Predictive Modeling of Corticosteroids Sensitivity in Sepsis Using a Supervised Learning Approach

Elisa Lannelongue, Agnès Grimaud, Charles Tillier, Djillali Annane

Predictive Modeling of Corticosteroids Sensitivity in Sepsis Using a Supervised Learning Approach

Elisa Lannelongue[1], Agnès Grimaud[1], Charles Tillier[1], Djillali Annane[2]

(1) Laboratoire de mathématiques de Versailles, Université Paris-Saclay, UVSQ, CNRS, Versailles, France

(2) Réanimation medico-chirurgicale, hôpital Raymond-Poincaré, AP–HP, Garches, France

Abstract Dealing with Sepsis poses a critical challenge in healthcare and necessitates rapid and well-adapted treatment responses. Corticosteroids have been used as a treatment but individual-level effects vary widely. This study aims at improving treatment efficacy by leveraging machine learning techniques to predict patients' sensitivity to corticosteroids. We use two comprehensive datasets of sepsis patients and follow the methodology proposed by Hellali et al. [2024] to evaluate four distinct model configurations. These configurations employ Logistic Regression and Random Forest algorithms, both with and without class balancing using the Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC) data augmentation to address mixed data types. Our findings consistently demonstrate that Random Forest models, particularly when paired with appropriate class balancing techniques, outperform other model configurations in predicting corticosteroid sensitivity using both datasets individually and combined. Notably, incorporating SMOTE-NC significantly enhances model performance, underscoring the importance of appropriately addressing imbalanced datasets in predictive modeling.

Keywords: sepsis, corticosteroids, machine learning, Random Forest, Logistic Regression, SMOTE-NC, class imbalance

1 Introduction

Sepsis has become a global issue (Fleischmann et al. [2016]) and is now one of the leading causes of morbidity and mortality in hospitalized patients. That is in part the result of recent improvements in medical care and access to hospitals where vulnerable patients can be exposed to a variety of pathogens, and are treated for disorders for which no treatments were available until recently, causing considerable mortality, costs and healthcare utilization (O'Brien et al. [2007], Rudd et al. [2020]).

The definition of sepsis has evolved through time, and it is now generally defined as a life-threatening organ dysfunction caused by a deregulated host response to an infection (van der Poll et al. [2021]). The signs and symptoms of sepsis are largely influenced by the virulence of the pathogen, the portal of entry, the susceptibility and response of the host, and the temporal evolution of the condition. Despite recent progress in clinical practices and the pharmaceutical industry, the incidence and mortality rates of sepsis have failed to decrease substantially over the last few decades (Rudd et al. [2020]). Early recognition and adapted intervention are essential to optimize patient outcomes, which is why developing tools to quickly assess what type of treatment is best adapted to each patient has become crucial.

Generally speaking, the Surviving Sepsis international consensus guidelines recommend starting antibiotic treatment within one hour from sepsis onset (Weiss et al. [2020]). However sepsis treatment can be difficult due to disease complexity in clinical context (Weiss et al. [2020]) and heterogeneity of the septic population (van der Poll et al. [2021]).

The present study aims at building and evaluating predictive models for patients' sensitivity to corticosteroids, a class of immuno-regulators that can be used to treat sepsis in cases where antibiotics and blood products have proved inefficient (Keh and Sprung [2004], Annane et al. [2009], Rochwerg et al. [2018]). The RECORDS project aims to investigate the effectiveness of corticosteroids in treating sepsis, as well as determine what biomarkers are involved in a patient's response to treatment. However, individual responses to corticosteroids are highly variable, and the precise factors and biomarkers of responsiveness have not been properly identified yet (Long and Koyfman [2017]).

Many studies have already been conducted using machine learning techniques to predict septic patient outcomes (Pirracchio et al. [2020], Fohner et al. [2019], Komorowski et al. [2022]), provide early diagnosis, or predict sensitivity to corticosteroids, yet the results are often inconclusive due to the heterogeneity of the effect of corticosteroids. Our modeling approach is to build a predictive model to detect corticosteroids sensitivity using the APROCCHSS and RECORDS datasets as in Hellali et al. [2024]. Both datasets contain information gathered from the Assistance Publique – Hôpitaux de Paris (APHP) from the day of admission to the hospital (day 0) to day 90 on patients diagnosed with sepsis to assess the efficacy of corticosteroids treatment and investigate the biomarkers influencing the patients' sensitivity to corticotherapy.

Corticosteroids have been used for years to treat sepsis in addition to hemodynamic and respiratory support and antibiotic administration. However, both the safety and efficacy of corticosteroids remain controversial (Fang et al. [2019]) and various systematic reviews and meta-analyses have either confirmed (Annane et al. [2015]) or refuted (Liang et al. [2021]) any survival benefit.

In 2018, two large studies with random control trials were conducted (Annane et al. [2018], Venkatesh et al. [2018]) and reported comprehensive analyses of the uses of corticosteroids in patients with sepsis yielding opposite results. In the APROCCHSS study, low doses of hydrocortisone and fludrocortisone were shown to reduce the 90-day mortality among patients with septic shock (Annane et al. [2018]), however in the ADRENAL study, a continuous infusion of hydrocortisone in patients undergoing mechanical ventilation did not result in lower mortality compared with patients receiving a placebo (Venkatesh et al. [2018]). Although these studies differed on many levels (gravity of illness, type of corticosteroids administered, etc..), the results remain ambiguous. The uncertainty about the efficacy of corticosteroids to treat sepsis has resulted in a wide variation in clinical practice and current guidelines provide only a weak recommendation for the use of corticosteroids in patients with septic shock when other treatments have failed (Rhodes et al. [2017]).

In recent years, medicine has witnessed the emergence of machine learning as a novel tool to analyze large amounts of data. Many retrospective studies have shown that machine learning models can be used to accurately predict sepsis and septic shock onset with good discrimination in retrospective cohorts (Fleuren et al. [2020]). These models were shown to perform better when including variables that have been known by clinicians to be important to sepsis determination and outperform the usual scoring tools (Moor et al. [2021]). However, the general assessment of machine learning models' performance is still limited due to the vast heterogeneity of studies.

Another type of approach based on data-driven investigations has become increasingly popular.

The aim is to take advantage of the increasing quantity of data available including clinical data and health history for individuals at risk and patients suffering from sepsis (Johnson et al. [2016]) using data mining. Machine learning models can be used to leverage the information available and make accurate predictions about which patient is developing sepsis or which patient is more likely to be sensitive to corticosteroids (Fleuren et al. [2020], Thorsen-Meyer et al. [2020]). For instance, multiple studies have successfully employed a variety of computational models to tackle the challenge of predicting sepsis at the earliest time point possible (McCoy and Das [2017], Barton et al. [2019], Kaji et al. [2019]).

Our main research questions are the following

- How can we improve predictions of patients' responsiveness to corticosteroids based on data collected during the first days of hospitalization?
- Can we improve model generalizability when training and testing on different cohorts?

In order to answer these questions, we used supervised learning techniques. Following previous study Hellali et al. [2024], we used a similar supervised learning approach using APROCCHSS and RECORDS data while refining the statistical analysis and leveraging modeling tools adapted to mixed type data.

The following section describes the data and data collection process as well as the preprocessing phase conducted in Hellali et al. [2024] and presents the system architecture of the data mining pipeline, the models used, and the model selection process. Part 3 presents the results, and Part 4 highlights the interpretations and limits of the results obtained through the application of the different approaches presented above and provides directions for future work. Part 5 concludes this study.

2 Material and Methods

Both datasets include categorical and numerical variables, which means that we are dealing with mixed type datasets. Our approach aims at taking this feature into account.

2.1 Datasets

The APROCCHSS dataset The APROCCHSS cohort data results from a randomized controlled trial study on patients diagnosed with sepsis upon admission to the hospital. The data was collected in the form of electronic case report forms, including all relevant personal and medical information for each patient (demographics, medical history, treatment details, etc...) for 90 days of hospitalization on 1240 patients, with 612 no placebo patients who were administered corticosteroids. For each patient, the database contains 5645 variables, including an indicator of whether they were treated using corticosteroids or not, as well as a label build with expert clinician's help indicating whether the patient is considered sensitive or resistant to corticosteroids in case they received the treatment (see Section 2.2 for more details on the label).

Before preprocessing, the data was reviewed by medical experts, and some patients were reclassified from cortico-resistant to cortico-sensible, and others were considered outliers and thus removed from the study. In the end, a total of 1234 sepsis patients remained in the database. Since our study aims to build a model to predict patients' responsiveness to corticosteroids, the present analysis only includes patients from the APROCCHSS database who received corticosteroids, which is a total of 612 patients. Thus the placebo patients were not used within this study. A description of the APROCCHSS database is given in Table 1.

Group	Cortico-sensitive	Cortico-resistant	Total
Treatment	233	379	612
Placebo	213	409	622
Total	446	788	1234

Table 1: Description of the APROCCHSS database including the distribution of cortico-sensitive and cortico-resistant patients among the participants.

The RECORDS dataset The RECORDS project involves an adaptive clinical trial aiming to assess the ability of biomarkers and algorithms derived from machine learning to predict patients' sensitivity to corticosteroids and thus personalize treatment options in case of sepsis. The RECORDS observational cohort used in this work project corresponds to the data collected during the first phase of the RECORDS trials in the form of case reports on patients diagnosed with sepsis upon admission to the hospital. So far, the RECORDS observational dataset includes a total of 747 patients, with 546 no-placebo patients who received corticosteroids.

The main characteristic of the RECORDS observational cohort is that the study took place in 2020 during the height of the COVID pandemic, which led clinicians to believe that there might be notable differences between the patients in the RECORDS cohort and the pre-pandemic APROCCHSS patients. A description of the RECORDS database is given in Table 2.

Group	Cortico-sensitive	Cortico-resistant	Total
Treatment	235	311	546
Placebo	81	120	201
Total	316	431	747

Table 2: Description of the RECORDS database including the distribution of cortico-sensitive and cortico-resistant patients among the participants.

2.2 Preprocessing

The raw databases used for this study contain a great number of variables that are not necessarily of interest to clinicians, or that can be transformed to facilitate interpretation. The preprocessing phase was divided into four main stages, including feature selection, missing values management, data labelling and class balancing. The first three preprocessing tasks were performed in Hellali et al. [2024].

Feature selection The first stage of preprocessing is to select variables of interest to clinicians. In our case, we used the same variables highlighted by experts as in Hellali et al. [2024]. In addition, since the aim is to predict sensitivity to corticosteroids in the earliest stage of sepsis, we selected features available close to the initial date of hospitalization, that is specifically at day 0, day 1, and a maximum of day 2.

After removing the variables corresponding to the variables after Day 2, as well as the variable of no interest to the experts, the total number of variables is reduced from 5645 to 238 features in APROCCHSS and from 21388 to 84 features in RECORDS. In both cases, the selected features can be categorized as either temporal or non temporal, with each category containing both numerical and categorical variables.

- Non temporal data corresponds to the metadata and more generally to the data about the current status of the patient and personal data (id, sex, weight, age, origin, date of hospitalization, and

treatments before hospitalization, etc...). These characteristics are recorded once at the time of admission (day 0).

- Temporal data refers to the monitoring data recorded over 90 days after the time of admission. These features are generally related to patients' vital signs and laboratory tests and were recorded at different intervals over 90 days of monitoring (site of infection and examination type are recorded before giving treatment, and features such as the SOFA score or treatments doses are recorded along every hospitalization day).

Managing missing values Both databases have a low rate of missing values, and we used the same methods rule-based replacement conventions to handle missing values as Hellali et al. [2024].

- Replacing missing values using rule-based conventions: If the missing value is associated with a dynamic, temporal feature, then the missing value is set to the previously recorded value for that feature. If the missing value is associated with a static feature, then that value is set to -1, to take into account the absence of information in the model.

Data labelling The data labelling is the same as in Hellali et al. [2024]. For every patient in both cohorts, after enrolling in the study on day 0, no placebo patients begin receiving corticosteroid treatment every 4 to 6 hours while their progress is monitored for 90 days maximum. The criteria established by APHP medical experts to determine whether a patient responds to corticotherapy or not is based on four indicators measured on day 14 of monitoring.

Patients are considered to be sensitive to corticosteroids (cortico-sensitive, label = 1) if the following criteria are met after 14 days of treatment :

- The patient did not die.
- The patient did not receive vasopressor treatment over the previous 24-hour period.
- The patient did not require mechanical ventilation over the previous 24-hour period.
- The patient's SOFA (sequential organ failure assessment) score measured on day 14 is less than 6. The SOFA score is used to track a person's status during the stay in an intensive care unit to determine the extent of a person's organ function or rate of failure. The score is based on six different scores, one each for the respiratory, cardiovascular, hepatic, coagulation, renal and neurological systems. The lower the SOFA score, the better is the patient's general state.

If any one of these criteria is not met, the patient is considered to be resistant to corticosteroids (cortico-resistant, label = 0). Patients for whom the label could not be computed were removed from the dataset.

Class balancing In our datasets, the number of patients who are considered sensitive to corticosteroids, the class of interest, is much smaller than the number of patients who are resistant to corticosteroids. This class imbalance can pose a problem since most classification algorithms do not perform as well when the data is skewed toward one class (Hasib et al. [2020]). To address the class imbalance problem and improve performance prediction in the minority class (cortico-sensitive patients), we used the SMOTE-NC (Fernández et al. [2018], Chawla et al. [2002]) method for mixed data types.

Dataset	APROCCHSS	RECORDS
Percentage of Cortico-sensitive patients	38%	43%
Total number of Cortico-sensitive patients	233	235
Total number of patients	612	546

Table 3: Table of the APROCCHSS and RECORDS data class imbalance

2.3 Modelling approach

In the present study, we considered a binary classification problem and the variable of interest corresponds to the patient's sensitivity to corticosteroids, which is encoded with the values $\{0, 1\}$, with 0 = cortico-sensitive and 1= cortico-resistant. We used four distinct types of model specification corresponding to four data configurations. We also tested other specifications, including using only day 0 data, day 0 and day 1 and we choose to present the following configuration because they were performing better.

- **First configuration** using APROCCHSS data from day 0 to day 2, including variables computed using the difference between daily variables, no placebo.
- **Second configuration** using RECORDS data from day 0 to day 2 including variables computed using the difference between daily variables, no placebo.
- **Third configuration** using APROCCHSS and RECORDS data from day 0 to day 2, no placebo, training on APROCCHSS and testing on RECORDS. Since the APROCCHSS dataset contains variables that are not present in the RECORDS dataset, for this configuration the APROCCHSS data is restricted only to the variables that are in the RECORDS dataset.
- **Fourth configuration** using APROCCHSS and RECORDS data from day 0 to day 2, no placebo, training and testing on both APROCCHSS and RECORDS. Both datasets are combined, and the variables missing from the RECORDS dataset are imputed using MissForests (Stekhoven and Bühlmann [2012]) and the values from APROCCHSS. We used the MissForest python package.
  The MissForest imputation method is well adapted to mixed-type data and can handle both continuous and categorical variables simultaneously, thus allowing us to account for possible relations between these variables. For this method, missing values are imputed using Random Forests trained on the observed parts of the dataset. For each variable, the missing values are imputed by fitting a Random Forest with observed values and predictors including the other variables and predicting the missing values by applying the trained Random Forest. The imputation procedure is repeated until the stopping criterion is met. The motivation behind this configuration is to make best use of the whole infomation contained in both APROCCHSS and RECORDS datasets.

Many standard machine learning techniques do not cope well with mixed-type datasets (Gutiérrez-Gómez et al. [2020]). More generally speaking, for each configuration, we compared the performances of several machine learning algorithms, including Logistic Regressions and Random Forests models which were both shown to perform well on these datasets in Hellali et al. [2024] and can effectively cope with mixed type data.

More specifically, for Random Forest models, we used the RandomForestClassifier from the scikit-learn package with the default parameter setting, including an initial choice of 100 estimators, with no specified maximum tree depth, which led to overfitting. For the Logistic Regression models, we

used the LogisticRegression classifier from the scikit-learn package with the default settings and we chose the liblinear solver to ensure the model convergence. The scores used to assess the models' prediction performances were evaluated by cross-validation (50 folds, with a 0.2 test/train ratio) using the cross_val_score function from scikit-learn, and the 95% confidence intervals were computed by bootstrapping.

In both cases, we built a baseline model, unbalanced and unscaled data, with cross-validation, training, and testing on the configuration's data and another model using SMOTE-NC data (scaled and balanced data depending on the type of variable). For each individual model, we used the same three-step approach:

- A default model with no parameter optimization.
- A tuned model with parameter optimization using cross-validation to search the best parameter from the parameter space available. We use the RandomizedSearchCV function of scikit-learn.
- A tuned model that includes feature selection to choose the variables contributing the most to the classification and removing non-contributing variables. For the Random Forest models we used the feature impurity criterion for selection that is implemented in the RandomForestClassifier function of the scikit-learn package, and for the Logistic Regression models, we used model coefficients representing the change in the log odds for a one-unit change in each predictor variable.

Feature Selection The main idea was to simplify the models, add interpretability, and identify problems with the data or modeling approach by calculating a score measuring how often a feature is used in the model and how much it contributes to the overall predictions. For Random Forest models, we used average impurity decrease (Louppe et al. [2013]) to select the features of the training dataset that are most predictive of the target variable. In the case of Logistic Regression models, we used the coefficient of the features in the decision function (Thomas et al. [2008]). However we generally did not observe any improvement in the models' performance in terms of prediction when using feature selection.

Evaluation metrics The first issue linked to class imbalance is that our model's predictive performances are not the same for both classes since the costs of different classification errors may vary as a function of the ratio between classes of interest. In the present study, we used three measures of predictive accuracy that we aim to maximize.

- **Accuracy** measures the number of correctly classified data instances over the total number of data instances. In the case of unbalanced data, the error rates derived from measured predictive accuracy are different whether we consider the predictive accuracy for the cortico-sensitive class or the cortico-resistant class. The accuracy might be high because the classifier learns to classify samples from the majority class accurately and fail to properly identify elements of the minority class without incurring a heavy cost to the accuracy metric.
- **Recall** : This measure shows the rate of misclassified items from the minority class. The recall is low in the case of unbalanced data because the classifier systematically fails to identify elements from the minority class, which in our case corresponds to cortico-sensitive patients.
- **AUC** : The AUC is the area under the ROC curve, which corresponds to the TP (True Positive) rate plotted against the FP (False Positive) rate. It is a standard metric to evaluate the performance of machine learning algorithms in classification problems at various threshold settings indicating how much the model is capable of distinguishing between classes.

3 Results

This section presents the modelling results for each configuration described above using the AUC, accuracy and recall as performance metrics. For each Random Forest model and Logistic Regression model, we present the baseline (no balancing operation) results with hyper parameter tuning and the results with class rebalancing using SMOTE-NC and hyper parameter tuning. We chose to only present the results for the models with hyper parameter tuning since we did not observe any difference between each step of the three step approach (for every model considered, the confidence intervals for all metrics were overlapping), but the tuned models were both not subject to overfitting and were generally more efficient computationally.

Configuration 1 To compare the baseline models, we first trained models using the APROCCHSS data from day 0 to day 2 with minimal transformations, including no balancing operation and then trained models using SMOTE-NC augmented data.

	Baseline (no balancing operation)			SMOTE-NC		
	AUC	Accuracy	Recall	AUC	Accuracy	Recall
Average value	0.804	0.727	**0.407**	0.840	0.746	**0.836**
Bootstrap CI 95%	[0.800, 0.807]	[0.724, 0.730]	**[0.400, 0.412]**	[0.836, 0.844]	[0.742, 0.750]	**[0.831, 0.841]**

Table 4: Performance chart of the Random Forest models for the first configuration specifications and hyperparameter tuning.

	Baseline (no balancing operation)			SMOTE-NC		
	AUC	Accuracy	Recall	AUC	Accuracy	Recall
Average value	0.777	0.718	**0.593**	0.806	0.739	**0.773**
Bootstrap CI 95%	[0.774, 0.780]	[0.714, 0.720]	**[0.586, 0.599]**	[0.801, 0.810]	[0.734, 0.743]	**[0.767, 0.779]**

Table 5: Performance chart of the Logistic Regression models for the first configuration specifications and hyperparameter tuning.

In all configuration 1 baseline models, the recall metric is always less than 0.6, which is coherent with the class imbalance problem that we expected to observe in this situation: the model is good at identifying the majority class that comprises most of the data, which is why the accuracy is high, and systematically misclassify elements from the minority class. The same observations go for the baseline Logistic Regression models as well, although we can see that both the accuracy and AUC values are better for Random Forest models with non-overlapping confidence intervals. The recall metric however is systematically higher for baseline Logistic Regression models, which is coherent with the idea that Logistic Regression models are less sensitive than Random Forest to the class imbalance problem (Guo et al. [2008]).

On the other hand, when applying the SMOTE-NC method to the APROCCHSS data, we can see a significant improvement in all metrics, and the recall becomes higher for the Random Forest models (SMOTE-NC recall for the hyper-tuning Random Forest model is in the intervals $[0.831, 0.841]$ whereas SMOTE-NC recall for the hyper-tuning Logistic Regression model is in the interval $[0.767, 0.779]$).

Configuration 2 To go further, and to have a first understanding of the RECORDS dataset, we adopted the same tools and modeling approach as in configuration 1 using RECORDS data.

	Baseline (no balancing operation)			SMOTE-NC		
	AUC	Accuracy	Recall	AUC	Accuracy	Recall
Average value	0.787	0.707	**0.572**	0.808	0.735	**0.801**
Bootstrap CI 95%	[0.782, 0.792]	[0.702, 0.712]	**[0.562, 0.581]**	[0.803, 0.811]	[0.730, 0.738]	**[0.794, 0.807]**

Table 6: Performance chart of the Random Forest models for the second configuration specifications and hyperparameter tuning.

	Baseline (no balancing operation)			SMOTE-NC		
	AUC	Accuracy	Recall	AUC	Accuracy	Recall
Average value	0.731	0.684	**0.607**	0.761	0.705	**0.761**
Bootstrap CI 95%	[0.725, 0.736]	[0.679, 0.689]	**[0.597, 0.615]**	[0.756, 0.765]	[0.701, 0.709]	**[0.754, 0.767]**

Table 7: Performance chart of the Logistic Regression models for the second configuration specification and hyperparameter tuning.

For both baseline models, the recall metrics are slightly better than they were for the baseline models on APROCCHSS data with recall in $[0.562, 0.581]$ for the Random Forest baseline and recall in $[0.597, 0.615]$ for the baseline Logistic Regression. However, in the present study, we consider a recall less than 0.7 is not satisfying for the detection of patients sensitive to cortico-steroids.

Comparing the baseline version of the Random Forest model on RECORDS data from day 0 to day 2 with the SMOTE-NC version, we observe a significant improvement of the recall measure in higher, non-overlapping confidence intervals (baseline recall is in $[0.562, 0.581]$, SMOTE-NC recall is in $[0.794, 0.807]$). The same observation is also valid for the baseline Logistic Regression model with recall in $[0.597, 0.615]$ and for the SMOTE-NC Logistic Regression model with recall in $[0.754, 0.767]$.And we can further observe that the Random Forest models systematically perform better than the Logistic Regression model for all metrics.

Configuration 3 In this configuration, models are trained on APROCCHSS data and tested on RECORDS data. Since the APROCCHSS preprocessed dataset contains more variables than the preprocessed RECORDS dataset, we reduced the APROCCHSS dataset to variables comparable to the ones contained in RECORDS.

	Baseline (no balancing operation)			SMOTE-NC		
	AUC	Accuracy	Recall	AUC	Accuracy	Recall
Average value	0.788	0.712	**0.582**	0.814	0.736	**0.803**
Bootstrap CI 95%	[0.783, 0.793]	[0.706, 0.716]	**[0.572, 0.591]**	[0.809, 0.817]	[0.731, 0.740]	**[0.796, 0.809]**

Table 8: Performance chart of the Random Forest models for the third configuration specifications and hyperparameter tuning.

	Baseline (no balancing operation)			SMOTE-NC		
	AUC	Accuracy	Recall	AUC	Accuracy	Recall
Average value	0.727	0.679	**0.592**	0.753	0.693	**0.730**
Bootstrap CI 95%	[0.721, 0.732]	[0.673, 0.683]	**[0.582, 0.600]**	[0.747, 0.757]	[0.688, 0.697]	**[0.722, 0.736]**

Table 9: Performance chart of the Logistic Regression models for the third configuration specifications and hyperparameter tuning.

In this case, the baseline model performs moderately well based on the AUC and accuracy metrics which are respectively in $[0.783, 0.793]$ and $[0.706, 0.716]$ for the Random Forests and $[0.721, 0.732]$ and $[0.673, 0.683]$ for the Logistic Regression. However, the recall metric is under 0.6 in both cases, which shows that these models are not well-equipped to detect items from the minority class.

When using the SMOTE-NC re-balancing method, all performance metrics for both models are significantly improved, with higher, non overlapping confidence intervals. More importantly, the recall metric for the Random Forest model goes from $[0.572, 0.591]$ for the baseline version to $[0.796, 0.809]$ for the SMOTE-NC version and the recall metric for the Logistic Regression model goes from $[0.582, 0.600]$ for the baseline version to $[0.722, 0.736]$ for the SMOTE-NC version. Furthermore, on all performance metrics considered, the Random Forest model with SMOTE-NC re-balancing is the best performing model on that configuration, which is in alignment with previous observations based on the results from the first two configurations. This satisfactory result tends to show that both datasets can be combined, to build more general classifications models.

Configuration 4 In this configuration, both datasets are combined, and the values missing from the RECORDS dataset are imputed using the MissForests imputation method on the APROCCHSS data from day 0 to day 2 data.

	Baseline (no balancing operation)			SMOTE-NC		
	AUC	Accuracy	Recall	AUC	Accuracy	Recall
Average value	0.858	0.814	**0.642**	0.899	0.832	**0.819**
Bootstrap CI 95%	[0.822, 0.891]	[0.776, 0.848]	**[0.563, 0.716]**	[0.871, 0.924]	[0.797, 0.864]	**[0.797, 0.864]**

Table 10: Performance chart of the Random Forest models for the fourth configuration specifications and hyperparameter tuning.

	Baseline (no balancing operation)			SMOTE-NC		
	AUC	Accuracy	Recall	AUC	Accuracy	Recall
Average value	0.722	0.651	**0.552**	0.762	0.690	**0.751**
Bootstrap CI 95%	[0.694, 0.750]	[0.627, 0.674]	**[0.502, 0.599]**	[0.730, 0.793]	[0.663, 0.716]	**[0.691, 0.808]**

Table 11: Performance chart of the Logistic Regression models for the fourth configuration specifications and hyperparameter tuning.

For all baseline models, the recall measures are below the 0.7 threshold, and the SMOTE-NC operation is required to improve the recall. The observed improvement is however systematically higher in the case of Random Forest models.

The Random Forest model with SMOTE-NC rebalancing in configuration 4 is the overall best performing model in this study with an AUC in $[0.871, 0.924]$ and a recall in $[0.797, 0.864]$. In this case, the baseline models were also moderately well performing based on the AUC and accuracy metrics which are respectively in $[0.822, 0.891]$ and $[0.776, 0.848]$ for the Random Forests and $[0.694, 0.750]$ and $[0.627, 0.674]$ for the Logistic Regression. However, the recall metric are around 0.6 in both cases, which shows that these models are not well-equipped to detect items from the minority class.

When using the SMOTE-NC re-balancing method, all performance metrics for both models are significantly improved, with higher, non overlapping confidence intervals. More importantly, the recall metric for the Random Forest model goes from $[0.563, 0.716]$ for the baseline version to $[0.797, 0.864]$ for the SMOTE-NC version and the recall metric for the Logistic Regression model goes from $[0.502, 0.599]$ for the baseline version to $[0.691, 0.808]$ for the SMOTE-NC version.

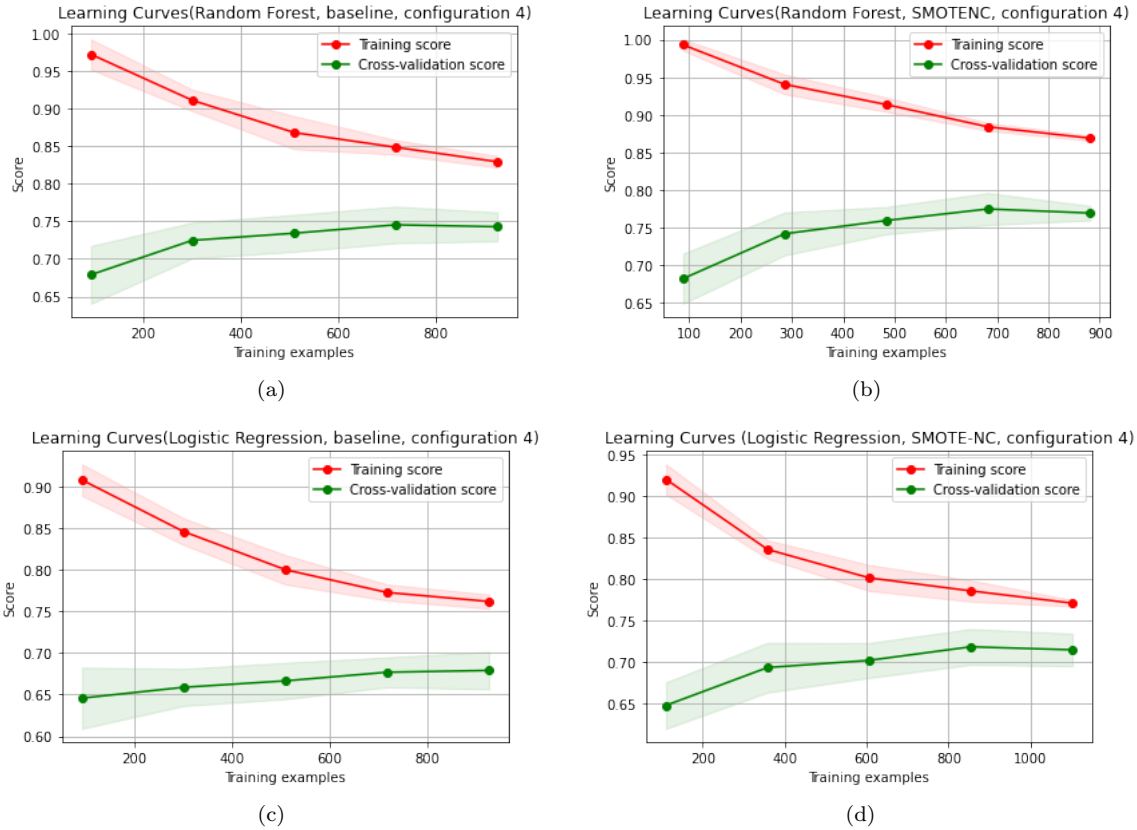

Figure 1: (a) and (b): training and testing learning curves for the baseline Random Forest and SMOTE-NC Random Forest models respectively with mixed APROCCHSS and RECORDS data from day 0 to day 2, no placebo (configuration 4); (c) and (d): training and testing learning curves for the tuned baseline Logistic Regression and SMOTE-NC Logistic Regression models respectively with mixed APROCCHSS and RECORDS data from day 0 to day 2, no placebo (configuration 4)

General Results Across all configurations, Random Forests were found to be the best-performing models. This could be explained by the fact that Random Forests are more sensitive to class imbalance, and the SMOTE method used previously is not well suited to the characteristic of the dataset (mixed type data). The performances of Random Forests using SMOTE-NC data are significantly improved and the predictive performance of the proposed algorithms are very satisfactory as evidenced by the evaluation metrics results. Another thing to note is that the confidence intervals are generally narrower in configurations 1, 2 and 3 than in configuration 4 : the size order of confidence interval length in configuration 1, 2 and 3 are 0.01-0.02 meanwhile it is 0.1 in configuration 4. This could be explained by the imputation of a large number of variables in configuration 4.

4 Discussion

The present study has shown that Random Forest models, when handling properly the class imbalance problem given the mixed nature of the data performed better than Logistic Regression models in terms of prediction accuracy, AUC and recall. In our case, the improvement of the classifier's sensitivity that is assessed by the recall comes at the price of a reduced increase in accuracy. This is due to

the fact that in being more sensitive and thus better classifying items from the minority class, the chances of missclassifying elements from the dominant class increase and the accuracy is thus reduced. Furthermore, we can see that mixing the RECORDS and APROCCHSS data as in configuration 3 and configuration 4 also yielded promising results, which might indicate that the trained classifier could be used on other cohorts in the future since the very good predictive results show the generalization ability of the suggested algorithms.

4.1 Data stratification

Since the effect of corticosteroids has been shown to be heterogeneous and highly variable from one individual to another, it might be interesting in further studies to look for relevant variables on which we could stratify the data. However this is a difficult question to answer and our data exploration phase using unsupervised learning to do clustering did not yield any conclusive results on that aspect. The aim of this study was to observe what are the most distinctive features of both datasets and see how these features relate to our variable of interest on cortico-sensitivity and asses what are the differences between both datasets and how these differences might relate to cortico-sensitivity. To better understand the underlying mechanisms of patients' sensitivity to corticosteroids, we only used the no-placebo data.

The clustering analysis did not yield conclusive results for two main reasons. The cluster separation was not good quality and overall unreliable and more importantly, there was no observable link between the clusters and label repartition. In addition, the variables influencing the most classification were the scores indicating the gravity of the patient's state (such as the SOFA, MACCABE, and KNAUS indicators), which were used to compute the label on day 14.

4.2 Variable selection : model interpretability and robustness

In the present study we used several metrics as presented above to assess each variables' influence on classification to improve model interpretability. However, we encountered several issues, including the fact that for each configuration and model, the variables that were assessed to be the most important were inconsistent with each others.

In order to use the most natural variable importance metric depending on the model, we used different approaches to evaluate variable importance from one model to the other. A possibility to allow for better comparison would be to adopt a similar approach to evaluate variable importance for both the Logistic Regression and the Random Forests models such as permutation importance for instance, which corresponds to the decrease in a model score when a single feature value is randomly shuffled. This measure is based on experiments on out-of-bag (OOB) samples in the case of Random Forests, and the main idea is to void the predictive power of a feature without changing its marginal distribution by randomly permuting the values of a feature in the OOB samples and seeing how it affects the overall model performance. The decrease in accuracy as a result of this permuting is averaged over all trees, and is used as a measure of the importance of the feature considered in the Random Forest.

4.3 Generalization issues : handling missing data and class imbalance with Random Forests

In previous studies Hellali et al. [2024], following clinicians' hypotheses, the APROCCHSS and RECORDS datasets were not mixed together, since it was suspected that there might be a difference due to the high number of covid patients in the RECORDS dataset. In addition, the only model involving both

datasets in Hellali et al. [2024] that was trained on APROCCHSS data and tested on RECORDS data systematically under-performed in comparison with other models and shows very poor predictive results. In the present study, we noticed that the observed under-performance of the previously considered models might have been partly due to the class imbalance problem that we addressed using the SMOTE-NC variant of the class rebalancing SMOTE method to cope with mixed-type data.

The overarching aim of the RECORDS project is to develop point-of-care testing to quickly assess a septic patient's chances of being sensitive to corticosteroid treatment. To do that, it is crucial to assess to what extent the classifier trained on RECORDS and APROCCHSS data can accurately predict the label for patients with different characteristics. The results on the RECORDS and APROCCHSS hybrid dataset showed that it is a reasonable hypothesis to assume that the data from RECORDS and APROCCHSS were generated by the same underlying distribution. This indicates that the trained classifier might perform adequately on other cohorts with different characteristics. To strengthen that hypothesis, further work might involve looking for multivariate tests of law comparison that might apply to our data sets.

The imputation problem One of the main stakes of using the RECORDS and APROCCHSS datasets conjointly is the handling of missing data since there are features recorded in the APROCCHSS dataset that are not present in the RECORDS trial data. In this case, the solution we used to join both datasets while retaining as much information as possible was to predict the missing features of RECORDS using Random Forest imputation on APROCCHSS data. However, this approach has several issues that are still unaccounted for in the present study:

- Generating a large number of unobserved variables for the RECORDS data based on APROCCHSS data might erase cohort-specific effects and create a false sense of generalizability that we cannot properly assess.
- Although our hybrid approach (configuration 4) yields very satisfactory results, it might not be immediately generalizable to other datasets. For instance, there could be additional key variables that are not well predicted, or variables introducing biases in the classification.

Injecting new RECORDS data into the analysis The main issue we encountered when trying to include RECORDS data into the analysis based on the APROCCHSS dataset was that since the RECORDS clinical study is still ongoing, the data we had access to for the present study was incomplete, and data collection strategy evolved since the APROCCHSS data collection period which means that:

- Some key variables present in the APROCCESS dataset are absent from the RECORDS trial data for now.
- The variable from the APROCCHSS dataset and the RECORDS data do not exactly match, which means that some variables do not have the same name, which makes the data alignment more difficult, or some variables that should be equivalent were collected differently, which introduces systematic differences between RECORDS and APPROCCHESS data.
- The larger RECORDS study in double blind has not been completed yet, so all the information about the data is not available.

Dealing with imbalanced data Random Forests are built on decision trees, and decision trees are sensitive to class imbalance in the sense that each tree is built on a "bag", and each bag is uniformly sampled with replacement from the training data, which means that each tree is biased on average in

the same direction and magnitude as the class imbalanced.

In the result section presented above, the strategy used to manage class imbalance for both Random Forests and Logistic Regression models was the SMOTE-NC method, which relies on generating synthetic samples of the minority class. In the case of Random Forests, we also considered different options to deal with the issue of class imbalance: two types of class weighting can help mitigate the effect of imbalance in classification tasks. The first technique is to weigh the tree-splitting criterion (Agusta et al. [2019]) and the other is to oversample or undersample data points during bootstrap sampling (Winham et al. [2013]). However, we did not observe significant differences between these methods in terms of predictive performance. Further work might involve comparing these methods using different metrics than the ones considered in the present study.

5 Conclusion

In the present study we defined a statistical approach to the supervised learning study results based on the APROCCHSS and RECORDS datasets. To obtain comparable results with those previously obtained (Hellali et al. [2024]), we carried out a study of several machine-learning models using appropriate tools to leverage mixed type and compared their performance on a binary classification task to build classifiers predicting patient's responsiveness to corticosteroids.

A second phase departing from the previous study explored different model specifications to assess the generalizability of the Logistic Regression and Random Forest models by testing it using additional data collected over time or from different patient groups belonging to the APROCCHSS and RECORDS cohort using tools adapted to mixed type data.

We saw that using the appropriate class balancing technique significantly improves models' performance in predicting corticosteroids sensitivity and allow for generalizations from cohort APROCCHSS to RECORDS. Further work might involve discussing the definition of the label with medical experts in light of previous machine learning studies and comparing classifiers based on different labels. In particular, the current label is based on variables assessing the gravity of the patient's state at a given moment. It might be useful to test different types of labels taking into account the evolution of the patient's state for instance.

6 Acknowledgments

The APROCCHSS study protocol and qualification of all investigators involved were approved by the Ethics Committee (Comité de Protection des Personnes, CPP) of Saint-Germain-en-Laye, France, on November 22, 2007. The RECORDS study protocol and qualification of all investigators involved were approved by the Ethics Committee (Comité de Protection des Personnes, CPP) of Dijon, France, on 6 April 2020. RECORDS is funded by public grant "Programme d'Investissements d'Avenir" (PIA), part of "France 2030", reference ANR-18-RHUS-0004.

RECORDS collaborators list : Alexandrou Antigoni, Annane Djillali, Arlt Birte, Badie Julio, Benghanem Sarah, Berdaguer Ferrari Fernando, Cerf Charles, Chelly Dagdia Zaineb, Chevret Sylvie, Colin Gwenhaël, Daniel Christel, Declercq Pierre-Louis, Delbove Agathe, Derridj Nawal, Devillier Philippe, Fleuriet Jérome, François Bruno, Garchon Henri-Jean, Godot Véronique, Grassin-Delyle

Stanislas, Grimaldi Lamiae, Grisolia Mathieu, Guitton Christophe, Helms Julie, Heming Nicholas, Herzog Marielle, Kamel Toufik, Kedad Zoubida, Lassalle Philippe, Lhermite Guillaume, Megarbane Bruno, Mekontso Dessap Armand, Mercier Emmanuelle, Meziani Ferhat, Mira Jean-Paul, Monchi Mehran, Monnet Xavier, Muller Grégoire, Plantefève Gaëtan, Quenot Jean-Pierre, Reignier Jean, Robine Adrien, Rottman Martin, Roux Anne-Laure, Schneider Francis, Siami Shidasp, Tissieres Pierre, Troché Gilles, Uhel Fabrice, Zeitouni Karine.

We thank Rahma Hellali, Zaineb Chelly Dagdia, Ahmed Ktaish, Karine Zeitouni and Djillali Annane for sharing their preprocessing code used in Hellali et al. [2024].

7 Declaration of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported here.

References

Zahra Putri Agusta et al. Modified balanced random forest for improving imbalanced data prediction. *International Journal of Advances in Intelligent Informatics*, 5(1):58–65, 2019.

Djillali Annane, Eric Bellissant, Pierre-Edouard Bollaert, Josef Briegel, Marco Confalonieri, Raffaele De Gaudio, Didier Keh, Yizhak Kupfer, Michael Oppert, and G Umberto Meduri. Corticosteroids in the treatment of severe sepsis and septic shock in adults: a systematic review. *Jama*, 301(22): 2362–2375, 2009.

Djillali Annane, Eric Bellissant, Pierre Edouard Bollaert, Josef Briegel, Didier Keh, and Yizhak Kupfer. Corticosteroids for treating sepsis. *Cochrane database of systematic reviews*, (12), 2015.

Djillali Annane, Alain Renault, Christian Brun-Buisson, Bruno Megarbane, Jean-Pierre Quenot, Shidasp Siami, Alain Cariou, Xavier Forceville, Carole Schwebel, Claude Martin, et al. Hydrocortisone plus fludrocortisone for adults with septic shock. *New England Journal of Medicine*, 378(9):809–818, 2018.

Christopher Barton, Uli Chettipally, Yifan Zhou, Zirui Jiang, Anna Lynn-Palevsky, Sidney Le, Jacob Calvert, and Ritankar Das. Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Computers in biology and medicine*, 109:79–84, 2019.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Fang Fang, Yu Zhang, Jingjing Tang, L Dade Lunsford, Tiangui Li, Rongrui Tang, Jialing He, Ping Xu, Andrew Faramand, Jianguo Xu, et al. Association of corticosteroid treatment with outcomes in adult patients with sepsis: a systematic review and meta-analysis. *JAMA internal medicine*, 179 (2):213–223, 2019.

Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

Carolin Fleischmann, André Scherag, Neill KJ Adhikari, Christiane S Hartog, Thomas Tsaganos, Peter Schlattmann, Derek C Angus, and Konrad Reinhart. Assessment of global incidence and mortality

of hospital-treated sepsis. current estimates and limitations. *American journal of respiratory and critical care medicine*, 193(3):259–272, 2016.

Lucas M Fleuren, Thomas LT Klausch, Charlotte L Zwager, Linda J Schoonmade, Tingjie Guo, Luca F Roggeveen, Eleonora L Swart, Armand RJ Girbes, Patrick Thoral, Ari Ercole, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive care medicine*, 46:383–400, 2020.

Alison E Fohner, John D Greene, Brian L Lawson, Jonathan H Chen, Patricia Kipnis, Gabriel J Escobar, and Vincent X Liu. Assessing clinical heterogeneity in sepsis through treatment patterns and machine learning. *Journal of the American Medical Informatics Association*, 26(12):1466–1477, 2019.

Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. In *2008 Fourth international conference on natural computation*, volume 4, pages 192–201. IEEE, 2008.

Leonardo Gutiérrez-Gómez, Frank Petry, and Djamel Khadraoui. A comparison framework of machine learning algorithms for mixed-type variables datasets: a case study on tire-performances prediction. *IEEE Access*, 8:214902–214914, 2020.

Khan Md Hasib, Md Sadiq Iqbal, Faisal Muhammad Shah, Jubayer Al Mahmud, Mahmudul Hasan Popel, Md Imran Hossain Showrov, Shakil Ahmed, and Obaidur Rahman. A survey of methods for managing the classification and solution of data imbalance problem. *arXiv preprint arXiv:2012.11870*, 2020.

Rahma Hellali, Zaineb Chelly Dagdia, Ahmed Ktaish, Karine Zeitouni, and Djillali Annane. Corticosteroid sensitivity detection in sepsis patients using a personalized data mining approach: a clinical investigation. *Computer Methods and Programs in Biomedicine*, 2024.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Deepak A Kaji, John R Zech, Jun S Kim, Samuel K Cho, Neha S Dangayach, Anthony B Costa, and Eric K Oermann. An attention based deep learning model of clinical events in the intensive care unit. *PloS one*, 14(2):e0211057, 2019.

Didier Keh and Charles L Sprung. Use of corticosteroid therapy in patients with sepsis and septic shock: an evidence-based review. *Critical care medicine*, 32(11):S527–S523, 2004.

Matthieu Komorowski, Ashleigh Green, Kate C Tatham, Christopher Seymour, and David Antcliffe. Sepsis biomarkers and diagnostic tools with a focus on machine learning. *EBioMedicine*, 86, 2022.

Huoyan Liang, Heng Song, Ruiqing Zhai, Gaofei Song, Hongyi Li, Xianfei Ding, Quancheng Kan, and Tongwen Sun. Corticosteroids for treating sepsis in adult patients: a systematic review and meta-analysis. *Frontiers in Immunology*, 12:709155, 2021.

Brit Long and Alex Koyfman. Controversies in corticosteroid use for sepsis. *The Journal of emergency medicine*, 53(5):653–661, 2017.

Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. *Advances in neural information processing systems*, 26, 2013.

Andrea McCoy and Ritankar Das. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ open quality*, 6(2), 2017.

Michael Moor, Bastian Rieck, Max Horn, Catherine R Jutzeler, and Karsten Borgwardt. Early prediction of sepsis in the icu using machine learning: a systematic review. *Frontiers in medicine*, 8: 607952, 2021.

James M. O'Brien, Naeem A. Ali, Scott K. Aberegg, and Edward Abraham. Sepsis. *The American Journal of Medicine*, 120(12):1012–1022, 2007. ISSN 0002-9343. doi: https://doi.org/10.1016/j.amjmed.2007.01.035. URL `https://www.sciencedirect.com/science/article/pii/S0002934307005566`.

Romain Pirracchio, Alan Hubbard, Charles L Sprung, Sylvie Chevret, Djillali Annane, et al. Assessment of machine learning to estimate the individual treatment effect of corticosteroids in septic shock. *JAMA network open*, 3(12):e2029050–e2029050, 2020.

Andrew Rhodes, Laura E Evans, Waleed Alhazzani, Mitchell M Levy, Massimo Antonelli, Ricard Ferrer, Anand Kumar, Jonathan E Sevransky, Charles L Sprung, Mark E Nunnally, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. *Intensive care medicine*, 43:304–377, 2017.

Bram Rochwerg, Simon J Oczkowski, Reed AC Siemieniuk, Thomas Agoritsas, Emilie Belley-Cote, Frédérick D'Aragon, Erick Duan, Shane English, Kira Gossack-Keenan, Mashari Alghuroba, et al. Corticosteroids in sepsis: an updated systematic review and meta-analysis. *Critical care medicine*, 46(9):1411–1420, 2018.

Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet*, 395(10219):200–211, 2020.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

D Roland Thomas, PengCheng Zhu, Bruno D Zumbo, and Shantanu Dutta. On measuring the relative importance of explanatory variables in a logistic regression. *Journal of Modern Applied Statistical Methods*, 7(1):4, 2008.

Hans-Christian Thorsen-Meyer, Annelaura B Nielsen, Anna P Nielsen, Benjamin Skov Kaas-Hansen, Palle Toft, Jens Schierbeck, Thomas Strøm, Piotr J Chmura, Marc Heimann, Lars Dybdahl, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *The Lancet Digital Health*, 2(4):e179–e191, 2020.

Tom van der Poll, Manu Shankar-Hari, and W Joost Wiersinga. The immunology of sepsis. *Immunity*, 54(11):2450–2464, 2021.

Balasubramanian Venkatesh, Simon Finfer, Jeremy Cohen, Dorrilyn Rajbhandari, Yaseen Arabi, Rinaldo Bellomo, Laurent Billot, Maryam Correa, Parisa Glass, Meg Harward, et al. Adjunctive glucocorticoid therapy in patients with septic shock. *New England Journal of Medicine*, 378(9): 797–808, 2018.

Scott L Weiss, Mark J Peters, Waleed Alhazzani, Michael SD Agus, Heidi R Flori, David P Inwald, Simon Nadel, Luregn J Schlapbach, Robert C Tasker, Andrew C Argent, et al. Surviving sepsis campaign international guidelines for the management of septic shock and sepsis-associated organ dysfunction in children. *Intensive care medicine*, 46:10–67, 2020.

Stacey J Winham, Robert R Freimuth, and Joanna M Biernacka. A weighted random forests approach to improve predictive performance. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(6):496–505, 2013.