorithms in terms of both internal and external validation for the MIMIC-III cohort. They developed a prediction tool using the different types of data available from electronic health records (EHRs): demographic data (such as age, gender, body mass index), vital signs (including the value closest to the time of ICU discharge and trends over the last 24 h prior to discharge), interventions during ICU admission, nursing scores, and diagnostic categories. They decided *a priori* on the use of a gradient boosting machine (GBM) algorithm and found that their machine learning model gave the highest AUROC score for predicting patients that had ever been readmitted (AUROC 0.76), followed by SWIFT (AUROC 0.65), and MEWS (AUROC 0.58) for the interval validation cohort. In a subsequent study [22], the same authors compared the intuition of the clinician versus a machine learning model built using data from the time of ICU discharge. Their automatic method outperformed humans by 0.07 AUROC points, reaching a score of 0.78.

Similarly, Pakbin et al. [7] created a model for predicting the risk of ICU readmission at a variety of time points using data available from EHRs in the MIMIC-III database. They focused on the last 24 h prior to ICU discharge, and used different classification algorithms, including logistic regression, random forest, and gradient-boosted decision trees. The best result was an AUROC of 0.84 with gradient-boosted decision trees.

Finally, Xue et al. [23] built graphs of the temporal patterns of the variables and then applied frequent subgraph mining, grouping similar subgraphs using Non-Negative Matrix Factorization. Experiments were carried out on a more balanced subset of MIMIC-III, in which readmissions occurred in 27% of the cases. The best result was an AUROC score of 0.66.

Table 1 summarizes the main characteristics of the related literature. The main differences between these works and ours are the technique used and the timing of data collection — in our work, we focus on baseline data and data collected during the ICU stay, while existing works rely mostly on data collected at the time of discharge.

Table 1
Comparative analysis of related work on ICU readmissions.

Author	Technique	Features	Period of data collection	Main results	Positive class (%, n)
Rojas et al. [2]	Gradient boosting machine	Demographic data, vital signs, laboratory tests, interventions during ICU admission, nursing scores, and diagnostic categories	Last 24 h prior to ICU discharge	AUROC 0.76	11% (2834)
Pakbin et al. [7]	XGBoost	Data available in an electronic health record, in the MIMIC-III database	Last 24 h prior to ICU discharge	AUROC 0.84; F-score 0.43	6.82% (3637)
Veloso et al. [14]	Clustering, k-means, k-medoids, x-means, DBSCAN	Demographic data, length of stay, laboratory tests	Time of discharge	Davies–Bouldin Index 0.50	3.5% (36)
Fialho et al. [11]	Fuzzy modeling combined with sequential forward selection	Vital signs and laboratory tests	Last 24 h prior to ICU discharge	AUROC 0.72; 68% recall; 73% specificity	13% (135)
Gajicet al. [20]	Logistic regression analysis	Length of stay in the ICU, source of patients admission, Glasgow Coma Scale, the evaluation of nursing care for respiratory problems, and the ratio between partial pressure of oxygen in arterial blood and the fraction of inspired oxygen.	Time of discharge	AUROC 0.75	8.8% (100)
Xue et al. [23]	Frequent subgraph mining on the trend graphs	Demographic data, vital signs, laboratory tests, interventions during ICU admission, and medications.	Entire duration of the ICU stay	AUROC 0.66	27% (310)

3. Data and pre-processing

We analyzed an anonymized dataset containing 11,805 adult patients from three ICUs in a Brazilian university hospital — Irmandade Santa Casa de Misericordia de Porto Alegre, from January 2013 to December 2018. The dataset was created using data available from the Epimed Monitor®, a web-based system used in many ICUs to store patient records and to evaluate performance and the quality of assistance given [24]. In order to evaluate only readmitted patients, we excluded 1879 patients who died during their first ICU admission. Our final dataset contained 9926 patients. Of those, 658 patients (6.6%) were readmitted to the ICU, and these made up our *positive* class. This selection process is depicted in Fig. 1. The original dataset had 185 attributes (or features) including demographics, LOS pre-ICU admission, comorbidities, severity indexes, interventions and organ support care during ICU stay, and laboratory results from ICU admission time (as shown in Appendix). Data on interventions and organ support care gave information on whether a patient had undergone a particular treatment during their ICU stay, since our dataset did not contain information on the duration of treatment. After removing attributes with redundant information or with more than 80% of missing values, we obtained a set of 134 attributes. The final complete dataset had a total of 2.7% missing values.

3.1. Attribute analysis

In order to assess the discriminative power of these attributes in detecting ICU readmissions, we calculated their Information Gain (IG), which measures the expected reduction in entropy (*i.e.,* uncertainty) considering just the feature and the class. The ten most and least discriminative attributes, as ranked by their IG are shown in Table 2. The best features in terms of IG were related to the admission status of the patients (*e.g.,* SAPS3 score, admission type and source, and respiratory failure), some data from chronic diseases (*e.g.,* chronic health status, steroid use, immunosuppression, and solid organ transplant) and LOS both prior to admission to the unit within the ICU. The worst features, according to IG, were mostly related to chronic diseases (*e.g.,* peripheral artery disease, dementia, asthma, complicated diabetes, rheumatic disease, chronic renal failure, and hyperthyroidism). Unlike in existing studies [4], age was not found to be a good predictor of ICU readmission.

Table 2
Best and worst attributes ranked by their Information Gain.

Most discriminative attributes	
0.109	Length of hospital stay prior to unit admission
0.047	Admission source
0.031	Admission type
0.029	SAPS3
0.016	Chronic health status
0.014	Respiratory Failure (first hour)
0.014	Steroids use
0.014	Respiratory Failure
0.012	Immunossuppression
0.012	Transplant solid organ
0,010	Unit length of stay
Least discriminative attributes	
3E−06	Peripheral artery disease
3E−06	Dementia
3E−06	Allogeneic bone marrow transplant
1E−06	Asthma
8E−07	Complicated diabetes
2E−07	Rheumatic disease
5E−08	Chronic renal failure — hemodialysis
2E−09	Hyperthyroidism
0E+00	Age
0E+00	Lowest systolic blood pressure (first hour)

3.2. Attribute selection and dimensionality reduction

Feature selection methods are used to choose the most important attributes without changing them, in order to optimize the performance of algorithms. Reducing the number of features reduces the complexity of the models, making them easier to understand and requiring less time for training.

Unlike feature ranking methods (such as IG, as discussed in Section 3.1), which evaluate each attribute in isolation, wrapper-based methods aim to select the best subset of features. They are able to identify and discard redundant attributes, and to select the combination that produces the best result. Wrapper methods are supervised (*i.e.,* they take the class attribute into consideration) and greedy search algorithms are used to avoid searching exhaustively for the best subset.

Dimensionality reduction methods also reduce the number of attributes in the dataset; however, the difference is that these methods create new attributes based on combinations of the original ones. In this study, we applied Principal Component Analysis (PCA), an unsupervised dimensionality reduction method. PCA initially normalizes all features in order to avoid a situation where attributes with a wide range dominate over those with smaller ranges. It then creates combinations (*i.e.,* principal components) with the maximum possible information to explain data variance. Vectors are created until 95% of the data variance is covered. In our dataset, the best principal component covered only 5% of the data variance; a low covariance was shown between features, and 126 principal components were required to cover 95% of our data variance.

The wrapper method with Naïve Bayes selected the following 15 attributes: age, length of hospital stay prior to unit admission, diabetes, previous myocardial infarction, liver transplant, lung transplant, hypothyroidism, focal neurologic deficit, rhythm disturbances, gastrointestinal bleeding (first hour), pulseless electrical activity (first hour), renal replacement therapy, pulseless electrical activity, acute atrial fibrillation, and discharge shift.

Both the IG and wrapper methods selected the length of hospital stay prior to unit admission as an important feature. In addition, while the IG method selected solid organ transplant, the wrapper method selected liver and lung transplant. The remaining selected attributes were different between these methods. Wrapper selected age and diabetes features, while IG ranked those attributes in the lowest ten.

Despite certain differences in feature selection between the IG and wrapper methods, both selected similar attributes, *i.e.,* of the same type. With the exception of the discharge shift (selected by wrapper but not by IG), and renal replacement therapy or unit LOS (selected only by IG) all other attributes selected were related the previous health status-related of the patient or the severity of acute disease (within the first hour of evolution), showing that these algorithms were able to predict ICU readmission very early during patient evolution.

The IG method discretizes numerical values, and many of our numerical attributes were discretized to a single category with all instances, leading to no IG from those attributes. In contrast, the wrapper method used the Naïve Bayes classifier, which uses Gaussian distributions for numerical attributes. If the Naïve Bayes classifier was changed to discretize numerical attributes, the wrapper method would select only categorical features.

3.3. Class imbalance

Machine learning algorithms typically assume that the number of instances in the different classes is similar, but in many real-life classification problems, the distribution of the instances can be disproportionate. When dealing with imbalanced data, machine learning algorithms are biased towards the majority class. Techniques that can be applied to deal with class imbalance include modification of the dataset (undersampling, oversampling), modification of the algorithm (adapting the algorithm to reduce the bias, through cost-sensitive classifiers), and hybrid methods (a combination of the previous two methods).

In this study, two different techniques were applied to deal with the class imbalance problem:

- **Cost-sensitive classifiers**: A cost-sensitive classifier can be adjusted to penalize false negatives, thus improving the true positive rate. In our experiments, the cost of classifying examples as false negatives was set to ten times higher than the cost of misclassifying false positive examples. This proportion was chosen since it approximates the imbalance ratio in our data.
- **Synthetic Minority Over-sampling Technique (SMOTE)**: This algorithm oversamples the positive class of the training data, thereby reducing the class imbalance. In this work, we were careful to separate the test set before applying oversampling only to the training instances in order to avoid the problem of data leakage (*i.e.,* when oversampled instances are found in both the training and test sets during the split of the validation folds). The algorithm used five nearest neighbors to synthesize the new instance, the percentage parameter for the creation of new instances was set to 1000%, again in an attempt to even out the two classes.

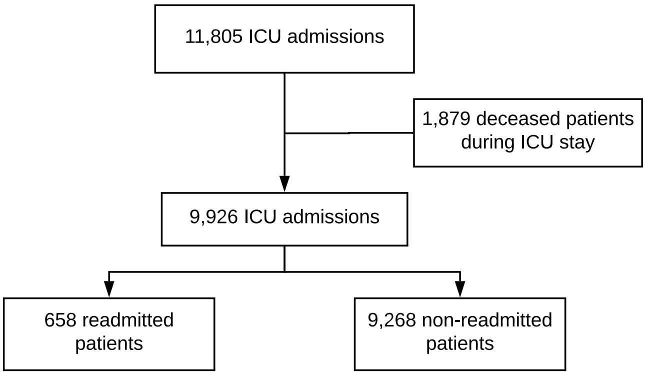
11,805 ICU admissions
1,879 deceased patients during ICU stay
9,926 ICU admissions
658 readmitted patients
9,268 non-readmitted patients

Fig. 1. Patients selected in this study.

4. Predicting ICU readmissions

The aim of this study is to predict ICU readmissions during hospitalization through the use of machine learning algorithms. This section presents the classification algorithms, the evaluation metrics used, and the results of the algorithms. We model the task of predicting ICU readmissions as a binary classification task, in which each instance represents data about a given patient and contains a number of attributes of interest. The classification algorithm builds a model from a set of training instances and then applies it to predict the class of the test instances.

4.1. Data representation

We applied the classification algorithms to several different versions of the dataset, as described below.

- The set of 80 attributes collected at admission (Arrival).
- The complete set containing 134 attributes (Complete).
- The dataset with dimensionality reduction (PCA) containing 126 attributes.
- The dataset with 15 attributes selected from the complete set of attributes using the wrapper method (Wrapper). (See Fig. 1.)

4.2. Classification algorithms

The following classifiers were evaluated in this study.

- **Naïve Bayes** (NB) — This is a classification algorithm that is based on the Bayes' Theorem application and uses the assumption of independence between attributes.
- **J48** — This is an implementation of the C4.5 algorithm in Weka. J48 is a decision tree algorithm that relies on measures of information entropy to split the instances into classes.
- **Random Forest** (RF) — This is a classifier based on a set of Decision Trees. Each tree is constructed with a subset of the data so that it differs as much as possible from one another. Their classifications are made from a majority decision among trained trees.

- **Sequential Minimal Optimization (SMO)** — This is an algorithm that implements John Platt's sequential minimal optimization algorithm for training a support vector classifier. It normalizes all attributes and breaks the problem into the series of smallest possible sub-problems, which are then solved analytically.
- **JRip** — This algorithm uses incremental reduced-error pruning (RIPPER) to find a set of rules that covers all the members of that class.
- **AdaBoost** (AB) — This is an ensemble meta-algorithm developed for binary classification that creates a strong classifier from a weak one. Boosting trains each new model instance to emphasize those training instances were misclassified in previous models. AdaBoost minimizes the exponential loss. In this study, the associated algorithm was Decision Stump, a one-level decision tree.
- **Logit Boost** (LB) — This is an ensemble meta-algorithm that applies a cost function of logistic regression to minimize the logloss. This algorithm was also associated with Decision Stump.
- **Iterative classifier** (ICO) — This is a meta-algorithm that uses cross-validation and optimizes the number of iterations for the given classifier. In our experiments, it was used to optimize the root mean square error of Logit Boost.

4.3. Experimental setup

Our experiments were run using Weka [25], a tool that implements a series of data mining algorithms including feature selection, dimensionality reduction, class balancing, and a series of classifiers. Weka's default parameters were used in all classification algorithms.

Missing values are treated differently by the different classes of algorithms. In the Naïve Bayes classifies, missing values are omitted because they are not taken into consideration in the Bayes formula. In the decision tree classifiers (J48 and RF), missing values are treated by returning the probability distribution of the labels under the attribute branch for which the value is missing. SMO replaces all missing values for nominal and numerical attributes with the modes and means of the training data. In JRip, a rule-based classifier, missing values fail any test included in the rules. In the classifiers that use Decision Stump as the base learner (AB, LB, and ICO), a missing value is treated as a separate value.

Combining the four sets of attributes, eight classification algorithms, and three alternatives for dealing with the class imbalance problem lead to a total of 96 experimental runs. Ten-fold cross-validation was used to mitigate variability in the metrics.

Evaluation Metrics. To evaluate the classifiers, the traditional evaluation metrics were used for those classification algorithms that compare the results of the classifier with the expected classes (ground truth). We argue that it is important to analyze the results using different metrics to allow for a deeper understanding of the results. Since our dataset is highly imbalanced, classifying all instances as belonging to the negative class would yield an accuracy of 93.37%, which is very high. For this reason, we did not use accuracy in our evaluation and opted to rely on more informative metrics, as follows.

- Kappa — This is a statistic that measures the agreement between the automatic classification and the expected classification controlling the effect of the agreement to have occurred by chance. Kappa values greater than zero mean that the classifier is getting a better result than chance.
- Precision — This is the proportion of instances that are truly positive, divided by the total instances classified as positive.
- Recall — This is the percentage of correctly classified instances of the positive class. The recall is equivalent to True Positive Rate (TP Rate).

Table 3
Results for all experimental runs with cost-sensitive learning.

Feature Set	Algorithm	Kappa	Precision	Recall	W-F1	U-F1	AUROC
Arrival	CS-AB	0.34	0.27	0.81	0.41	0.66	0.91
	CS-ICO	0.34	0.27	0.80	0.41	0.66	0.91
	CS-J48	0.37	0.33	0.60	0.42	0.68	0.75
	CS-Jrip	0.33	0.27	0.81	0.40	0.65	0.85
	CS-LB+DS	0.35	0.28	0.80	0.41	0.66	0.91
	CS-NB	0.28	0.23	0.68	0.35	0.63	0.85
	CS-RF	**0.48**	0.58	0.46	**0.51**	**0.74**	0.91
	CS-SMO	0.37	0.31	0.73	0.43	0.68	0.81
Complete	CS-AB	0.33	0.27	0.80	0.40	0.65	0.90
	CS-ICO	0.34	0.27	0.82	0.41	0.66	0.91
	CS-J48	0.36	0.33	0.57	0.42	0.68	0.76
	CS-Jrip	0.34	0.27	0.80	0.40	0.65	0.84
	CS-LB+DS	0.34	0.27	0.79	0.40	0.66	0.91
	CS-NB	0.22	0.19	0.69	0.30	0.59	0.83
	CS-RF	0.46	0.64	0.40	0.49	0.73	**0.92**
	CS-SMO	0.45	0.32	0.73	0.45	0.69	0.81
PCA	CS-AB	0.24	0.21	0.65	0.32	0.61	0.82
	CS-ICO	0.23	0.21	0.60	0.31	0.60	0.81
	CS-J48	0.25	0.28	0.34	0.31	0.63	0.64
	CS-Jrip	0.25	0.22	0.59	0.32	0.61	0.73
	CS-LB+DS	0.25	0.22	0.64	0.33	0.61	0.82
	CS-NB	0.12	0.14	0.42	0.21	0.54	0.72
	CS-RF	0.27	**0.72**	0.19	0.29	0.63	0.87
	CS-SMO	0.45	0.34	0.73	0.46	0.70	0.81
Wrapper	CS-AB	0.22	0.18	**0.94**	0.31	0.56	0.87
	CS-ICO	0.27	0.22	0.83	0.35	0.61	0.88
	CS-J48	0.28	0.24	0.61	0.35	0.63	0.66
	CS-Jrip	0.28	0.22	0.85	0.35	0.62	0.84
	CS-LB+DS	0.25	0.20	0.85	0.33	0.59	0.88
	CS-NB	0.36	0.38	0.46	0.41	0.68	0.83
	CS-RF	0.28	0.30	0.38	0.33	0.64	0.85
	CS-SMO	0.39	0.29	0.61	0.39	0.66	0.75

- F1 — This is a harmonic mean between precision and recall. The reported values are based on the F1 macro, and may be weighted or unweighted. Unweighted F1 (U-F1) is the simple mean of the values obtained for each class, while the weighted version (W-F1) is the average weighted by the number of instances for each class. Since our data are unbalanced, the unweighted measure is preferable because it attaches the same importance to both classes.
- AUROC — This is the area under the ROC curve, which measures the performance of the classifier over all classification thresholds.

5. Results

Our results are listed in Table 3, with the best result for each metric shown in bold. These scores refer to the runs in which cost-sensitive classification was used to mitigate the effects of class imbalance. Next, we analyze the different components of the classification approach.

Sets of Attributes. When evaluating the performance of the different sets of attributes used by the classification algorithms (Table 3), we can see that overall, the Arrival set yielded the best results in terms of kappa, W-F1, and U-F1. The Complete set was slightly better according to AUROC. The best recall was shown in a run where the wrapper method was used to select the best subset of attributes, and the best precision was found for a PCA run. Fig. 2 shows the mean results across all experimental runs in which the given set of attributes was used. Note that these average results differ from Table 3, which shows the scores for a particular experimental run. The Arrival set was consistently the best performer across all metrics, and this is an important finding which confirms our hypothesis. The dimensionality reduction method (PCA) and the feature selection algorithm (wrapper) did not provide significant gains. The only exception was the improvement in recall produced by the wrapper method.

Classifiers. According to Table 3, the top-performing algorithms were RF, AB, and ICO. The RF classifier achieved the best scores for most

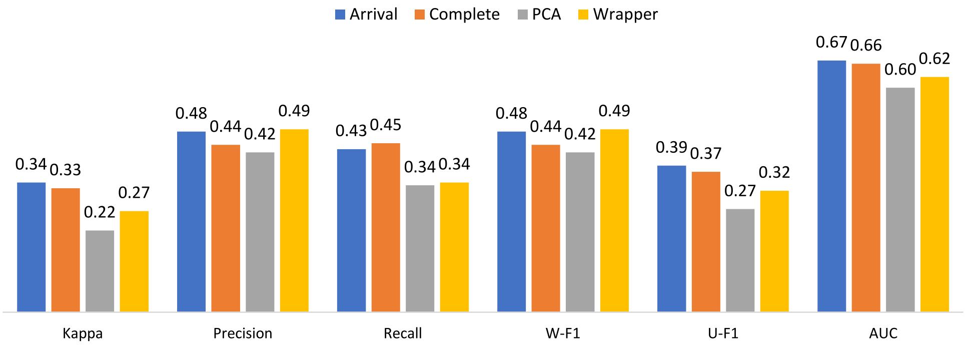

Fig. 2. Performances for the different sets of attributes — mean results across all experimental runs.

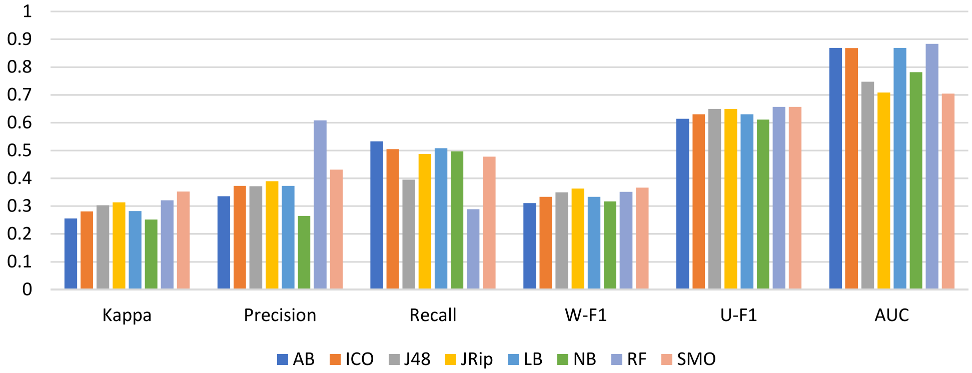

Fig. 3. Performance by classification algorithm — mean results across all experimental runs in which the algorithm was used.

metrics (precision, U-F1, W-F1, and AUROC) and was better at predicting the negative class. For the positive class, AB was the best performer, reaching a recall/TP rate of 0.94 using the attributes selected by the wrapper method. Thus, AB would be the algorithm of choice in cases in which higher sensitivity is desired. Fig. 3 compares the average performances of the classification algorithms under the different evaluation metrics. The bars reflect the average results of all experimental runs in which the specific algorithm was used. It is noticeable that there is no single algorithm that outperforms the others across all metrics. While SMO yields the best kappa by a small difference, RF clearly gives the best precision. On the other hand, RF stands out as the worst performer in terms of recall, followed by J48. For both variations of F1, the results were more homogeneous across classifiers. Regarding AUROC, the best results were achieved by AB, ICO, LB, and RF. In view of the results across all metrics, AB and ICO demonstrated a good balance of scores. We conclude that these two algorithms are the best for the task of predicting ICU readmission.

Effects of approaches for solving class imbalance. In order to assess the gain of those methods that aim to solve the class imbalance problem, we compared them against each other using the Original unbalanced dataset. These results are summarized in Fig. 4. Both SMOTE and the cost-sensitive versions of the classification algorithms give improvements in the classification quality. These improvements were considered statistically significant based on a t-test for kappa, recall, and W-F1 (p-values < 0.01). However, this improvement was at the expense of a significant loss in precision (p-values < 0.001), and the net result was no significant overall gain in AUROC. Cost-sensitive classification was shown to be a better alternative for solving the class imbalance problem, and outperformed SMOTE in all metrics except for precision. In terms of recall, the difference is significantly in favor of cost-sensitive classification.

Table 4 shows the confusion matrices for the ICO classifier using the original and cost-sensitive versions of the algorithm. The matrices demonstrate that cost-sensitive classifiers were able to increase the number of true positives by 386 (58.66%). However, the number of false positives also increased by 1428 (15.41%). We argue that for the purposes of predicting ICU readmissions, false positives are less harmful than false negatives, since being extra vigilant and keeping a patient in the ICU longer is preferable the early discharge of a patient who could be readmitted.

Table 4
Confusion matrices.

ICO + cost-sensitive				ICO No cost-sensitive			
		Predicted				Predicted	
		1	0			1	0
Actual	1	539	119	Actual	1	153	505
	0	1544	7724		0	116	9152

Analysis of misclassified instances. We also compared the true positives against the false negatives to get a deeper understanding of the classification errors. We observed that the algorithm tends to misclassify patients with shorter lengths of hospital stay prior to admission and a shorter unit LOS. Patients in the false negative class had a lower prevalence of steroid use, a lower incidence of respiratory failure, and were usually admitted to the ICU after a surgical procedure, while the true positive class was mainly admitted for medical reasons.

Comparison with other scores and studies. SOFA and SAPS3 scores are widely used to predict ICU mortality. We built classifiers using our best-performing algorithms using each of these scores individually, and compared them against a classifier that uses the set of attributes collected at admission. Fig. 5 shows the ROC curves for the three alternatives. SOFA was the worst performer with an AUROC of 0.54 (95% CI [0.52–0.56]). SAPS3 scored 0.72 (95% CI [0.69–0.74]), while the Arrival set of attributes achieved 0.91 (95% CI [0.89–0.92]). These differences were found to be statistically significant in a t-test with a 99% confidence interval. The poor results yielded by SAPS3 and SOFA were as expected, given that these scores were not designed to predict ICU readmissions. Nevertheless, our work showed that these scores

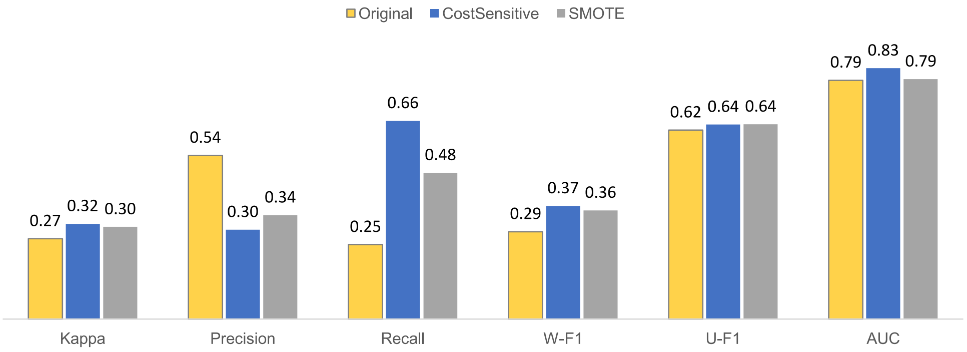

Fig. 4. Performance of the alternatives for solving class imbalance — mean results across all experimental runs.

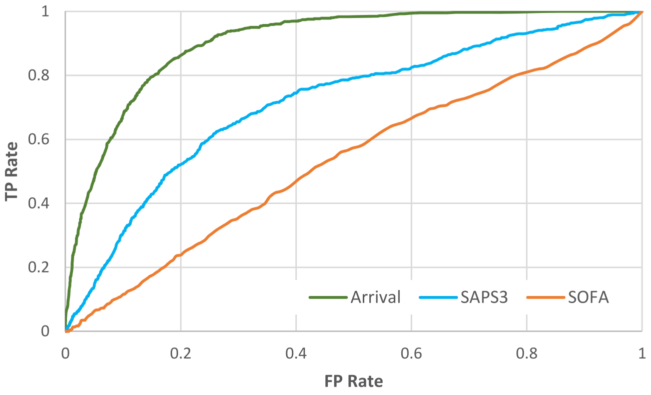

Fig. 5. ROC curves using SAPS3, SOFA, and the attributes collected at arrival.

could be combined with other metrics to provide good classification performance. The SWIFT score, which was designed to assess the risk of readmission, could not be used in our work as it relies only on information available at discharge.

Our results cannot be directly compared to results published in the literature surveyed here (shown in Table 1) since the datasets used in the experiments are not the same. However, it is worth mentioning that our AUROC scores outperform those of existing works [2,7,11,20,23] by as much as 19 percentage points for the same prediction task.

Limitations. Our study has certain limitations. It is an observational study with few ICUs, and thus, many of our findings might be explained by a low variability between practices in these units. All of the data were manually input into the system by health care providers, and there is therefore a (small) risk of error. In addition, we used the same clinical datasets to test the different methods, thereby limiting the generalizability of our findings.

Some of our variables, such as interventions and organ support, were modeled as categorical data, and their time-varying nature was therefore not taken into account. Future work may use time-series data in order to enrich our classification models and improve the results, based on existing evidence [26,27].

Although we achieved better results than previous studies that used patient data at discharge, it was not possible to prove that the admission variables gave better results than the discharge variables in our sample. This makes it more complex to establish comparisons between our findings and those of several other studies that rely on a SWIFT score calculated at ICU discharge. However, as we have shown, our results surpass most of those reported in other studies [2,7,11,20].

In addition, our approach tends to misclassify patients with shorter lengths of hospital stay prior to admission and shorter unit LOS. This suggests that these attributes, which indicate a chronic status for these patients, might have a disproportionate weight in predicting the outcome in such a high-risk group of patients. As a consequence, this could prevent the detection of important effects from other relevant attributes, and may oversimplify potentially explanatory attributes for the outcome of interest.

Another limitation is that no distinction is made in our data as to whether the readmissions were planned. This information could potentially be used to improve the accuracy of classification.

6. Conclusion

The goal of this article was to address the task of predicting ICU readmissions. This is a hard task, as several attributes of critical care, at both patient level and organizational level, can impact on this outcome. Most previous studies that have assessed the risk of ICU readmission using data mining and machine learning have used data collected at ICU discharge, and have found interesting aspects regarding this high risk population, although their performance in terms of predicting failed discharge has been limited.

Our hypothesis was that information collected at the time of admission would suffice to predict patients at high risk of clinical deterioration after discharge from the ICU. We experimented with different sets of attributes, classification algorithms, and approaches for dealing with class imbalance. The results confirmed our hypothesis, as the characteristics of the patient that were recorded at ICU admission were able to predict the risk of readmission outperforming previously published models using data only available at ICU discharge, such as SWIFT.

Further research is needed to determine whether these findings are consistent with other datasets in other settings, and to compare them to the performance of prediction tools using data at ICU discharge as an alternative or an additive approach. Furthermore, external validation of the algorithmic performance would help in an understanding of the differences between studies that have assessed the risk of ICU readmission and the potential strengths and weaknesses associated with our dataset. External validation would also enable us to assess the relative importance of admission characteristics and early events occurring in the ICU in terms of the prediction of ICU readmission in different settings, and to compare these with our findings.

Funding

This work was partially supported by CNPq/Brazil and by CAPES, Brazil Finance Code 001.

Appendix. List of attributes used in the analysis

Unit Code, Age, Gender, ICU Readmission, Hospital Readmission, ICU Discharge Code, Unit Destination Code, Length of Hospital Stay Prior to Unit Admission, UnitLengthStay, Hospital Destination Code, Admission Source Code, Admission Type Code, Charlson Comorbidity Index, MFI score, Saps3, SOFA Score, Chronic Health Status Name, CHF NYHA Class 2–3, CHF NYHA Class4, CRF No Dialysis, CRF Dialysis, Cirrhosis Child A or B, Cirrhosis Child C, Hepatic Failure, Solid Tumor Locoregional, Solid Tumor Metastatic, Anatomic Tumor Site Name, Hematological Malignancy, Immunossupression, Severe COPD, Steroids Use, AIDS, Arterial Hypertension, Asthma, Diabetes Uncomplicated, Diabetes Complicated, Angina, Previous MI, Cardiac Arrhythmia, DeepVenous Thrombosis, PeripheralArtery Disease, Chronic Atrial Fibrilation, Rheumatic Disease, Stroke Sequelae, Stroke No Sequelae, Dementia, Tobacco Consumption, Alcoholism, Psychiatric Disease, Morbid Obesity, Malnourishment, Peptic Disease, Solid Organ Transplant, Autologous BMT, Allogeneic BMT, Cardiac Transplant, Combined Liverkidney Transplant, Combined Pancreaskidney Transplant, Liver Transplant, Intestinal Transplant, Pancreas Transplant, Lung Transplant, Kidney Transplant, Hypothyroidism, Hyperthyroidism, Dyslipidemias, Chemotherapy, Radiation Therapy, History of Pneumonia, Coma, Seizures, Focal Neurologic Deficit, IntracranialMassEffect, Hypovolemic or Hemorrhagic Shock, Septic Shock, Rhythm Disturbances, Anaphylactic, mixed or undefined Shock, DigestiveAcuteAbdomen, Severe Pancreatitis, Liver Failure, Transplant Solid Organ, Trauma Multiple Trauma, Cardiac Surgery, Neurosurgery, Respiratory Failure (1st hour), Mechanical Ventilation (1st hour), NonInvasive Ventilation (1st hour), Vasopressors (1st hour), Cardiac Arrhythmias (1st hour), Cardiopulmonary Arrest (1st hour), Acute Kidney Injury (1st hour), Renal Replacement Therapy (1st hour), Gastrointestinal Bleeding (1st hour), Intracranial Mass Effect (1st hour), Neutropenia (1st hour), Asystole (1st hour), Pulseless Electrical Activity (1st hour), Ventricular Sustained Cardiopulmonary (1st hour), Acute Atrial Fibrilation (1st hour), Atrial Flutter (1st hour), Ventricular Sustained Arrhythmia (1st hour), Respiratory Failure, Mechanical Ventilation, NonInvasive Ventilation, Vasopressors, Cardiac Arrhythmias, Cardiopulmonary Arrest, Acute Kidney Injury, Renal Replacement Therapy, Gastrointestinal Bleeding, Intracranial Mass Effect, Neutropenia, Asystole, Pulseless Electrical Activity, Ventricular Sustained Cardiopulmonary, Acute Atrial Fibrilation, Atrial Flutter, Ventricular Sustained Arrhythmia, Lowest Systolic Blood Pressure (1st hour), Highest Heart Rate (1st hour), Highest Respiratory Rate (1st hour), Lowest Glasgow Coma Scale (1st hour), Lowest Platelets Count (1st hour), Highest Pa O2 (1st hour), Highest Fi O2 (1st hour), Urea, Non Invasive Ventilation, Mechanical Ventilation, Mechanical Ventilation Duration, Tracheotomy, Vasopressors, Renal Replacement Therapy, Central Venous Catheter, Bladder Catheter, Healthinsurance, Discharge Shift

References

[1] J.C. Forte, A. Perner, I.C. van der Horst, The Use of Clustering Algorithms in Critical Care Research to Unravel Patient Heterogeneity, Springer, 2019.

[2] J.C. Rojas, K.A. Carey, D.P. Edelson, L.R. Venable, M.D. Howell, M.M. Churpek, Predicting intensive care unit readmission with machine learning using electronic health record data, Ann. Am. Thorac. Soc. 15 (7) (2018) 846–853.

[3] A.L. Rosenberg, T.P. Hofer, R.A. Hayward, C. Strachan, C.M. Watts, Who bounces back? Physiologic and other predictors of intensive care unit readmission, Crit. Care Med. 29 (3) (2001) 511–518.

[4] C.R. Ponzoni, T.D. Corrêa, R.R. Filho, A. Serpa Neto, M.S. Assunção, A. Pardini, G.P. Schettino, Readmission to the intensive care unit: incidence, risk factors, resource use, and outcomes. A retrospective cohort study, Ann. Am. Thorac. Soc. 14 (8) (2017) 1312–1319.

[5] A.L. Woldhek, S. Rijkenberg, R.J. Bosman, P.H. van der Voort, Readmission of ICU patients: A quality indicator? J. Crit. Care 38 (2017) 328–334.

[6] S.E. Brown, S.J. Ratcliffe, J.M. Kahn, S.D. Halpern, The epidemiology of intensive care unit readmissions in the united states, Am. J. Respir. Crit. Care Med. 185 (9) (2012) 955–964.

[7] A. Pakbin, P. Rafi, N. Hurley, W. Schulz, M.H. Krumholz, J.B. Mortazavi, Prediction of ICU readmissions using data at patient discharge, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, IEEE, 2018, pp. 4932–4935.

[8] T.R. McMillan, R.C. Hyzy, Bringing quality improvement into the intensive care unit, Crit. Care Med. 35 (2) (2007) S59–S65.

[9] A. Garland, Improving the ICU, Chest 127 (6) (2005) 2151–2164.

[10] B.J. Mortazavi, N.S. Downing, E.M. Bucholz, K. Dharmarajan, A. Manhapra, S.-X. Li, S.N. Negahban, H.M. Krumholz, Analysis of machine learning techniques for heart failure readmissions, Circ. Cardiovasc. Qual. Outcomes 9 (6) (2016) 629–640.

[11] A.S. Fialho, F. Cismondi, S.M. Vieira, S.R. Reti, J.M. Sousa, S.N. Finkelstein, Data mining using clinical physiology at discharge to predict ICU readmissions, Expert Syst. Appl. 39 (18) (2012) 13158–13165.

[12] R. Sadeghi, T. Banerjee, W. Romine, Early hospital mortality prediction using vital signals, Smart Health 9 (2018) 265–274.

[13] L. Turgeman, J.H. May, A mixed-ensemble model for hospital readmission, Artif. Intell. Med. 72 (2016) 72–82.

[14] R. Veloso, F. Portela, M.F. Santos, A. Silva, F. Rua, A. Abelha, J. Machado, A clustering approach for predicting readmissions in intensive medicine, Proc. Technol. 16 (2014) 1307–1316.

[15] R. Maharaj, M. Terblanche, S. Vlachos, The utility of ICU readmission as a quality indicator and the effect of selection, Crit. Care Med. 46 (5) (2018) 749–756.

[16] J. Futoma, J. Morris, J. Lucas, A comparison of models for predicting early hospital readmissions, J. Biomed. Inform. 56 (2015) 229–238.

[17] C. Hebert, C. Shivade, R. Foraker, J. Wasserman, C. Roth, H. Mekhjian, S. Lemeshow, P. Embi, Diagnosis-specific readmission risk prediction using electronic health data: a retrospective cohort study, BMC. Med. Inform. Decis. Mak. 14 (1) (2014) 65.

[18] R. Amarasingham, F. Velasco, B. Xie, C. Clark, Y. Ma, S. Zhang, D. Bhat, B. Lucena, M. Huesch, E.A. Halm, Electronic medical record-based multicondition models to predict the risk of 30 day readmission or death among adult medicine patients: validation and comparison to existing models, BMC. Med. Inform. Decis. Mak. 15 (1) (2015) 39.

[19] J. Billings, J. Dixon, T. Mijanovich, D. Wennberg, Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients, BMJ 333 (7563) (2006) 327.

[20] O. Gajic, M. Malinchoc, T.B. Comfere, M.R. Harris, A. Achouiti, M. Yilmaz, M.J. Schultz, R.D. Hubmayr, B. Afessa, J.C. Farmer, The stability and workload index for transfer score predicts unplanned intensive care unit patient readmission: initial development and validation, Crit. Care Med. 36 (3) (2008) 676–682.

[21] M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T.H. Kyaw, B. Moody, R.G. Mark, Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database, Crit. Care Med. 39 (5) (2011) 952.

[22] J.C. Rojas, P.G. Lyons, M. Kilaru, K.A. Carey, L.R. Venable, J. Picart, L. Mccauley, V. Arora, D.P. Edelson, M.M. Churpek, Man vs. machine: Comparison of a machine learning algorithm to clinician intuition for predicting intensive care unit readmission, in: Critical Care: As You Like It-ICU Management and Processes of Care, American Thoracic Society, 2019, p. A2459.

[23] Y. Xue, D. Klabjan, Y. Luo, Predicting ICU readmission using grouped physiological and medication trends, Artif. Intell. Med. 95 (2019) 27–37.

[24] F.G. Zampieri, M. Soares, L.P. Borges, J.I.F. Salluh, O.T. Ranzani, The epimed monitor ICU database®: a cloud-based national registry for adult intensive care unit patients in Brazil, Rev. Bras. de Ter. Intensiv. 29 (4) (2017) 418–426, http://dx.doi.org/10.5935/0103-507x.20170062.

[25] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.

[26] Z. Zhang, J. Reinikainen, K.A. Adeleke, M.E. Pieterse, C.G. Groothuis-Oudshoorn, Time-varying covariates and coefficients in Cox regression models, Ann. Transl. Med. 6 (7) (2018).

[27] T.M. Therneau, P.M. Grambsch, Modeling survival data: Extending the Cox model, Springer-Verlag, New York, 2000, p. 350.