Corticosteroid sensitivity detection in sepsis patients using a personalized data mining approach: a clinical investigation

Rahma Hellali, Zaineb Chelly Dagdia, Ahmed Ktaish, Karine Zeitouni, Djillali Annane

Corticosteroid sensitivity detection in sepsis patients using a personalized data mining approach: a clinical investigation

Rahma Hellali[a], Zaineb Chelly Dagdia[a], Ahmed Ktaish[b], Karine Zeitouni[a], Djillali Annane[c]

[a]*Université Paris-Saclay, UVSQ, DAVID, Paris, France*
[b]*GCI, France*
[c]*Réanimation medico-chirurgicale, hôpital Raymond-Poincaré, AP–HP, Garches, France*

Abstract

Background and Objective: Sepsis is a life-threatening disease with high mortality, incidence, and morbidity. Corticosteroids (CS) are a recommended treatment for sepsis, but some patients respond negatively to CS therapy. Early prediction of corticosteroid responsiveness can help intervene and reduce mortality. In this study, we aim to develop a data mining methodology for predicting CS responsiveness of septic patients.
Methods: We used data from a randomized controlled trial called APROCCHSS, which recruited 1241 septis patients to study the effectiveness of corticotherapy. We conducted a thorough study of multiple machine learning models to select the most efficient prediction model, called "signature". We evaluated the performance of the signature using precision, sensitivity, and specificity values.
Results: We found that Logistic Regression was the best model with an AUC of 72%. We conducted further experiments to examine the impact of additional features and the model's generalizability to different groups of patients. We also performed a statistical analysis to analyze the effect of the treatment at the individual level and on the population as a whole.
Conclusions: Our data mining methodology can accurately predict cortico-sensitivity or resistance in septis patients. The signature has been deployed into the Assistance Publique – Hôpitaux de Paris (APHP) information system as a web service, taking patient information as input and providing a prediction of cortico-sensitivity or resistance. Early prediction of corticosteroid responsiveness can help clinicians intervene promptly and improve patient outcomes.

Keywords: Sepsis, Corticosteroids, Sensitive or resistant, Machine learning, Clinical decision support system.

1. Introduction

Sepsis is a life-threatening organ dysfunction caused by dysregulation of the host's response to infection that necessitates special medical treatment in the intensive care units of hospitals [1]. The sepsis disease life cycle involves three main stages: Systemic Inflammatory Response Syndrome (SIRS), severe sepsis, and septic shock [2]. In order to fight infection, the body releases additional immune system chemicals into the bloodstream when they are damaged. This is called the released immune system stage of sepsis, which can be extremely dangerous if it progresses rapidly. Severe sepsis occurs if the initial sepsis is not treated or does not meet treatment and may impact organ function. The symptoms of septic shock are similar to those of severe sepsis, but they also include a significant drop in blood pressure. This drop in blood pressure can lead to heart failure, stroke, other organ failures, respiratory failure, and even death [3]. The global sepsis case number is difficult to determine. An estimation has been made in 2017 indicating there were 48.9 million cases and 11 million deaths due to sepsis recorded worldwide, which represented about 20% of all deaths worldwide [4]. In France, 2019 incidence was 403/100,000 (357 in 2015), mortality 23%, disability 15%, and cost ~16,000€ per patient [5]. Also, maternal sepsis happens when sepsis occurs during pregnancy, during or after childbirth, or after a miscarriage. We also talk about neonatal sepsis when newborns are affected by the sepsis disease. Although highly preventable, maternal and neonatal sepsis remains one of the most serious causes of death in pregnant women and newborns[1].

With the significant progress of the clinical best practices and the pharmaceutical industry, the risk of death was considerably reduced. However, the number of death cases still increases depending on the global number of varied hospitalized cases [6]. Therefore, the major challenge behind the sepsis mortality decrease is how to administrate the right treatment at the right time.

Corticosteroid (CS) therapy has been shown to be related to the majority of sepsis patients in whom age, sex, disease severity, type of infection, source of infection, or type of pathogen do not influence survival benefit. However, the response to corticotherapy depends on the patient. The precise factors and biomarkers of responsiveness are various and complex [7]. The purpose of this paper is to conduct, in the context of a real-world blinded randomized study, called RECORDS, a thorough clinical investigation on patients' responsiveness to CS by apply-

[1] `https://www.who.int/health-topics/sepsis#tab=tab_1`

ing a robust data mining approach. Sepsis patients' data have been gathered from the Assistance Publique – Hôpitaux de Paris (APHP)[2], the largest French hospital system in Europe and one of the largest in the world, with the goal to predict if sepsis patients' are CS sensitive or CS resistant; thus contributing to a better understanding of this condition.

We adopted a consistent data mining approach and developed a prediction model, referred to as "signature", which aims to promptly identify the responsiveness of patients to corticotherapy. This signature guiding corticotherapy is then used by the APHP clinicians as one of the biomarker strata during the randomization process [8]. In fact, after the inclusion of patients, the RECORDS trial protocol stipulates the use of eleven biomarkers to randomize the patients that are: CIRCI, Endocan, GILZ, DUSP-1, MDW, Transcriptomic SRS2, Endotype B, COVID-19, Influenza, other respiratory viruses, and Cutaneous vasoconstrictor response to glucocorticoids. Besides these tests, machine learning algorithms are also considered as biomarkers. Two algorithms have been deployed, our signature and another one trained on an international database [9]. The randomization follows a bayesian process to assign the treatment arm to the patient. This flowchart is highlighted in Figure 1; where our signature is referred to as "AI signature 1".

Numerous studies have employed machine learning techniques to predict sepsis, evaluate sensitivity to corticosteroids, and explore related topics. These studies are elaborated upon in what follows.

- **Corticosteroid sensitivity prediction**

  In [10], experts showed that adults suffering from severe sepsis do not benefit from using hydrocortisone. After analyzing several statistical tests like Fisher, chi-square, t-test, Mann-Whitney U, and log-rank, no significant differences in 28-, 90-, or 180-day mortality rates can be found between patients who received the steroid or not. The 28-day mortality of all sepsis patients has not improved after hydrocortisone treatment, as shown in [11]. Despite having demonstrated that this treatment reduces cardiovascular organ failure, it did not decrease mortality.

  However, a comparative study on the effect of hydrocortisone in patients with septic shock was performed in [12]. Using a logistic regression model adjusted for variable stratification, authors demonstrated that patients receiving hydrocortisone therapy had a more rapid resolution of septic shock. Furthermore, the work of [13] provides an effect analysis of hydrocortisone plus fludrocortisone therapy in resolving organ failure in adults suffering from septic shock. They used regression models, chi-square, and t-tests to examine the effect of the test substance on the incidence of fatal events. Their research revealed that the use of hydrocortisone plus fludrocortisone was associated with a lower rate of all-cause mortality compared to placebo at 90 days, discharge from Intensive Care Unit (ICU) and hospital, and 180 days.

  In [9] authors applied an ensemble machine learning approach and statistical analysis to study the Individual Treatment Effect (ITE). They found that treating adults with septic shock with corticosteroids resulted in a positive impact on these patients.

  Authors in [14] identified a subgroup of pediatric septic shock patients who may benefit from corticosteroid treatment. The study used a combination of prognostic and predictive strategies based on biomarkers to assign study subjects to different endotypes. The primary endpoint was a complicated course, defined as the persistence of two or more organ failures at day seven of septic shock or 28-day mortality. The study found that among patients with intermediate to high risk of mortality, corticosteroids were associated with a significant reduction in the risk of a complicated course. The authors suggested that their findings support the use of precision medicine strategies to identify patients who are most likely to benefit from corticosteroids.

- **Predictive models and sepsis prediction**

  Authors in [15, 16, 17, 18, 19] conducted a systematic review of machine learning models for sepsis prediction. They used different data types and machine learning models to analyze data when diagnosing sepsis symptoms. In particular, authors in [20] considered that early fluid therapy in sepsis patients may not be effective and may possibly cause major side events since the fluid is not responding. The purpose of this work was to construct prediction models for evaluating fluid responsiveness in sepsis patients using the MIMIC III dataset and associated matched waveform datasets during the entire ICU stay duration of each patient. In order to extract high-frequency continuous waveform data, they created a pipeline, and waveform properties were incorporated into the prediction models. The best Area under the ROC curve (AUC) value was registered for Random Forest when no waveform information was supplied with a value of 84% compared to XGBoost [21] with a value of 64%, Linear SVM [22] with a value of 72%, and SVM Polynomial [23] with a value of 65%; with mean arterial blood pressure and age being the main determining factors.

  In [24], authors proposed a machine learning method for early and effective diagnosis of sepsis using metabolic data from blood samples. The proposed method, which combines random forest feature selection with a kernel extreme learning machine improved by a chaotic fruit fly optimization algorithm, achieved a recognition rate of 81.6%, sensitivity of 89.57%, and specificity of 65.77%. The study also identified five biomarkers that showed promising diagnostic potential for sepsis. The results suggested that the proposed methodology can be a useful diagnostic tool for clinical decision support.

[2] `https://www.aphp.fr/`

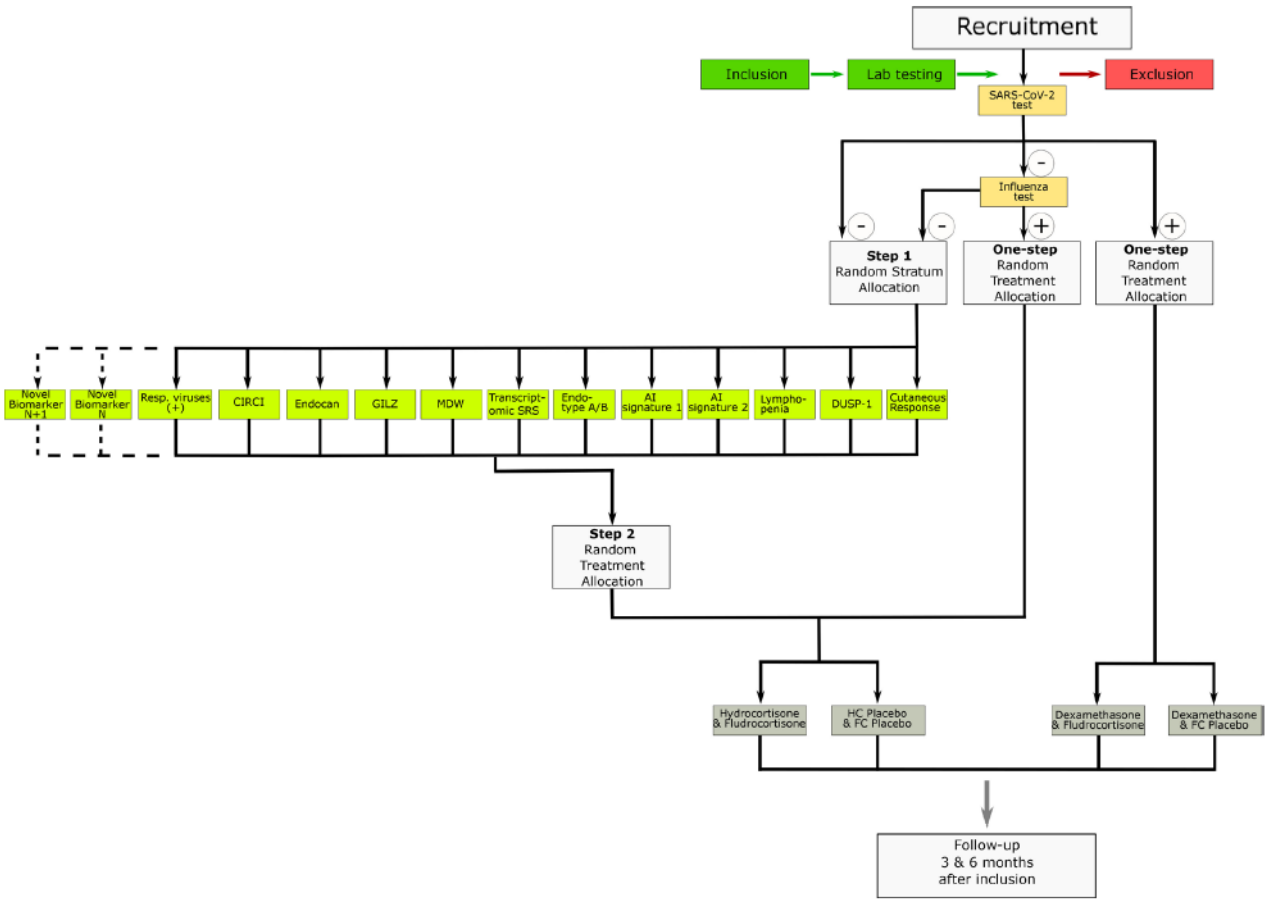

Figure 1: Flowchart illustrating the recruitment and randomization procedure of the study

In [25], authors presented the effectiveness of a broad range of standard and ensemble machine learning techniques for predicting cardiac arrest for adult sepsis patients. The analysis was done using a systematic approach. The input features incorporate the patients' vital signs time series, recorded over a period of 30 hours for each individual.

Authors in [26] employed a Random Forest (RF) classifier to identify patients at an early stage who were at risk of sepsis. The RF model was trained to distinguish septic and non-septic patients using 132 features which were extracted from physiological measurements such as heart rate, blood pressure, mean arterial pressure, and temperature. Features were retrieved from a moving window of 3 to 6 hours. The authors also showed that the RF model can be applied with the inclusion of additional characterization of leukocytes.

In [27], authors address the urgent need for early sepsis detection by discussing the PhysioNet/Computing in Cardiology Challenge 2019, which aimed to develop open-source algorithms for this purpose. The challenge involved participants submitting algorithms to a cloud-based testing environment where they were evaluated using a novel clinical utility-based metric, rewarding early predictions and penalizing late or missed predictions and false alarms. Data from over 60,000 ICU patients with a range of clinical variables were sourced and analyzed using Sepsis-3 clinical criteria for sepsis onset. The study's findings indicate that various computational approaches can predict sepsis onset several hours before clinical recognition, although generalizability across different hospital systems remains a challenge.

- **Related topics**

Authors in [28] presented a new approach to deduce optimal treatment policies for septic patients by using continuous state-space models and deep reinforcement learning. The predictive models were developed to identify two primary medical interventions: the volume of intravenous (IV) fluid (adjusted for fluid tonicity) and the maximum vasopressor (VP) dosage administered within a 4-hour timeframe. The model was able to learn clinically interpretable treatment policies and could reduce absolute patient mortality in the hospital by up to 3.6% over observed clinical policies. The learned treatment policies could aid intensive care clinicians in medical decision-making and improve the likelihood of patient survival. Outside the sepsis context but still related to corticother-

apy, authors in [29] used gradient-boosted decision-tree models to identify which COVID-19 patients treated with either corticosteroids or remdesivir had improved survival times. The models were trained and tested on electronic health records from 2364 adult patients in 10 US hospitals. After adjusting for confounding, the study found that neither corticosteroids nor remdesivir use was associated with increased survival time in the overall population or in the subpopulation that received supplemental oxygen. However, the machine learning models identified subpopulations in which both corticosteroids and remdesivir were significantly associated with an increase in survival time. Specifically, in these subpopulations, corticosteroids and remdesivir had hazard ratios of 0.56 and 0.40, respectively, with both results having a p-value of 0.04. These findings suggest that machine learning methods can improve patient outcomes and allocate resources during the COVID-19 crisis by identifying which patients are most likely to benefit from these treatments.

The presented state-of-the-art review sets the stage for the novel contributions of our research. While previous studies have made substantial strides in understanding sepsis and corticosteroid responsiveness, our research takes a unique path by presenting a real case study in collaboration with the APHP hospital, bridging the gap between cutting-edge research and the clinical environment. What sets this work apart is its commitment to predicting corticosteroid responsiveness as early as day 0 when the patient enters the hospital, utilizing a comprehensive range of individual traits, from demographics to clinical factors.

Moreover, the novelty of this study is underscored by our comprehensive experimental approach. We conducted experiments systematically, incrementally adding patient features to predictive models, assessing their impact on prediction effectiveness, and notably, doing so within a timeframe of up to two days. This brief evaluation period for feature impact is a distinctive aspect of our work, as previous studies often extended this duration considerably, typically up to day 7, 28, and 90 [12, 13, 9, 14, 10, 8, 11]. By focusing on the initial 48 hours, we provide a unique perspective on the early prediction of corticosteroid responsiveness, offering clinicians a practical and real-time window to make informed treatment decisions.

Furthermore, this study delves into uncovering the individual-level effects of corticotherapy treatment through an examination of both statistical analysis and predictive models. To accomplish this, we generate predictions for patients who received corticotherapy and those who received a placebo. Subsequently, the calculated values of individual treatment effects represent the disparity between these two sets of predictions. A notable contribution of our research is our innovative interpretation of the statistical study. In contrast to existing literature, our calculations are based on correctly classified patients, enhancing the precision of our statistical analysis. This approach not only refines the accuracy of our findings but also lays the groundwork for future research directions, including a detailed exploration of patient characteristics associated with beneficial and detrimental treatment outcomes.

What further emphasizes the novelty and real-world impact of our work is the deployment of our signature, which serves as a guide for corticotherapy, within the APHP hospital. This signature, denoted as "AI signature 1" in Figure 1, is actively used by APHP clinicians as one of the biomarker strata during the randomization process, as previously highlighted in the beginning of Section 1. This brings cutting-edge predictive models directly into the hands of clinicians for timely and informed decisions.

To give a clearer context about the way how our signature is used at APHP, in a related study [8], our collaborative specialists at APHP have provided a study flowchart (Figure 1) describing the recruitment and randomization process. Initially, patients are stratified based on various criteria, including disease strata (e.g., COVID-19, influenza virus, non-influenza respiratory virus), and specific biomarkers or signatures. These biomarkers are selected based on prior research, observational data, and national/international guidelines. In the next step, each stratum is precisely defined according to specific criteria relevant to the chosen biomarkers. For instance, the COVID-19 stratum may involve the randomization of patients to receive dexamethasone in combination with fludrocortisone or a placebo. Subsequently, the protocol identifies a set of candidate biomarkers/signatures (e.g., CIRCI, endocan, MDW, lymphocyte count, transcriptome sepsis response signature, and machine learning algorithms (i.e., including our signature "AI signature 1")) that are used for patient stratification. These biomarkers play a crucial role in distinguishing various patient profiles and treatment response characteristics. The final step involves treatment assignments based on the randomization process, which can include the administration of corticosteroid treatment or a placebo. This step is designed to assess the effects of the treatment based on the strata and biomarkers.

Based on the aforementioned studies, it is evident that early prediction of corticosteroid responsiveness is crucial for optimizing patient care and improving clinical outcomes. Our research aims to study this critical issue by developing a predictive model that can accurately predict corticosteroid responsiveness as early as possible. With respect to this aim, we study the effect of patients' features, as early as day two of treatment, on the prediction effectiveness of CS responsiveness.

This is accomplished by gathering patients' data and by using signatures of CS sensitivity based on each individual's traits, such as genetic, demographic, and clinical factors [9, 30]. We aim to improve the accuracy and efficiency of early prediction, which will ultimately aid in individualized treatment and better patient outcomes. This study has significant implications for the field of critical care medicine, as it provides a novel approach for early identification of corticosteroid responsiveness, leading to improved patient care and outcomes.

In our work, we conducted a thorough study of several machine learning models, including investigating models' configurations (e.g., hyperparameter tuning, top feature selection), and compared their performance in terms of recognizing CS responsiveness. From a clinical and data mining perspectives, we aim to answer the following fundamental questions:

- **RQ1:** Can predictive models recognize sepsis responders and non-responders to corticosteroid treatment?
- **RQ2:** How can patient features affect the accuracy of the obtained results?
- **RQ3:** What is the corticotherapy treatment effect at the individual patient level and at all treated patients?
- **RQ4:** Can the learned model (i.e., the signature) be generalized to different sepsis cohorts?

The rest of the paper is organized as follows: Section 2 describes the system architecture of our data mining pipeline, the preprocessing, the APHP sepsis data preparation, the model's actual deployment within the Assistance Publique – Hôpitaux de Paris information system, and the experimental setup. In Section 3, we thoroughly discuss the obtained results. Section 4 presents the discussion of the results, where we delve deeper into the implications of our findings. Section 5 concludes this paper and provides directions for future work.

2. Materials and methods

2.1. General system

In this section, we provide an introduction to the RECORDS project, in which we are conducting our research, and then explain our data mining methodology. We describe the system architecture that supports our clinical investigation.

2.1.1. Project background and knowledge acquisition

RECORDS[3], a national research project coordinated by APHP, lies at the crossroads of University Hospital and Industrial partners. The aim of RECORDS is the rapid detection of a patient's sensitivity or resistance to the treatment of sepsis with corticosteroids. The project's clinical trial is an adaptive clinical trial to assess the ability of biomarkers and algorithms derived from machine learning to define a patient's corticosteroid resistance and thus optimize their management.

This project uses data collected in a unique manner to properly analyze the severity of sepsis cases. Collecting information on patient demographics, health outcomes, and samples, led to the construction of a first sepsis cohort, dubbed APROCCHSS [13, 31], which provides a unique resource for medical research. These data sources have been used as a bootstrap of the analysis process in RECORDS. Indeed, our initial signature is learned on the APROCCHSS cohort.

To conduct our work on the APROCCHSS cohort, and to be able to develop the signature that will identify specific resistance or sensitivity to cortico-systemic drugs, knowledge acquisition seems essential. The sources of knowledge, in this work, include medical literature, the APHP domain expertise, and eventually the sepsis cohort.

2.1.2. System architecture

Our system's general architecture, composed of six key steps, is depicted in Figure 2. The first step in this process covers the acquisition of patients' sepsis related data. Next, data preprocessing is performed covering data cleaning and feature engineering. Features are selected for further processing based on whether or not the patient received corticosteroid treatment. Data preparation presents the third step in our data mining pipeline. It includes labeling the data, splitting the data into training and testing, data scaling, and class balancing. In the fourth step, machine learning algorithms are applied to choose the best-performing one in detecting patient CS sensitivity. This step includes four configurations using hyperparameters and feature selection methods. Then, in order to interpret and evaluate the considered models, statistical analyses are performed using different measures. Finally, in the last step, we present the deployment of our selected machine learning model within the APHP information system; where our model was packaged within an Application Programming Interface (API) using Flask[4]. It takes as input the patient's clinical data and returns whether the patient is sensitive or resistant to Corticosteroids. The following sections provide detailed information about each of these steps.

2.2. Clinical data description

2.2.1. The APROCCHSS cohort

The patients' data have been collected through an electronic Case Report Form (e-CRF) upon admission to the hospital. The medical staff enters all relevant personal and medical information for each patient, including demographics, medical history, current condition, and treatment details, into the e-CRF. This data collection process is time-sensitive, with a maximum window of 90 days to ensure the accuracy and completeness of the information recorded. The data collected through the e-CRF is used to build the APROCCHSS cohort and will require thorough data preprocessing and data preparation for machine learning. Figure 3 presents the flowchart that shows the screening process of the patient cohort.

The APROCCHSS cohort comprises data from 1241 patients, described by 5645 health characteristics (i.e., risk factors) with a specification of whether they were treated with corticosteroids or a placebo. Since we need to know the resistance/sensitivity of the patient to the treatment before actually taking the drug, only data before the hospitalization (i.e., Day 0) have been considered to build the machine learning signature. Table 1 presents the list of relevant selected features at Day 0 along with their description and their corresponding format. These features will undergo further preprocessing.

2.2.2. Characteristics of the cohort

The APROCCHSS cohort results from a randomized controlled trial. The advantage of randomization is that it limits selection bias, thus allowing known and unknown prognostic

[3] `https://www.fhu-sepsis.uvsq.fr/rhu-records`

[4] `https://flask.palletsprojects.com/en/2.2.x/`

Table 1: Considered set of relevant features at Day 0

Reference	Description	Format
DATINF	Diagnosis date	Precision = JJ/MM/YYYY, Min = DATHOSP (Hospital admission date), Max = Current date
SITINF	Infection location	0 = Lung, pleura, 1 = Peritoneal, 2 = Urogenital, 3 = Central Nervous System (CNS), 4 = Endocarditis, mediastinum, 5 = Sepsis, 6 = Soft tissue, 7 = Bones and joints, 8 = Other
ID1191S12V11	Indicates if the patient has bacteriological documentation	0 = No, 1 = Yes
EXAMINF_CHOICE	Indicates which examination has been performed on the patient	1 = Blood culture, 2 = Stool culture, 3 = Cytobacteriological examination of urine , 4 = Sinus puncture, 5 = Bronchial sample, 6 = Biopsy, 7 = Catheter, 8 = cerebrospinal fluid (CSF), 9 = Operation site, 10 = plevre, 11 = Peritoneum, 12 = Joint, 13 = Soft tissue, 14 = Prosthesis, 15 = Legionellosis diagnosis
ID1191S12V13_CHOICE	Indicates the type of medical Radiology the patient took	1 = Radio, 2 = ultrasound, 3 = Magnetic resonance imaging, 4 = Scanner
ID1191S12V12	Indicates the probability of infection	0 = Unlikely, 1 = Likely, 2 = Certainty
PREBROINF	Indicates the type of bronchial sampling the patient take	0 = Bronchial Aspiration, 1 = Brush $\geq$ 1.000 UFC/ml, 2 = Brush $<$ 1.000 UFC/ml, 3 = Combicath $\geq$ 1.000 UFC/ml, 4 = Combicath $<$ 1.000 UFC/ml, 5 = Expectoration, 6 = Bronchoalveolar lavage, 7 = Bronchoalveolar lavage
SEX	Indicates patient sex	1 = Male, 2 = Female
PATWGHT	Indicates the weight of the patient	Min = 36, Max = 154
ORIGIN	Indicates the patient ORIGIN	1 = City, 2 = Hospital, 3 = Institution
AGE	Indicates patient age	Min = 18, Max = 97
SOFA_INC2	Used in intensive care to identify and track the status of a patient in organ failure [32] and indicates the worst value calculated in the 6 hours prior to inclusion	Min = 8, Max = 17
KNAUS_J0	Activity and medical follow-up in the six months prior to admission	1 = Stage D Major activity restriction due to illness, including bedridden or hospitalized patients, 2 = Stage C Chronic illness causing significant but not total activity restriction, 3 = Stage B Moderate or moderate activity limitation due to illness (limited work activities), 4 = Stage A Good health, no activity limitation
MACCABE_J0	Description of the patient's condition before the episode leading to ICU	1 = Absence of underlying disease or underlying disease not life-threatening, 2 = Underlying disease life-threatening within 5 years, 3 = Underlying disease estimated to be fatal within one year
SOFA_ADM	Indicates the worst case value up to 3 hours after admission	Min = 2, Max = 16
ID1191S12V3	Indicates the temperature of the patient	Min = 10.0, Max = 50.0
IGS3_ADM	Index of Gravity simplified that indicates the worst case value up to 3 hours after admission	Min = 32, Max = 132
GLYCEMIC	Indicates the level of Blood Glucose	Unit: mmol/L
LACTATES_J0	Lactic acid is a blood glucose metabolite produced by body tissues when they are lacking oxygen	Unit: mmol/L
IGSII_ADM_TYP	Indicates the admission type of the patient	0 = Scheduled surgery, 6 = Medical, 8 = Unscheduled surgery

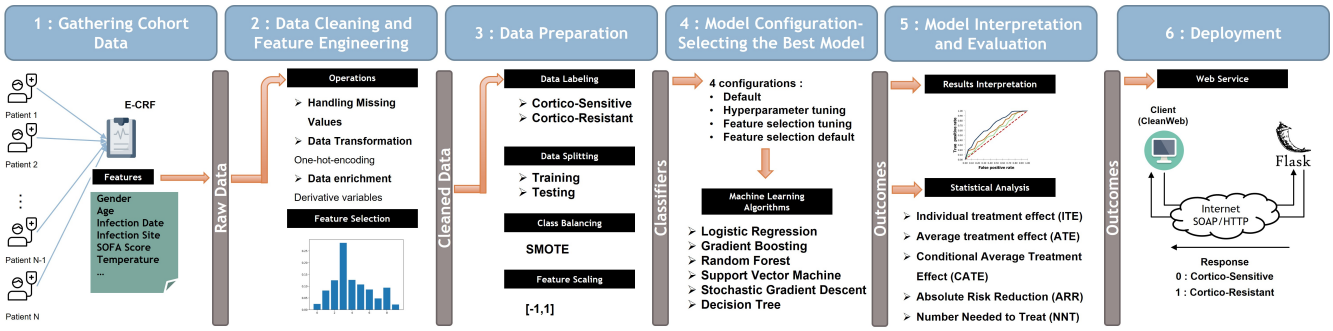

Figure 2: The system architecture

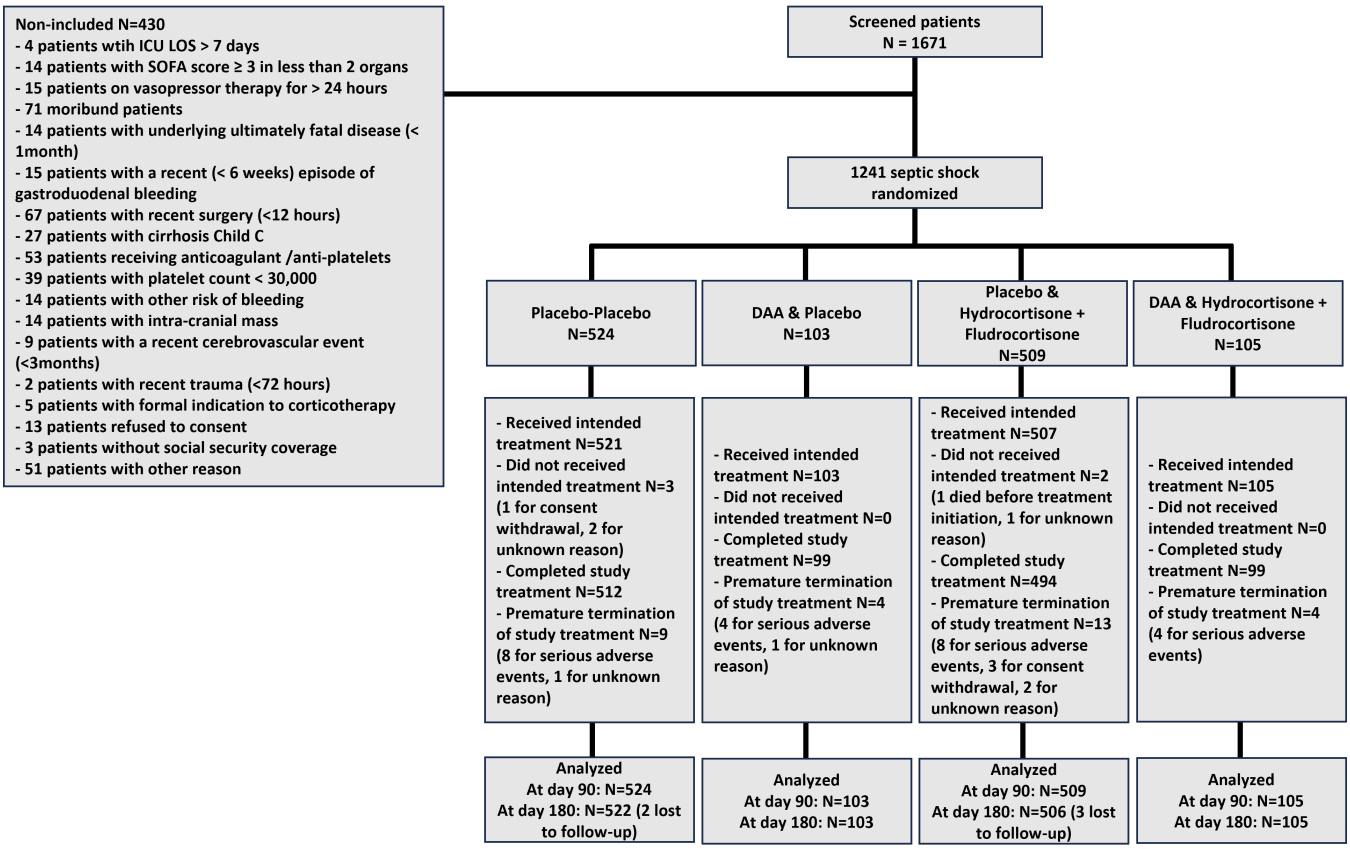

Figure 3: Patient APROCCHSS cohort screening process flowchart

factors to be evenly distributed between groups [33]. As previously mentioned, APROCCHSS contains information about 1241 sepsis patients. After reviewing the data by the medical experts, some patients were reclassified from cortico-resistant to cortico-sensible and some others were removed as they do not align with some defined criteria (see Section 2.5.1). It is also to be noted that one patient withdrew his/her consent to use his/her data. As a result, a total of 1234 sepsis patients remained in the cohort. As the aim of this clinical study is to primarily focus on investigating CS responsiveness, we will only include patients who received corticotherapy; a total of 612 patients. The remaining 622 patients who were treated with placebo will be included in the phase where we will conduct the evaluation of the treatment effectiveness.

2.3. Data cleaning and feature engineering

Data cleaning and feature engineering are essential steps in data analysis, particularly in critical fields like health. These tasks can impact the accuracy and effectiveness of machine learning models and ultimately affect the decision-making process. This study encountered several challenges in dealing with the collected raw sepsis data.

2.3.1. Feature selection

Sepsis is a time-sensitive disease, where early identification and treatment greatly increase the chances of survival. In order to optimize early detection, this study focuses on utilizing features available at the earliest stage, specifically at Day 0 of

hospitalization (Table 1), for predicting patients' sensitivity to corticotherapy. However, it is worth noting that additional features until Day 2 of hospitalization have been also considered in our experimental setup, in order to provide a more comprehensive understanding of the patients' health status during treatment with corticotherapy, and to study the impact that such features (from Day 0 to Day 2) can have on the prediction generated by the considered machine learning models. This will be later discussed in Section 2.7 and Section 3.

As we only consider the patient's data that are known at maximum Day 2, feature selection starts by pruning the unused variables related to the observations made during the hospitalization, until Day 90 (which reduces the number to 238 out of the initial pool of 5645 features). Then, we rely on the expertise of the medical team to select 24 essential features among the 238.

There are two categories of the collected data: static, referring to the metadata, and dynamic referring to the follow-up/monitoring data (Table 1). The first category consists of data on the current status of the patient as well as personal data such as identifier, sex, weight, age, origin, date of hospitalization, and whether or not he/she was given an antibiotic before Day 0. These characteristics are recorded during the time of admission and do not change during hospitalization. The second category includes dynamic (the monitoring) features related to patient vital signs and laboratory tests that can be recorded one or more times per day during hospitalization. Examples of these features that have been recorded one time include infection date, infection site, and examination type, which are mostly collected before giving the treatment. Examples of features registered along all the hospitalization days and which are related to patients sensitivity to treatment include SOFA score, ventilation, vasopressor usage, and given treatment dose.

2.3.2. Handling missing values

Patients data are manually entered into the e-CRF by trained personnel. The APROCCHSS cohort has a low rate of missing data, but in order to effectively handle the missing cases, we have implemented two approaches. If the missing value is associated with a temporal feature, such as a measurement taken at a specific time point, we use the last recorded value as a substitute. For example, if the "GLYCEMIC" (level of Blood Glucose) feature at Day 5 is missing, it will be replaced with the value recorded at Day 4. On the other hand, if the missing value is associated with a non-temporal feature, we replace it with -1 as a remarkable constant value. For instance, for the SITINF feature, referring to the infection location, its actual observation value ranges between 0 and 8. If its value is missing then we replace it with -1 as its actual observation value cannot be -1. By doing so, the absence of infection will be taken into account in the model.

2.3.3. Data enrichment

Data enrichment in the medical field refers to the process of creating new variables based on existing ones in order to further describe the data and improve the accuracy of prediction systems. It helps in identifying patterns and relationships in the data that were not previously visible. According to the APHP medical specialists' indications, we created the variable AR_INF_Type that indicates the origin of infection and which was derived from the DATINF (diagnossis date) and DATHOSP (Hospital admission date) variables using the following formula:

$$\text{AR\_INF\_Type} = \begin{cases} 1, & \text{if } (\min(\text{DATINF}, \text{DATHOSP})) \leq 2 \\ 2, & \text{otherwise} \end{cases} \quad (1)$$

where 1 refers to community-acquired (i.e., infection is acquired outside the hospital) and 2 refers to hospital acquired. Also, the cortisol variable values are being corrected using a dataset provided by medical experts, which contains accurate values for this feature. This correction process is performed by merging the corrected cortisol dataset with APROCCHSS and then replacing the original cortisol values with the corrected ones. This ensures that our prediction models use the most accurate and reliable data available.

Knowing whether a patient took an antibiotic before being admitted to the hospital is crucial for accurate diagnosis and treatment. This information is important not only for the machine learning model but also for the medical staff's decision making regarding appropriate dosages during the patient's hospital stay. In order to capture this information, a new feature named ANTIBIOTIC needs to be created. A value of 1 of this feature indicates that the patient received an antibiotic and a value of 0 indicates that he/she did not. This feature creation is based on the following criteria:

$$\text{ANTIBIOTIC} = \begin{cases} 1, & \text{if } ((\text{DADEBTT\_J0} \geq \text{-7}) \text{AND} (\text{DADEBTT\_J0} < 0)) \\ & \text{OR}((\text{DAFINTT\_J0} \geq \text{-7}) \text{AND} (\text{DAFINTT\_J0} < 0)) \\ & \text{OR}((\text{DADEBTT\_J0} \leq \text{-7}) \text{AND} (\text{DAFINTT\_J0} \geq \text{-7}))) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The variables ADEBTT_J0 and DAFINTT_J0 refer to the start and end dates of antibiotic use prior to hospital admission (Day 0), respectively. The new variable "ANTIBIOTIC" contains data on 550 patients who took antibiotics before hospital admission and 690 patients who did not.

2.4. Data transformation

For data transformation, we utilized two methods. The first is based on the use of an ordinal encoding method for features related to infection, where a value of 1 was assigned to non-null feature values and 0 for null values. Date features were also transformed to numeric values based on Day 0. The second method concerns the use of one-hot-encoding implemented through the Pandas library in Python[5]. For example, the SEX feature was converted into two new features, SEX1 and SEX2. SEX1 was filled with a value of 1 if the patient was a male and SEX2 was filled with a value of 0 if the patient was a female. One-hot-encoding was applied to the ORIGIN, SEX, DATINF, SITINF, ID1191S12V11, EXAMINF CHOICE,

[5] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

ID1191S12V13_CHOICE, ID1191S12V12, PREBROINF, and IGSII_ADM_TYP features presented in Table 1. As a result, the number of features increased from 24 to 127 after one-hot-encoding was applied.

2.5. Data preparation

After being processed, data needs to be properly prepared before using machine learning models. The purpose of data preparation is to create a training model that accurately predicts the sensitivity of new patients to corticotherapy. Data preparation includes: labeling data, splitting data, balancing class distribution, and scaling features.

2.5.1. Data labeling

After enrolling a patient in the study on Day 0, he/she begin receiving corticosteroid treatment or a placebo every 4 to 6 hours while monitoring a series of features that indicate the patient's progress. Each patient is monitored for 90 days and feature values are recorded daily. The APHP medical experts have established clear criteria for determining whether a patient will respond to corticotherapy or not [13, 31]. Specifically, patients are classified as cortico-sensitive (i.e., responders) if all of the following four criteria are met after 14 days of treatment:

- The patient did not die,
- The vasopressor treatment is absent for at least 24 hours,
- The patient is free from mechanical ventilation for a least 24 hours,
- The SOFA score is less than 6.

If the criteria are not met, the treatment response is considered negative, meaning the patient is cortico-resistant or a non-responder. Therefore, data is labeled as 1 or 0, indicating whether the patient responded or did not respond to the treatment on Day 14. Finally, in order to maintain the integrity of our data, and with respect to the medical experts guidelines, patients who did not align with the rule mentioned above were removed from the cohort; leading to 1234 patients as previously mentioned. Table 2 highlights the distribution of patients in the APROCCHSS cohort.

2.5.2. Data splitting

To analyze the performance of our machine learning models, the data need to be split between training and testing. In our setup, we used 80% of data to fit and train our models and optimize their parameters. After training, we used the remaining 20% of the data for testing. This applied when using the same cohort for training and testing. The training and testing were performed on specific cohorts as presented in Table 3 with a thoroughly defined experimental protocol that will be detailed in Section 2.7.3.

2.5.3. Class balancing

The training set contains more cortico-resistant patients (62%) than cortico-sensitive patients (38%). In order to achieve the best accuracy for both classes, it was necessary to balance the training data to have equal amounts of cortico-resistant and cortico-sensitive patients. To do so, we have applied the Synthetic Minority Over-sampling Technique (SMOTE) [34].

2.5.4. Feature scaling

The final step in data preparation is feature scaling. As the cohort's features have different units of measurement and different ranges of values, we have converted all feature values to the same scale of [0,1]. This process is important as it can help prevent bias and improve the performance of the machine learning models.

2.6. Deployment of the signature

This section describes the deployment of a web service for predicting cortico-sensitivity or resistance in sepsis patients. The deployment was accomplished using Flask, a Python-based web service framework, and hosted on the Assistance Publique – Hôpitaux de Paris information system. The web service takes patient data as input and provides a prediction of cortico-sensitivity or resistance as output. Figure 4 presents the deployment phase. In practice, the deployment of the sepsis prediction model, which is based on Logistic Regression with the SMOTE variant and Feat_Imp_Default configuration is achieved through the use of the APHP client application CleanWeb. CleanWeb communicates with the Flask web service by sending patient data through an HTTP request. The preprocessing script within the Flask web service prepares the data and applies the signature, which then generates a prediction of cortico-sensitivity or resistance. The output is returned to the CleanWeb client application, enabling healthcare professionals at the APHP to easily access the predictions generated by the signature. This integration of the prediction model into the APHP information system provides a seamless and efficient process for accessing the predictions. As previously mentioned, the signature will allow the APHP health specialists to stratify the patient groups during the randomization process. The prediction algorithm is considered as one of the biomarkers in this process, and thus will be integrated in future tests at the patient's bedside to reinforce the other biological tests.

It is to be noted that the selection of Logistic Regression with the SMOTE variant and Feat_Imp_Default will be explained later in the results section (Section 3).

2.7. Experimental setup

In this section, we will present a comprehensive description of the experimental setup for our approach to detect Corticosteroid sensitivity. This includes information on the considered cohorts, the machine learning algorithms utilized to predict a patient's susceptibility to corticotherapy, the experimental protocol followed, and the performance evaluation metrics that encompass both metrics related to the evaluation of machine learning models, as well as metrics used to analyze the treatment's effects at both the individual and population levels.

Table 2: Distribution of patients in APROCCHSS and RECORDS

Cohorts	APROCCHSS				RECORDS-OBSERVATIONAL			
Group	Features	Sensitive	Resistant	Total	Features	Sensitive	Resistant	Total
Corticosteroid	5645	233	379	612	21388	235	311	546
Placebo		213	409	622		81	120	201
Total	5645	446	788	1234	21388	316	431	747
Characteristic	APROCCHSS randomized controlled trial				Patients affected by COVID-19			

Table 3: Experimental protocol used on APROCCHSS and RECORDS

	Features	Training		Testing	
		APROCCHSS	RECORDS	APROCCHSS	RECORDS
Mod 1	Day 0	X		X	
Mod 2	Day 0 & Day 1	X		X	
Mod 3	Day 0 & Day 1 & Day 2 & Diff(Day 2, Day 1)	X		X	
Mod 4	Day 0		X		X
Mod 5	Day 0	X			X

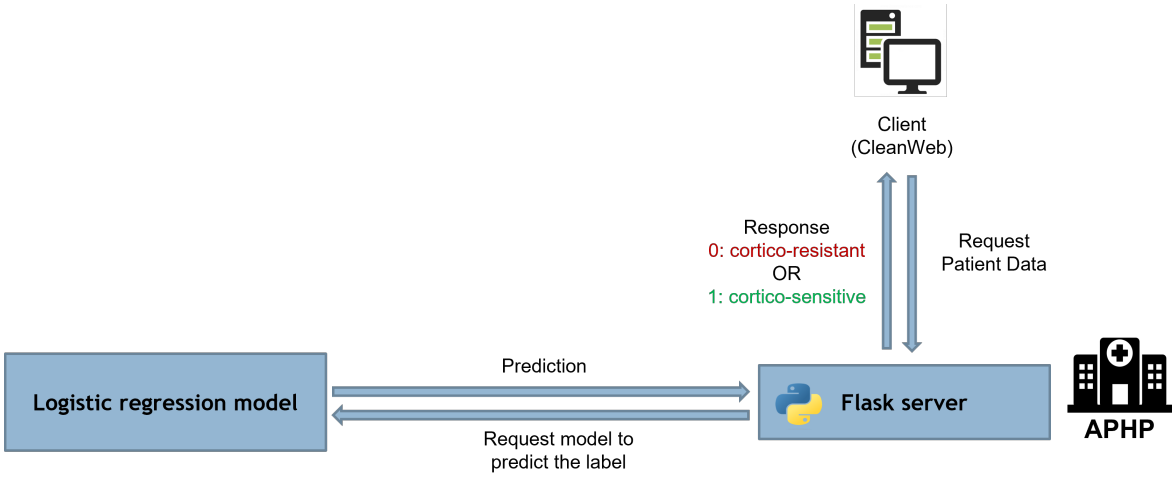

Figure 4: Deployment mode of our system

2.7.1. Considered cohorts

New patients are recruited to form the RECORDS randomized clinical trial (RCT) study during the RECORDS project which is a PIA (Investments for the Future program) project. The targeted size of the cohort at the end of the 5-year project is 1800 patients recruited in 25 intensive care units in France. Sequential intermediate analyses will occur every 500 patients to identify relevant predictive biomarkers/signatures.

During the initial period of the project, an observational study was conducted, which consisted in collecting the same clinical data as in the running period but without randomization (i.e., having treated and placebo groups). It is worth mentioning that the observational cohort coincided with the period when the COVID-19 pandemic was the highest. The CS therapy was absent in the first period of the pandemics whereas it was highly recommended in the second half. This allowed us to make a parallel with non-treated and treated patients and test the previously learned signature from the APROCCHSS cohort.

Several processing steps were applied to the RECORDS observational cohort in order to use the same parameters in APROCCHSS cohort. Therefore, we identified a mapping between the two cohorts with some modifications for multiple features. This observational cohort is characterized by the fact that most patients were affected by COVID-19. Table 2 shows the breakdown of sensitive and resistant patients to corticotherapy as well as the number of initial variables. RECORDS contains 747 sepsis patients. Unlike the APROCCHSS cohort, after reviewing RECORDS data by the medical experts, the number of patients remains unchanged. This study will only include patients who received corticotherapy, a total of 546 patients. The remaining 201 patients who were treated with placebo will be discarded. Figure 5 presents the flowchart that shows the screening process of the RECORDS cohort's patients.

As detailed in Section 2.2.2, our models will be tested on the APROCCHSS cohort which resulted from a randomized controlled trial. The objective behind using two different cohorts in our study is to evaluate the generalization of the prediction model in terms of detecting CS responsiveness when confronted to different data characteristics (i.e., COVID-19 effect).

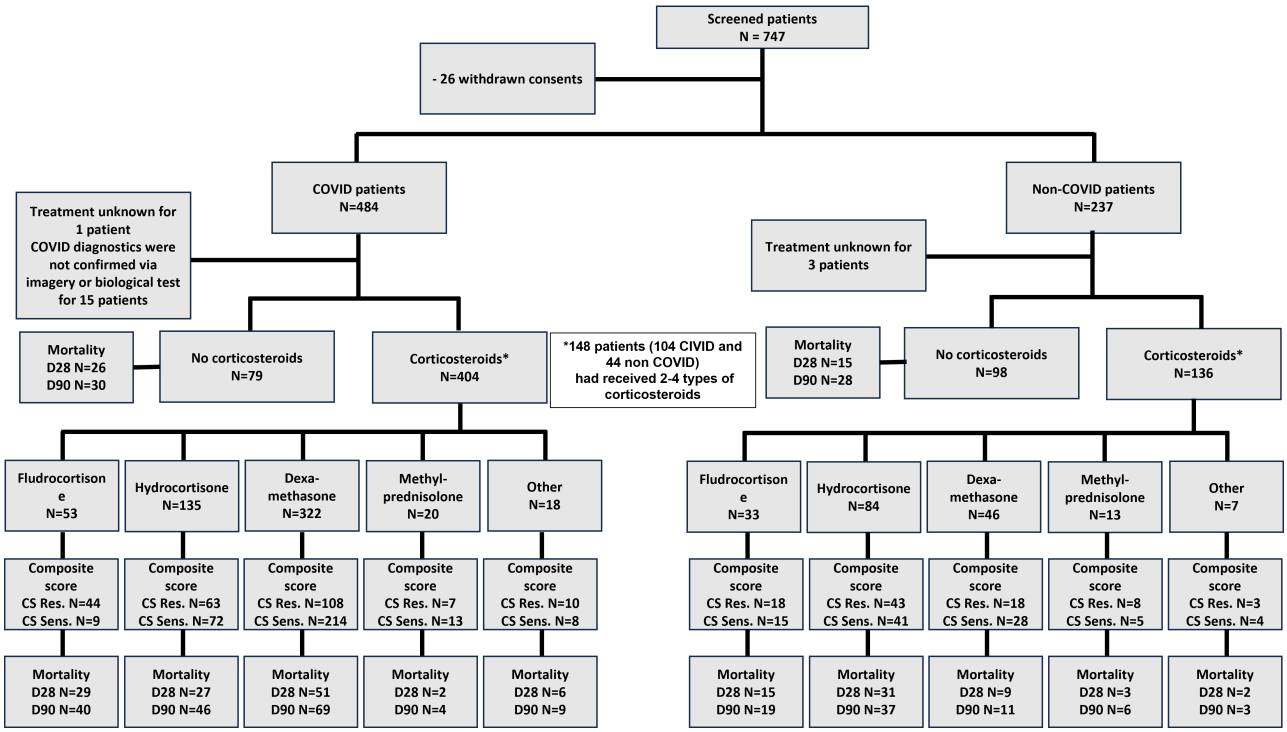

Figure 5: Patient RECORDS cohort screening process flowchart

2.7.2. *Machine learning models*

In our conducted experiments, we explored a range of machine learning algorithms to choose the best-performing one in detecting patient CS sensitivity. Particularly, we investigated the performance of the following classifiers: Logistic Regression (LR) [35], Gradient Boosting (GB) [36], Random Forest (RF) [37], Support Vector Machine (SVM) [38], Stochastic Gradient Descent (SGD) [39], and Decision Tree (DT) [40].

For each model, we set two test configurations: default and optimized. The default version tests the performance of each classifier with its default hyperparameters defined in the scikit-learn library [41]. The optimized version, on the other hand, aims to improve the efficiency of the default variant through two strategies: hyperparameter tuning and forward feature selection. In the hyperparameter tuning strategy, we used the RandomizedSearchCV technique [6] to select the best combination of values for the model's hyperparameters.

The forward feature selection strategy[7], on the other hand, aims to identify the key features that contribute the most to the model's performance. In our experiments, both the hyperparameter tuning and forward feature selection strategies were applied separately and in combination to improve the performance of the models. Experiments were run in the APHP secure environment, including the secure access to the APROCCHSS and RECORDS cohorts. We used Jupyter framework with a 64-bit Linux operating system of 188 GB RAM.

[6] `https://scikit-learn.org/`

[7] `https://scikit-learn.org/stable/modules/feature_selection.html`

2.7.3. *Experimental protocol*

Our experimental protocol is divided into two stages. The first stage focuses on learning an initial signature that aims to answer the question of the efficacy of machine learning in distinguishing sepsis responders and non-responders to corticosteroid treatment. The second stage is devoted to exploring the impact of using additional features collected over time on the accuracy of the model, as well as to investigate the generalizability of the model when applied to different patient groups belonging to the different two cohorts: APROCCHSS and RECORDS. The two stages are outlined as follows:

- Experiment 1: We evaluated the performance of various machine learning models using Day 0 features from the APROCCHSS cohort. The models were tested on four different preprocessed data versions: (i) "Original" (i.e., without scaling or balancing), (ii) "Scaled", (iii) "balanced" (SMOTE), and (iv) "balanced and scaled" (SMOTE/Scaled). In addition, we configured each model into four variants: (i) "default" version which represents the default version of the model, (ii) the "Hyper-Tuning" version which shows the version of the model after running the hyperparameter tuning process, (iii) the "Feat_Imp_Tunning" version with forward feature selection and hyperparameter tuning, and (iv) the "Feat_Imp_Default" version which specifies the default version of the model after performing forward feature selection but no hyperparameter tuning. Based on the obtained experimental results, we selected the best-performing model, referred to as "BestMod". Conducting this experiment will allow us answer **RQ1**, mentioned in

Section 1, which is related to the capacity of the machine learning models to recognize sepsis responders and non-responders to corticosteroid treatment.

- Experiment 2: The goal of this experiment is to test the performance of the best model returned by the first experiment, "BestMod", under different conditions and settings. Specifically, we implemented "BestMod" in five model variants (Mod1 to Mod5), where each model takes as input a different set of features or samples in training and/or testing. The aim is to evaluate the gain/loss in performance when incorporating additional features known at Day 1 and/or Day 2, or when applying "BestMod" to different patients' characteristics (APROCCHSS vs RECORDS). Table 3 presents a list of models and the feature sets used in our protocol, which will help us understand the impact of these factors on model performance. As seen from Table 3, Mod1, Mod4, and Mod5 take the same set of features corresponding to Day 0. They are trained either on different cohorts (Mod1 and Mod5 trained with APROCCHSS and Mod4 trained with RECORDS) or on a different testing cohort (Mod1 tested with APROCCHSS and Mod4 and Mod5 tested with RECORDS). Mod2 takes the feature sets corresponding to Day 0 and Day 1. Mod3 uses features ranging from Day 0 to Day 2 with additional ones reflecting the difference between the values of features of Day 2 and those of Day 1. Table 4 summarizes the list of features used in each experimental setting. Conducting this experiment will allow us answer **RQ2** and **RQ4**, mentioned in Section 1, which are tied to the influence of the considered features on the obtained results, and to the model's generalization when it is learned on different cohorts, respectively.

2.7.4. Applied statistical framework

Our experimental study involves algorithms with a non-deterministic nature, meaning they may produce varying results across multiple runs. To address this stochastic element of the results, the utilization of statistical testing becomes necessary. Given that we are comparing more than two algorithms, we have opted to employ the Kruskal-Wallis test [42]. This test serves as a non-parametric alternative to the ANOVA test, eliminating the need to assess whether the cohort follows a normal distribution. We conducted 30 experimental runs and analyzed the statistical differences in the results with a confidence level of 95% ($\alpha = 0.05$). The Kruskal-Wallis test evaluates the null hypothesis H0, which posits that the outcomes of all algorithms are samples from continuous distributions with equal medians. It is juxtaposed with the alternative hypothesis, H1, suggesting that they are not. The p-value from the Kruskal-Wallis test represents the probability of rejecting the true null hypothesis H0 (a type I error). If the p-value is less than or equal to $\alpha(\leq 0.05)$, we accept H1 and reject H0. Conversely, a p-value greater than $\alpha(> 0.05)$ leads to the opposite conclusion. To conduct this test, we employed the Python routine kruskal(). However, one limitation of this test is its inability to identify a single algorithm that may significantly differ from the others, while the rest remain statistically similar. To address this issue, we incorporated another Python function, namely posthoc_dunn(), to perform additional pairwise comparisons, enabling the detection of statistical differences between algorithms in a one-versus-one manner.

2.7.5. Performance assessment

In this section, we will introduce the metrics adopted to evaluate the performance of our Corticosteroid sensitivity detection approach. This includes the metrics used to measure the treatment's effect on individual patients and the overall population, as well as the metrics used to evaluate the performance of the machine learning models used in our protocol.

Metrics to evaluate treatment effectiveness. Individual Treatment Effect (ITE) and Average Treatment Effect (ATE) are two important metrics used to evaluate the effectiveness of treatment in randomized controlled trials. To compute ITE, we compare the outcome of patients who received corticosteroid treatment (T=1) with the outcome of patients who did not receive the treatment (T=0). We calculate predictions for patients who received corticotherapy and those who received a placebo, and subsequently, the computed values of individual treatment effect represent the disparity between these two predictions. ITE can be used to identify patients who are most likely to benefit from a treatment and can help to personalize treatment decisions [43]. ATE, on the other hand, measures the average effect of a treatment across all patients in the study. It can be used to estimate the overall benefit of the treatment and can help to determine whether a treatment is effective for a population as a whole. ITE is defined as follows [43]:

$$ITE = Y(X = 1) - Y(X = 0) = Y(T = 1) - Y(T = 0) \quad (3)$$

where, $Y(T = 1)$ denotes the outcome value for the patients who took the corticosteroid treatment (i.e., $X = 1$), and $Y(T = 0)$ denotes the outcome of not receiving the CS treatment referring to those who took the placebo (i.e., $X = 0$). ATE is defined as follows [43]:

$$ATE = E(Y(T = 1)) - E(Y(T = 0)) \quad (4)$$

where $E(Y(T = 1))$ is the average outcome of receiving the CS treatment, and $E(Y(T = 0))$ is the average outcome of not receiving the CS treatment; referring to those who took the placebo. Additionally, we used the Conditional Average Treatment Effect (CATE) which measures the ATE within a certain subgroup or condition. It is calculated by taking the average treatment effect within a specific subgroup or condition, rather than across the entire population.

The Absolute Risk Reduction (ARR) [44] and the Number Needed to Treat (NNT) [45] are two additional metrics that we have used to evaluate the effectiveness of the CS treatment. ARR is the difference in the incidence of an event between the treated group and the control group. It is a measure of the absolute benefit of the treatment, and is calculated by subtracting the incidence of the event in the control group from the incidence of the event in the treatment group. The event for

Table 4: Set of features used in each experimental setting on APROCCHSS and RECORDS

	Features	Total
Mod 1, Mod 4 and Mod 5	ORIGIN, SEX, IGSII_ADM_TYP, AR_Cortico, PATWGHT, AGE, SOFA_INC2, KNAUS_J0, MACCABE_J0, SOFA_ADM, ID1191S12V3, IGS3_ADM, GLYCEMIE, LACTATES_J0, Cort_DiffMax, Cort_Rep, AR_INF_Type, SITINFx, ID1191S12V11_SITINFx, DATINF_SITINFx, ID1191S12V12_SITINFx, ID1191S12V13_SITINFx, ID1191S12V13_CHOICE1Y_SITINFx, EXAMINF_CHOICEx, PREBROINFx	24
Mod 2	Same to Mod 1 + DOSMAXAD_J1[1], DOSMAXDOB_J1[2], DOSMAXDOP_J1[3], DOSMAXNA_J1[4], DOSMINAD_J1[5], DOSMINDOB_J1[6], DOSMINDOP_J1, DOSMINNA_J1[9], DOSTERL_J1,[10]DOSVASOP_J1,[11]'SOFA_J1', VENTIL_J1, VIVANT_J1	37
Mod 3	Same to Mod 2 + DOSMAXAD_J2[1], DOSMAXDOB_J2[2], DOSMAXDOP_J2[3], DOSMAXNA_J2[4], DOSMINAD_J2[5], DOSMINDOB_J2[6], DOSMINDOP_J2, DOSMINNA_J2[9], DOSTERL_J2,[10]DOSVASOP_J2,[11]SOFA_J2, VENTIL_J2[7], VIVANT_J2[8], DOSMAXAD_Diff_1_2, DOSMAXDOB_Diff_1_2, DOSMAXDOP_Diff_1_2, DOSMAXNA_Diff_1_2, DOSMINAD_Diff_1_2, DOSMINDOB_Diff_1_2, DOSMINDOP_Diff_1_2, DOSMINNA_Diff_1_2, DOSTERL_Diff_1_2, SOFA_Diff_1_2	60

[1] Max dose Adrenaline (mg/h), [2] Dose maxi dobutamine (µg/kg/min) [3] Max dose dopamine (µg/kg/min)
[4] Max dose Noradrenaline (mg/h) [5] Minimum dose Adrenaline (mg/h) [6] Minimum dose dobutamine (µg/kg/min)
[7] Assisted ventilation [8] Indicates whether or not the patient is alive [9] Minimum Noradrenaline dose (mg/h)
[10] Terlipressin (IU/h) [11] Vasopressin (IU/h)

ARR in our study is a significant clinical outcome such as a patient's positive response to corticosteroid treatment. ARR quantifies the difference in the incidence of this event between the treated group (corticosteroid treatment) and the control group (placebo). ARR can also be calculated as the negative of ATE, which aligns with its definition. The formula for ARR is [44]:

$$ARR = P(\text{event in control group}) - P(\text{event in treatment group}) \quad (5)$$

or simply:

$$ARR = -ATE \quad (6)$$

On the other hand, NNT is a measure of the number of patients that need to be treated in order to prevent one event. NNT is calculated as the reciprocal of ARR. The formula for NNT is [45]:

$$NNT = \frac{1}{ARR} \quad (7)$$

Both ARR and NNT are used to evaluate the effectiveness of a treatment, but they provide different perspectives on the treatment's performance. These performance assessment metrics will allow us respond to **RQ3**, mentioned in Section 1, which is tied to the effectiveness of the treatment.

Other evaluation metrics and metrics interpretation. To evaluate the performance of our machine learning models under different configurations and settings, we used the following common metrics: accuracy (Ac), precision (Pr), recall (Re), F1 score, and the Area Under the Curve (AUC). Referring to [46], in case where AUC > 0.9 then it indicates a highly accurate model, if $0.7 \leq$ AUC < 0.9 then this indicates a moderately accurate model, and if AUC < 0.7 then this indicates a low accurate model.

As previously presented, we have used metrics to evaluate treatment effectiveness. The ITE values can be positive, negative or zero. In our study, the outcome is binary. Therefore, the ITE values are reported as 1, 0, or -1, rather than a specific numerical value. A value of -1 for ITE indicates that the treatment had a beneficial effect on the individual patient. A value of 0 indicates that the treatment had no effect on the individual patient, and a value of 1 indicates that the treatment had a detrimental effect on the individual patient.

Similarly, the values of ATE can be positive, negative or zero. A negative ATE value indicates that the treatment had a beneficial effect on the population as a whole, a positive ATE value indicates that the treatment had a detrimental effect on the population as a whole, and a zero ATE value indicates that the treatment had no effect on the population as a whole. It is important to note that even if the ATE is positive, it does not mean that the treatment would work well for all patients. Some individual patients may not benefit from the treatment, or even may experience negative effects. The ATE is an average of all patients in the study and it might not reflect the individual patient experience.

The values of ARR can be positive, negative, or zero. A negative ARR value indicates that the incidence of the event is higher in the treatment group than in the control group, which means that the treatment is not effective in reducing the risk of the event and can be harmful. A positive ARR value indicates that the incidence of the event is lower in the treatment group than in the control group, which means that the treatment is effective in reducing the risk of the event. A zero ARR value indicates that there is no difference in the incidence of the event between the treatment group and the control group, which

means that the treatment is not effective in reducing the risk of the event.

A positive NNT with a low value indicates that the treatment is very effective in reducing the risk of the event and that a small number of patients need to be treated to prevent one event. A positive NNT with a high value indicates that the treatment is less effective in reducing the risk of the event and that a large number of patients need to be treated to prevent one event. It is to be noted that a negative NNT reflects the number of patients needed to harm. An NNT value of infinity means that the treatment is not effective in reducing the risk of the event at all and that no matter how many patients are treated, the event will not be prevented.

3. Results

In this section, we present the obtained results with respect to our defined experimental protocol.

3.1. Experiment 1: empirical results

The results obtained from different machine learning models with various settings are presented in Table 5. From the table, it can be observed that:

- Most of the models in our study demonstrated that when the most relevant features were selected in combination with hyperparameter tuning (i.e., Feat_Imp_Tuning), they achieved the highest level of AUC in their predictions. This is likely because feature selection and hyperparameter tuning are both important steps in optimizing the performance of machine learning models. Feature selection helps to identify the most important variables that have the greatest impact on the outcome, while hyperparameter tuning helps to optimize the performance of the model by selecting the best set of parameters for a given dataset. When used together, these two techniques can help improve the performance of the model and reduce overfitting.

  In the results presented in Table 5, the best AUC values were obtained by models that used this combination of feature selection and hyperparameter tuning. For example, Gradient Boosting, and Support Vector Machine achieved an AUC value of 71%, and Stochastic Gradient Descent and Decision Tree achieved an AUC of 70%. Random Forest registered 71% with the Hyper_Tuning setting. However, unexpectedly, the best AUC value was registered for Logistic Regression with a value of 72% when the most relevant features were selected within the default version (i.e., Feat_Imp_Default). Figure 6 shows a better visualization of these results.

- From Table 5, it can be observed that in most cases (all models except Stochastic Gradient Descent), the models performed better when using the SMOTE and SMOTE/Scaled variants compared to the other configurations (i.e., Original and Scaled). The SMOTE and SMOTE/Scaled variants improved the performance of the machine learning models. This is likely because the SMOTE technique helps to balance the dataset by creating synthetic samples of the minority class, which in turn increases the diversity of the data and provides a more representative sample of the sepsis population. The better the data quality, the more informative the dataset becomes, which in turn makes the machine learning model more capable of performing its prediction task.

- To supplement the above, by analyzing the performance metrics of the different models, the accuracy, precision, recall, and F1 metrics, we can also conclude that the best model which achieved the best results is Logistic Regression. Particularly, Logistic Regression registered 72%, 75%, 72%, and 73% for the accuracy, precision, recall, and F1-score, respectively, in comparison to Random Forest with 74%, 71%, 71%, and 71%, Gradient Boosting with 71%, 68%, 69%, and 68%, Support vector machine with 72%, 74%, 72%, and 72%, Stochastic Gradient Descent with 72%, 68%, 68%, and 68% and finally, Decision Tree with 69%, 66%, 68%, and 67% for the accuracy, precision, recall, and F1-score, respectively.

- Based on the data provided in Table 5, it is clear that the computed p-values fall below the significance level $\alpha = 0.05$. For each classifier, we calculated the p-values for various input data configurations, including original, scaled, SMOTE, and SMOTE Scaled. These p-values range from 1.267057943824278E-25 to 1.2670579438242684E-25. Furthermore, the p-values for each classifier consistently remain at the value of 1.267057943824278E-25. Consequently, we can conduct a statistical comparison among the algorithms under consideration. Table 6 indicates that the logistic regression algorithm exhibits statistical differences compared to Gradient Boosting, Support Vector Machine, Stochastic Gradient Descent, and Decision Tree. Notably, logistic regression demonstrates superior performance when compared to these algorithms, especially Gradient Boosting and Support Vector Machine. This observation aligns with the results obtained from the experiments.

- In conclusion, Logistic Regression is considered to be the best model for our study based on the analysis of the accuracy, precision, recall, F1 metric, and AUC. With the SMOTE variant and Feat_Imp_Default configuration, Logistic Regression achieved the highest values among all the models and hence selected as the "BestMod" to be used for the rest of our experimental protocol.

3.2. Experiment 2: empirical results and discussion

The results obtained from applying the Logistic Regression model with the SMOTE setting on the previously defined different variant models (Table 3) are presented in Table 7. The following observations can be made:

- Similarly to the results obtained in Experiment 1, in most configurations, the models achieved the highest AUC when using the combination of selecting the

most important features with hyperparameter tuning (i.e., Feat_Imp_Tuning). Even with more features (Mod 2 and Mod 3) and by considering different cohorts, this configuration proves its importance and impact on the efficiency of the Logistic Regression model. From Table 7, Mod 1 (i.e., the best model) achieved an AUC value of 72%, Mod 2 achieved an AUC value of 75%, Mod 3 achieved an AUC value of 76%, and Mod 4 achieved an AUC value of 69%.

- From Table 7, it can be observed that the best AUC score was registered for Mod 3 with the value of 76% when the most relevant features are selected from Day 0, Day 1, Day 2, and the difference between Day 2 and Day 1 (see Table 3) using the same Feat_Imp_Tuning configuration. This confirms the fact that training the model with more features increases its performance.

- In addition to the AUC value, by evaluating the remaining performance metrics (i.e., accuracy, precision, recall, and F1 score) of the different models' variants, we can also conclude that Mod 3 is the best variant model that achieved the best results. Especially, the accuracy, precision, recall, and F1 score are set to 75%, 73%, 76%, and 73% with Mod 3 in comparison to Mod 1 with 72%, 75%, 72%, and 73%, Mod 2 with 74%, 73%, 76%, and 73%, Mod 4 with 69%, 69%, 70%, and 69%, and Mod 5 with 52%, 56%, 55%, and 50% for the accuracy, precision, recall, and F1 score, respectively.

Table 5: Evaluation of different machine learning models on APROCCHSS using different configurations in terms of mean values generated across 30 runs

Classifier	Model	Original					Scaled					Smote					Smote/scaled					p-value	p-value / model	p-value Global
		Auc.	Ac.	Pr.	Re.	F1	Auc.	Ac.	Pr.	Re.	F1	Auc.	Ac.	Pr.	Re.	F1	Auc.	Ac.	Pr.	Re.	F1			
Logistic Regression	Default	0.67	0.71	0.67	0.67	0.67	0.62	0.65	0.67	0.65	0.66	0.67	0.66	0.71	0.66	0.67	0.61	0.63	0.60	0.61	0.60	1.267057943824278 E-25	1.267057943824278 E-25	8.749844824 E-37
	Hyper-Tuning	0.67	0.72	0.68	0.68	0.68	0.63	0.67	0.68	0.67	0.67	0.65	0.63	0.69	0.63	0.64	0.64	0.66	0.63	0.64	0.63	1.267057943824278 E-25		
	Feat_Imp_Tuning	0.71	0.72	0.68	0.69	0.68	0.67	0.72	0.71	0.72	0.71	0.72	0.71	0.69	0.72	0.97	0.65	0.66	0.63	0.65	0.64	1.267057943824278 E-25		
	Feat_Imp_Default	0.69	0.73	0.70	0.70	0.70	0.64	0.67	0.68	0.67	0.68	**0.72**	**0.72**	**0.75**	**0.72**	**0.73**	0.63	0.64	0.62	0.63	0.62	1.267057943824278 E-25		
Gradient Boosting	Default	0.62	0.68	0.63	0.62	0.63	0.64	0.68	0.64	0.64	0.64	0.66	0.69	0.65	0.66	0.66	0.62	0.68	0.63	0.62	0.63	1.2670579438242684 E-25	1.267057943824278 E-25	
	Hyper-Tuning	0.60	0.72	0.72	0.60	0.60	0.66	0.70	0.66	0.66	0.66	0.63	0.66	0.62	0.63	0.62	0.65	0.71	0.66	0.65	0.66	1.267057943824278 E-25		
	Feat_Imp_Tuning	0.62	0.73	0.72	0.62	0.62	0.68	0.72	0.68	0.68	0.68	**0.71**	**0.71**	**0.68**	**0.69**	**0.68**	0.69	0.72	0.69	0.69	0.69	1.267057943824278 E-25		
	Feat_Imp_Default	0.67	0.71	0.67	0.67	0.67	0.68	0.72	0.68	0.68	0.68	0.67	0.69	0.66	0.67	0.66	0.68	0.73	0.69	0.68	0.68	1.267057943824278 E-25		
Random Forest	Default	0.50	0.67	0.34	0.50	0.40	0.50	0.67	0.34	0.50	0.40	0.67	0.70	0.66	0.67	0.67	0.69	0.71	0.67	0.69	0.68	1.2670579438242684 E-25	1.267057943824278 E-25	
	Hyper-Tuning	0.65	0.72	0.68	0.65	0.66	0.65	0.72	0.68	0.65	0.66	**0.71**	**0.74**	**0.71**	**0.71**	**0.71**	0.66	0.71	0.66	0.66	0.66	1.2670579438242684 E-25		
	Feat_Imp_Tuning	0.68	0.76	0.73	0.68	0.70	0.68	0.76	0.73	0.68	0.70	0.70	0.72	0.69	0.70	0.70	0.69	0.72	0.69	0.69	0.69	1.2670579438242684 E-25		
	Feat_Imp_Default	0.50	0.67	0.34	0.50	0.40	0.50	0.67	0.34	0.50	0.40	0.64	0.62	0.62	0.64	0.61	0.66	0.66	0.64	0.66	0.64	1.2670579438242684 E-25		
Support Vector Machine	Default	0.64	0.65	0.68	0.67	0.67	0.63	0.69	0.67	0.66	0.66	0.68	0.71	0.68	0.67	0.68	0.65	0.69	0.69	0.69	0.69	1.267057943824278 E-25	1.267057943824278 E-25	
	Hyper-Tuning	0.69	0.73	0.69	0.69	0.69	0.63	0.67	0.67	0.66	0.66	0.66	0.69	0.69	0.67	0.68	0.67	0.70	0.70	0.68	0.69	1.267057943824278 E-25		
	Feat_Imp_Tuning	0.69	0.73	0.69	0.69	0.69	0.65	0.72	0.66	0.65	0.66	0.60	0.66	0.66	0.67	0.66	**0.71**	**0.72**	**0.74**	**0.72**	**0.72**	1.267057943824278 E-25		
	Feat_Imp_Default	0.56	0.64	0.64	0.67	0.64	0.65	0.72	0.67	0.65	0.65	0.63	0.67	0.67	0.61	0.62	0.68	0.69	0.66	0.67	0.66	1.267057943824278 E-25		
Stochastic Gradient Descent	Default	0.56	0.69	0.64	0.56	0.54	0.59	0.62	0.58	0.59	0.58	0.52	0.35	0.67	0.52	0.28	0.58	0.59	0.57	0.58	0.57	1.26705794382427 E-25	1.267057943824278 E-25	
	Hyper-Tuning	0.62	0.64	0.61	0.62	0.61	0.59	0.61	0.58	0.59	0.58	0.49	0.43	0.49	0.49	0.43	0.61	0.62	0.60	0.61	0.59	1.26705794382427 E-25		
	Feat_Imp_Tuning	**0.70**	**0.72**	**0.68**	**0.68**	**0.68**	0.59	0.60	0.58	0.59	0.58	0.67	0.61	0.65	0.67	0.61	0.62	0.67	0.62	0.62	0.62	1.26705794382427 E-25		
	Feat_Imp_Default	0.64	0.61	0.62	0.64	0.60	0.63	0.63	0.62	0.63	0.61	0.67	0.61	0.65	0.67	0.61	0.64	0.64	0.62	0.63	0.62	1.26705794382427 E-25		
Decision Tree	Default	0.58	0.63	0.58	0.58	0.58	0.59	0.63	0.59	0.59	0.59	0.56	0.58	0.55	0.56	0.55	0.56	0.58	0.55	0.56	0.55	1.2670579438242684 E-25	1.267057943824278 E-25	
	Hyper-Tuning	0.54	0.59	0.54	0.54	0.54	0.54	0.59	0.54	0.54	0.54	0.57	0.62	0.57	0.57	0.57	0.56	0.61	0.56	0.56	0.56	1.2670579438242684 E-25		
	Feat_Imp_Tuning	0.64	0.65	0.62	0.64	0.63	0.62	0.66	0.61	0.62	0.62	0.67	0.69	0.66	0.67	0.66	**0.70**	**0.69**	**0.66**	**0.68**	**0.67**	1.2670579438242684 E-25		
	Feat_Imp_Default	0.59	0.63	0.59	0.59	0.59	0.57	0.61	0.57	0.57	0.57	0.55	0.56	0.54	0.55	0.54	0.55	0.56	0.54	0.55	0.54	1.2670579438242684 E-25		

- In most cases, models performed better when using the APROCCHSS cohort and by considering additional features compared to models applied to RECORDS. Let us recall that the RECORDS cohort includes patients affected by COVID-19 (Section 2.7.1). This data property could have an influence on the model distribution and therefore impacted the prediction results.

- By comparing the results of Mod 1 and Mod 4 having the same features at Day 0 but trained with different cohorts, the application of the initial signature to the RECORDS cohort presents less accurate predictions. With the Feat_Imp_Tuning configuration, Mod 1 achieved an AUC value of 72%, however, Mod 4 achieved an AUC value of 70% with the same setup. This might be due to the COVID-19 effect that is present on the RECORDS cohort.

- In all configuration setups (i.e., default, hyperparameter tuning, Feat_Imp_Tuning, and Feat_Imp_Default), Mod 5 presents less accurate results compared to other variants. For example, the model achieved an AUC value of 55% with default and Hyper_tuning, which are considered to be among the most performing configurations. This can also be explained by the fact of including COVID-19 patients in the RECORDS cohort. Therefore, a generalization may not be applied to different sepsis causes; as other factors might influence the behaviour of the machine learning model. We conclude that viral versus bacterial sepsis may lead to different signatures.

- Based on the data provided in Table 7, it is clear that the computed p-values fall below the significance level α = 0.05. For each mod, we calculated the p-values for various configurations, including Default, Hyper_tuning, Feat Imp Tuning, and Feat Imp Default. These p-values range from 1.26705794382425E-25 to 1.2670579438242684E-25. Furthermore, the p-values for each mod consistently remain at the value of 3.334325930239449E-31. Consequently, we can conduct a statistical comparison among the algorithms under consideration. Table 8 indicates that Mod 3 exhibits statistical differences compared to Mod 4 and Mod 5. We also notice a statistical difference between Mod 1 and Mod 2 when added additional features. Notably, Mod 3 demonstrates superior performance when compared to the other experiments. This observation aligns with the results obtained from the experiments. However, we have noticed that there is no statistical difference between Mod 3 and Mod 2. This necessitates further investigation. This encourages us to analyze the added features per day in the initial stage and then enhance the model with the difference in feature values between consecutive days.

3.3. Evaluation of treatment effectiveness

Evaluating treatment effectiveness will allow us to respond to **RQ3**: What is the effect of corticotherapy on individual treated and all treated patients? To complement the analysis conducted in Experiment 1 and Experiment 2, we will analyze the effect of corticotherapy on individual treated patients and all treated patients as a whole using the treatment effectiveness metrics defined in Section 2.7.5. We will measure ITE, ATE, ARR, and NNT, for the four different Logistic Regression variants (based on SMOTE) using various feature sets (Mod 1 to Mod 4) with the four different configurations: default, Hyper_Tuning, Feat_Imp_Tuning, and Feat_Imp-Default. The results of these metrics are presented in Table 10. Mod 5 was not considered in the analysis as it was trained and tested using a separate cohort, and its results would not provide meaningful insights into the conclusions of our study.

To ensure the accuracy and reliability of our findings and evaluations, the reported results for all statistical metrics are based solely on patients who were correctly predicted by the models. The calculation of the metrics' values is based on the same sample set. To achieve this, we first identified the correctly classified patients and subsequently determined the intersection of all patients. This approach not only maintains a consistent sample size but also ensures uniformity in patient identifiers. Based on the results shown in Table 10, the ITE measure may vary considerably depending on the configuration used for the Logistic Regression. The obtained results can be interpreted as follows:

- For Mod 1, for the different configurations (i.e., default, hyperparameter tuning, Feat_Imp_Tuning, and Feat_Imp_Default), the percentage of patients with an ITE = 0, varies from 77.33% to 87.67%, indicating that it is difficult to decide whether the treatment is effective or not. In addition, the positive impact of the treatment on patients indicating ITE = -1 shows a small variation between the different settings, ranging from 9% to 15.67%. Patients having a detrimental effect when using the treatment show a relatively low percentage, varying from 2.33% to 9.33%. Figure 7 shows a better visualization of these results.

- For Mod 2 (Figure 7), for the different configurations, the percentage of patients with an ITE = 0 varies from 82% to 85.33%. In addition, the beneficial impact of the treatment on patients indicating ITE = -1 shows a slight variation between the configurations, ranging from 8.67% to 10.33%. Patients having a detrimental effect using the treatment (ITE = 1) show a relatively low percentage, varying from 4.67% to 8%.

- Figure 7 shows the results obtained from Mod 3. The percentage of patients with an ITE = 0 varies from 84.33% to 88%, indicating that it is difficult to decide whether the treatment is effective or not. The positive impact of the treatment on patients (i.e., ITE = -1) presents a low percentage, ranging from 3.33% to 8.33%, similarly to the percentage reflecting the negative effect which varies from 7% to 9.33%.

- For Mod 4 (Figure 7), for the different configurations, the percentage of patients with an ITE = 0, varies from 73%

Table 6: Dunn's test performance for experiment 1 – p-values based on mean AUC values generated across 30 runs

	Logistic regression	Gradient boosting	Random forest	Support vector machine	Stochatic Gradient Descent	Decision Tree
Logistic regression	1.0	1.8092987633461E-28	0.355944493430778	9.1472313444385E-05	1.7481136093207E-10	2.2262610723402E-18
Gradient boosting	1.8092987633461E-28	1.0	2.2262610723402E-18	1.7481136093207E-10	9.1472313444385E-05	0.355944493430778
Random forest	0.355944493430778	2.2262610723402E-18	1.0	0.355944493430778	9.1472313444385E-05	1.7481136093207E-10
Support vector machine	9.1472313444385E-05	1.7481136093207E-10	0.355944493430778	1.0	0.355944493430778	9.1472313444385E-05
Stochatic Gradient Descent	1.7481136093207E-10	9.1472313444385E-05	9.1472313444385E-05	0.355944493430778	1.0	0.355944493430778
Decision Tree	2.2262610723402E-18	0.355944493430778	1.7481136093207E-10	9.1472313444385E-05	0.355944493430778	1.0

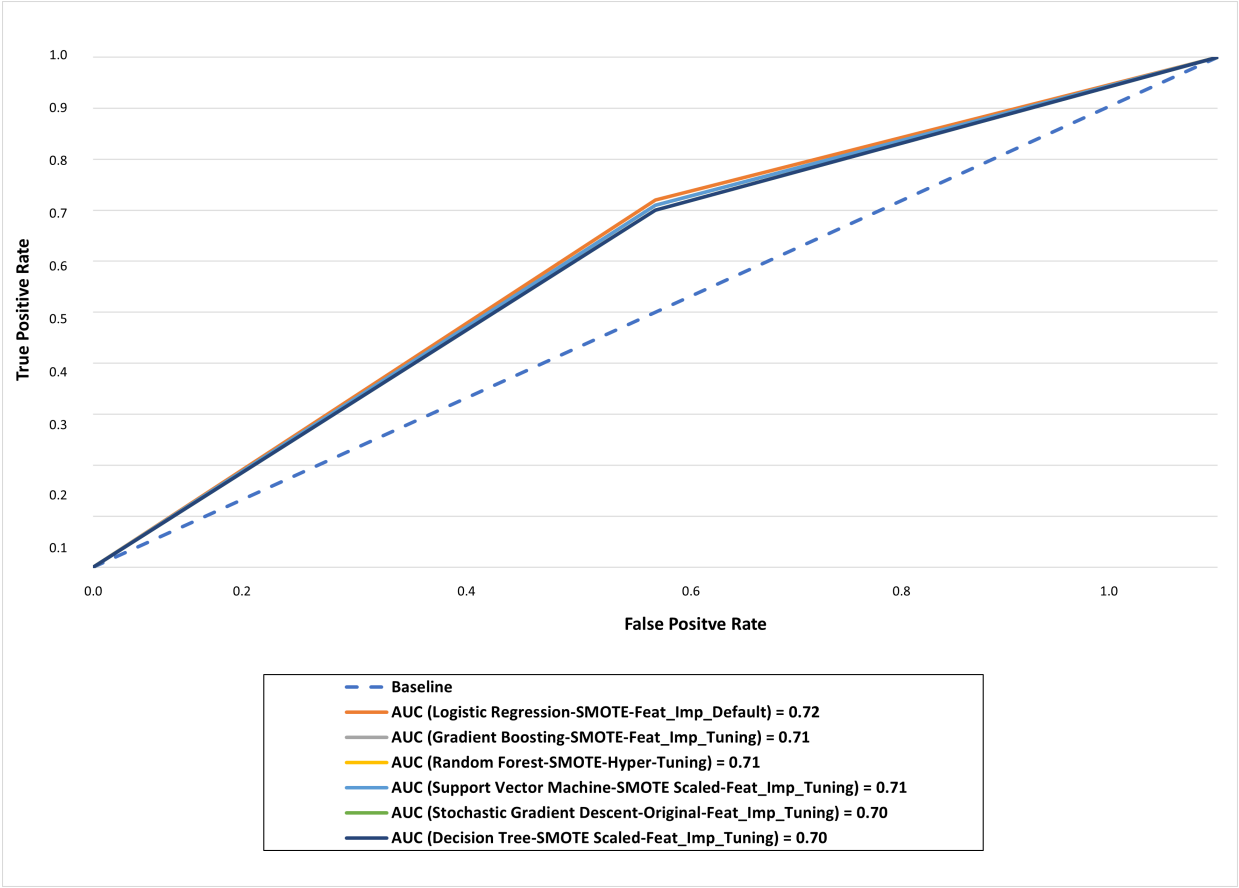

Figure 6: ROC curves pf the Best Model applied on APROCCHSS cohort

to 81%, indicating that it is difficult to decide whether the treatment is effective or not. In addition, the positive impact of the treatment on patients indicating ITE = -1 shows a small variation between the different settings, ranging from 11% to 14%. Patients having a detrimental effect when using the treatment show a relatively low percentage, varying from 8% to 13%.

- From Table 10, it can be observed that the high percentage with positive treatment effect was obtained by models (except for Mod 4) that used especially feature selection combined with either hyperparameter tuning or the default model (i.e., Feat_Imp_Tuning and Feat_Imp_Default). This confirms the outcomes found in Experiments 1 and Experiments 2 tied to the best configurations. However, for Mod 3, we can notice a significant decrease in the beneficial effect on the individual patient values (ITE = -1), observed for all configurations, compared to Mod 1 and Mod 2 recorded values. For example, for ITE = -1, Mod 3 achieved only a percentage of 3.33%, 4.67%, 8.33%, and 8.33%, in comparison to Mod 1 which achieved 9.33%, 9%, 15.67%, and 10% and to Mod 2 which achieved 10%, 8.67%, 10.33%, and 10% for the Default, Hyperparameter tuning, Feat_Imp_Tuning, and Feat_Imp_Default configurations, respectively.

As for the ATE interpretations, results may be interpreted in a similar way to ITE; as follows.

- From Table 10, for Mod 1 and Mod 2, for all configurations, the negative ATE values indicate that the treatment had a beneficial effect on the population as a whole. Mod 1 ATE values vary from -0.09 to 0 while for Mod 2 ATE values vary from -0.05 to -0.02.

- In Mod 3, both Default and Hyper_Tuning configurations exhibit positive ATE values of 0.05, suggesting potential

Table 7: Evaluation of model variants applied on APROCCHSS and RECORDS cohorts based on mean values generated across 30 runs

Mod	Default					Hyper_Tuning					Feat_Imp_Tuning					Feat_Imp_Default					p-value	p_value Gobal
	Auc.	Ac.	Pr.	Re.	F1	Auc.	Ac.	Pr.	Re.	F1	Auc.	Ac.	Pr.	Re.	F1	Auc.	Ac.	Pr.	Re.	F1		
Mod1	0.69	0.73	0.70	0.70	0.70	0.64	0.67	0.68	0.67	0.68	**0.72**	**0.72**	**0.75**	**0.72**	**0.73**	0.63	0.64	0.62	0.63	0.62	1.267057943824278 E-25	3.334325930239449 E-31
Mod 2	0.68	0.69	0.67	0.69	0.66	0.68	0.68	0.66	0.68	0.65	**0.75**	**0.74**	**0.73**	**0.76**	**0.73**	**0.75**	**0.74**	**0.73**	**0.76**	**0.73**	1.26705794382425 E-25	
Mod 3	0.69	0.68	0.67	0.69	0.67	0.65	0.65	0.64	0.66	0.63	**0.76**	**0.75**	**0.73**	**0.76**	**0.73**	0.74	0.74	0.72	0.75	0.72	1.2670579438242684 E-25	
Mod 4	0.67	0.67	0.67	0.68	0.67	0.67	0.67	0.68	0.68	0.67	**0.69**	**0.69**	**0.69**	**0.7**	**0.69**	0.68	0.65	0.69	0.69	0.69	1.267057943824278 E-25	
Mod 5	**0.54**	**0.52**	**0.56**	**0.55**	**0.5**	**0.54**	**0.52**	**0.56**	**0.55**	**0.5**	0.49	0.52	0.48	0.49	0.41	0.52	0.55	0.54	0.52	0.46	1.2670579438242684 E-25	

Table 8: Dunn's test performance for experiment 2 – p-values based on mean AUC values generated across 30 runs

	Mod 1	Mod 2	Mod 3	Mod 4	Mod 5
Mod 1	1.0	4.789919835201660E-07	0.0634364490986529	0.0634364490986529	4.789919835201660E-07
Mod 2	4.789919835201660E-07	1.0	0.0634364490986529	2.647075928862760E-15	9.468492290506880E-27
Mod 3	0.0634364490986529	0.0634364490986529	1.0	4.789919835201660E-07	2.647075928862760E-15
Mod 4	0.0634364490986529	2.647075928862760E-15	4.789919835201660E-07	1.0	0.0634364490986529
Mod 5	4.789919835201660E-07	9.468492290506880E-27	2.647075928862760E-15	0.0634364490986529	1.0

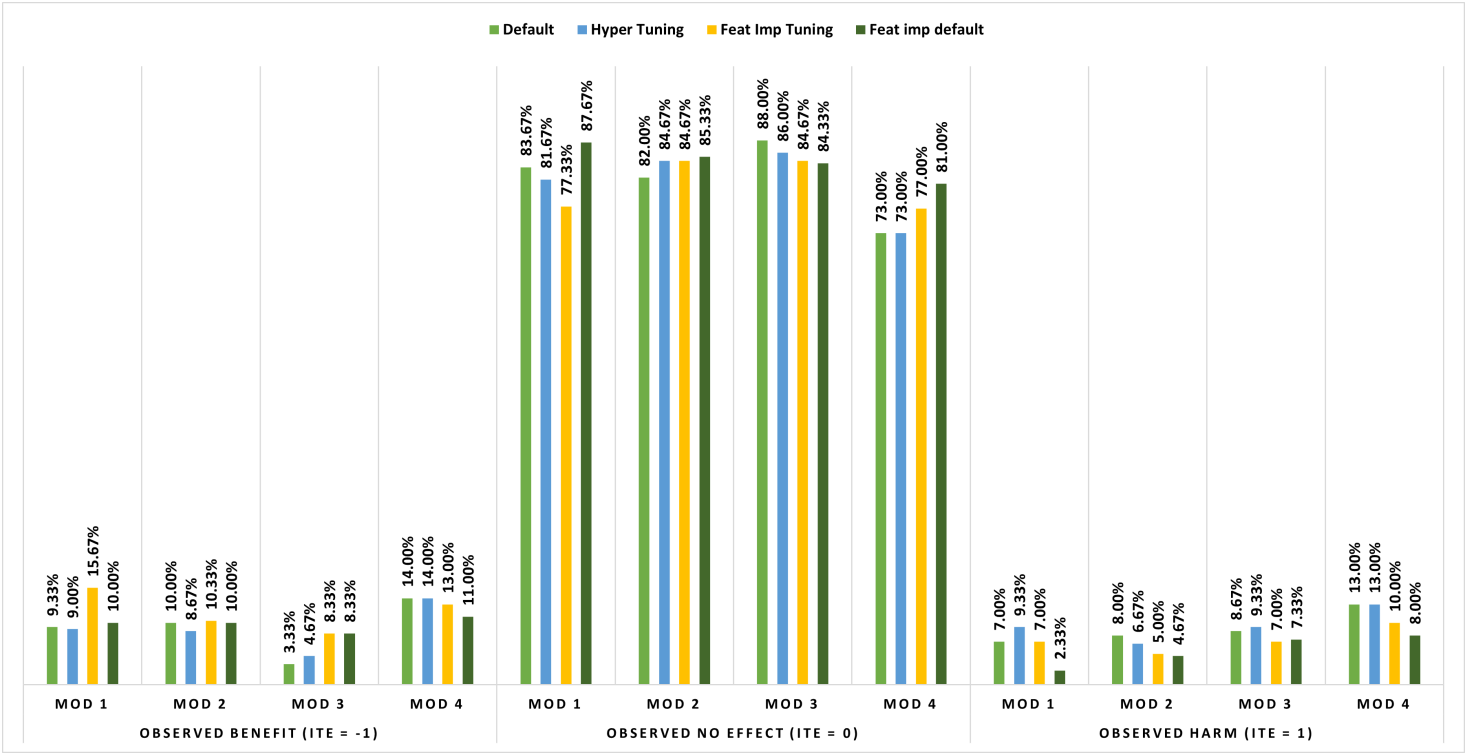

Figure 7: Percentage ITE

Table 9: Statistical analysis: APROCCHSS vs RECORDS

Cohort	Treatment / No treatment	Number of Female Patients	Number of Male Patients
APROCCHSS	Corticosteroid	187	355
	Placebo	178	395
RECORDS	Corticosteroid	136	329
	No treatment	51	121

Table 10: Statistical metrics for the different modes

		Default	Hyper_Tuning	Feat_Imp_Tuning	Feat_Imp_Default
Mod 1	Correctly classified patients	300			
	ITE	(-1, 9.33%) (0, 83.67%) (1, 7.0%)	(-1, 9.0%) (0, 81.67%) (1, 9.33%)	(-1, 15.67%) (0, 77.33%) (1, 7.0%)	(-1, 10.0%) (0, 87.67%) (1, 2.33%)
	ATE	-0.02	0	-0.09	-0.08
	ARR	0.02	0	0.09	0.08
	NNT	50	-	11.11	12.5
Mod 2	Correctly classified patients	300			
	ITE	(-1, 10.0%) (0, 82.0%) (1, 8.0%)	(-1, 8.67%) (0, 84.67%) (1, 6.67%)	(-1, 10.33%) (0, 84.67%) (1, 5.0%)	(-1, 10.0%) (0, 85.33%) (1, 4.67%)
	ATE	-0.02	-0.02	-0.05	-0.05
	ARR	0.02	0.02	0.05	0.05
	NNT	50	50	20	20
Mod 3	Correctly classified patients	300			
	ITE	(-1, 3.33%) (0, 88.0%) (1, 8.67%)	(-1, 4.67%) (0, 86.0%) (1, 9.33%)	(-1, 8.33%) (0, 84.67%) (1, 7.0%)	(-1, 8.33%) (0, 84.33%) (1, 7.33%)
	ATE	0.05	0.05	-0.01	-0.01
	ARR	-0.05	-0.05	0.01	0.01
	NNT	-20	-20	100	100
Mod 4	Correctly classified patients	80			
	ITE	(-1, 14.0%) (0, 73.0%) (1, 13.0%)	(-1, 14.0%) (0, 73.0%) (1, 13.0%)	(-1, 13.0%) (0, 77.0%) (1, 10.0%)	(-1, 11.0%) (0, 81.0%) (1, 8.0%)
	ATE	-0.01	-0.01	-0.03	-0.03
	ARR	0.01	0.01	0.03	0.03
	NNT	100.00	100.00	33.33	33.33

harm to the overall patient population due to the treatment. Conversely, for Feat_Imp_Tuning and Feat_Imp_Default configurations, negative ATE values of -0.01 are consistently observed, indicating potential benefit for all patients from the treatment.

- For Mod 4, for all configurations, the negative ATE values indicate that the treatment had a beneficial effect on the population as a whole which varying from -0.03 to -0.01.
- By analyzing the results of Mod 1, Mod 2, and Mod 3, we noticed that the longer a model uses patients' monitoring data (i.e., more features), the benefit will decrease. As noticed above, negative values are observed for Mod 1 and Mod 2, with better values achieved for Mod 1, in contrast to Mod 3 which registered positive values for ATE. Therefore, additional data is necessary for a more confident analysis and to refine the approach.

From Table 10, we can also interpret the results of the Number Needed to Treat (NNT) according to the Absolute Risk Reduction (ARR) (see Equation 7) measure. The results' interpretation can be given as follows:

- For Mod 1 and Mod 2, for the different configurations, the positive NNT values mean that the treatment is effective in reducing the risk of the event. Mod 1 NNT values vary from 11.11 to 50 meaning that for the default model's configuration, for instance, we have to treat 50 patients with CS to prevent one additional bad outcome. For Mod 2 NNT values vary from 20 to 50. It is to be noted that there is no other therapy for sepsis, so in this case, we cannot decide whether the values of NNT are low or high. This NNT interpretation is also based on the ARR measures which present positive values, for Mod 1 and Mod 2. This shows that the incidence of the event is lower in the treatment group than in the control group which means that the treatment is effective in reducing the risk of the event. Mod 1 ARR values vary from 0 to 0.09 while for Mod 2 ARR values vary from 0.02 to 0.05.
- In Mod 3, NNT values for Default and Hyper_Tuning configurations are negative, with a magnitude of 20, indicating the number of patients needed to potentially experience harm. This is in line with the negative ARR (Absolute Risk Reduction) values of -0.05, suggesting a higher event incidence in the treatment group compared to the control group. In this context, the treatment does not appear effective and might even be harmful. Conversely, for Feat_Imp_Tuning and Feat_Imp_Default configurations, NNT values are positive at 100, signifying the

Table 11: Statistical metrics for the different modes applied on male data

		Default	Hyper_Tuning	Feat_Imp_Tuning	Feat_Imp_Default
Mod 1	Correctly classified patients	209			
	ITE	(-1, 16.27%) (0, 80.38%) (1, 3.35%)	(-1, 16.75%) (0, 79.9%) (1, 3.35%)	(-1, 17.22%) (0, 76.56%) (1, 6.22%)	(-1, 19.62%) (0, 78.95%) (1, 1.44%)
	ATE	-0.13	-0.13	-0.11	-0.18
	ARR	0.13	0.13	0.11	0.18
	NNT	7.69	7.69	9.09	5.56
Mod 2	Correctly classified patients	209			
	ITE	(-1, 15.69%) (0, 80.88%) (1, 3.43%)	(-1, 15.12%) (0, 80.98%) (1, 3.9%)	(-1, 13.88%) (0, 81.34%) (1, 4.78%)	(-1, 13.88%) (0, 81.82%) (1, 4.31%)
	ATE	-0.12	-0.11	-0.09	-0.1
	ARR	0.12	0.11	0.09	0.1
	NNT	8.33	9.09	11.11	10.00
Mod 3	Correctly classified patients	209			
	ITE	(-1, 12.14%) (0, 80.58%) (1, 7.28%)	(-1, 11.27%) (0, 78.92%) (1, 9.8%)	(-1, 12.5%) (0, 80.77%) (1, 6.73%)	(-1, 11.96%) (0, 79.9%) (1, 8.13%)
	ATE	-0.05	-0.01	-0.06	-0.04
	ARR	0.05	0.01	0.06	0.04
	NNT	20.00	100.00	16.67	25.00
Mod 4	Correctly classified patients	95			
	ITE	(-1, 10.53%) (0, 68.42%) (1, 21.05%)	(-1, 14.74%) (0, 64.21%) (1, 21.05%)	(-1, 8.42%) (0, 72.63%) (1, 18.95%)	(-1, 11.58%) (0, 73.68%) (1, 14.74%)
	ATE	0.11	0.06	0.11	0.03
	ARR	-0.11	-0.06	-0.11	-0.03
	NNT	-9.09	-16.67	-9.09	-33.33

number of patients needed to potentially benefit. This corresponds to lower event incidence in the treatment group, indicating an effective reduction in the risk of the event.

- From Table 10, we can also notice that for these first three models (Mod 1, 2, and 3), when applied to APROCCHSS, the mean ARR has an average of 0.25. This means that the average reduction in risk with corticosteroid treatment for sepsis patients is estimated to be 25% in the APROCCHSS cohort. For Mod 4, results show, when applied to RECORDS, that the mean ARR has an average of 0.02. This means that the average reduction in risk with corticosteroid treatment for sepsis patients is estimated to be 2% in the RECORDS cohort.

To further study and analyze these treatment effectiveness measures, we have used the Conditional Average Treatment Effect (CATE) measure [47]. In our analysis, CATE is reported based on the distribution of females and males over treatment and placebo groups. Table 9 shows some statistical analysis of APROCCHSS and RECORDS reflecting the considered groups for CATE analysis.

- From Table 11, for the male group study, it can be observed that for Mod 1, Mod 2, and Mod 3, for the different configurations, the percentage of patients with an ITE = -1, varies from 11.27% to 19.62%, indicating that the treatment had a positive impact on the individual patient. This is endorsed by the obtained negative ATE values which vary from -0.01 to -0.18, and by the positive ARR values which vary from 0.01 to 0.18. Also, the positive NNT values, varying from 5.56 to 100, indicate that the treatment is effective in reducing the risk of the event. Same as above, we are unable to determine if the values of NNT in this trial are low or high because no other treatment for sepsis is currently being examined.

In Mod 4, for the different configurations, the percentage of patients with an ITE = -1, varies from 8.42% to 14.74%, indicating that the treatment had a positive impact on the individual patient. This is endorsed by the obtained positve ATE values which vary from 0.03 to 0.11, and by the negative ARR values which vary from -0.11 to -0.03. Also, the negative NNT values, varying from -33.33 to -9.09, indicate that the treatment is not effective in reducing the risk of the event. Same as above, we are unable to determine if the values of NNT in this trial are low or high because no other treatment for sepsis is currently being examined.

Table 12: Statistical metrics for the different modes applied on female data

		Default	Hyper_Tuning	Feat_Imp_Tuning	Feat_Imp_Default
Mod 1	Correctly classified patients	118			
	ITE	(-1, 10.71%) (0, 70.54%) (1, 18.75%)	(-1, 11.5%) (0, 73.45%) (1, 15.04%)	(-1, 15.79%) (0, 66.67%) (1, 17.54%)	(-1, 11.4%) (0, 76.32%) (1, 12.28%)
	ATE	0.08	0.04	0.02	0.01
	ARR	-0.08	-0.04	-0.02	-0.01
	NNT	-12.50	-25.00	-50.00	-100.00
Mod 2	Correctly classified patients	118			
	ITE	(-1, 9.4%) (0, 78.63%) (1, 11.97%)	(-1, 9.32%) (0, 77.12%) (1, 13.56%)	(-1, 8.47%) (0, 78.81%) (1, 12.71%)	(-1, 8.47%) (0, 79.66%) (1, 11.86%)
	ATE	0.03	0.04	0.04	0.03
	ARR	-0.03	-0.04	-0.04	-0.03
	NNT	-33.33	-25.00	-25.00	-33.33
Mod 3	Correctly classified patients	118			
	ITE	(-1, 10.17%) (0, 74.58%) (1, 15.25%)	(-1, 8.47%) (0, 72.03%) (1, 19.49%)	(-1, 12.71%) (0, 70.34%) (1, 16.95%)	(-1, 11.02%) (0, 71.19%) (1, 17.8%)
	ATE	0.05	0.11	0.04	0.07
	ARR	-0.05	-0.11	-0.04	-0.07
	NNT	-20.00	-9.09	-25.00	-14.29
Mod 4	Correctly classified patients	100			
	ITE	(-1, 7.94%) (0, 82.54%) (1, 9.52%)	(-1, 8.75%) (0, 80.0%) (1, 11.25%)	(-1, 13.75%) (0, 75.0%) (1, 11.25%)	(-1, 12.5%) (0, 77.5%) (1, 10.0%)
	ATE	0.02	0.03	-0.03	-0.03
	ARR	-0.02	-0.03	0.03	0.03
	NNT	-50.00	-33.33	33.33	33.33

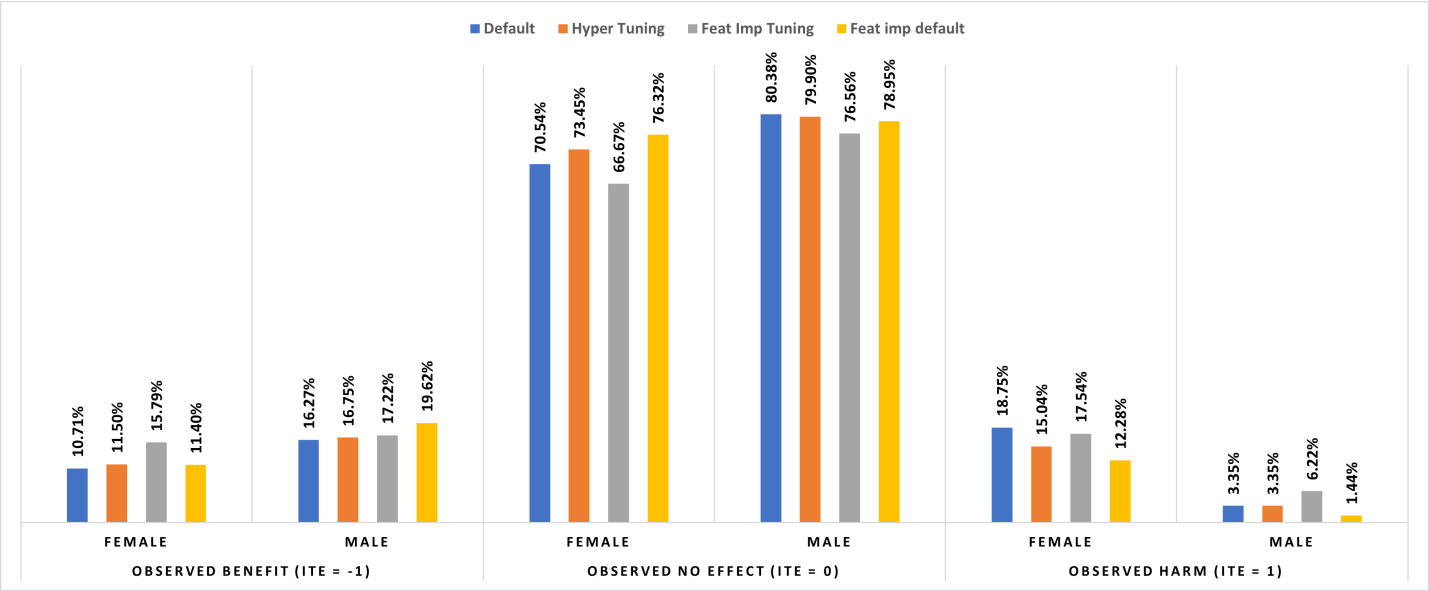

Figure 8: Percentage CATE male and female for Mod 1 with APROCCHSS

- From Table 12, for the female group study, it can be observed that for Mod 1, Mod 2, and Mod 3, the negative impact of the treatment on patients and the whole population, was for all configurations. The percentage of patients with an ITE = -1 varies from 8.47 to 15.79% with an ATE varying from 0.01 to 0.11. For Mod 4, for all configurations, patients had a beneficial effect when using the treatment indicating ITE = -1; ranging from 7.94 to 13.75%. Negative ARR values for Mod1, Mod 2, and Mod 3, indicate that the incidence of the event is higher in the treatment group than in the control group, which means that the treatment is not effective in reducing the risk of the event (values vary from -0.11 to -0.01). A negative NNT varying from -100 to -9.09 reflects the number of patients needed to harm.

- From Figures 8 and 9, it can be noticed that corticotherapy has a more positive effect (ITE = -1) on males than on females. For Mod 1 (Figure 8), for all configurations, the negative percentage of ITE values for males varies from 16.27% to 19.62% while for females values vary from 10.71% to 15.79%. For Mod 2 (Figure 9), for all configurations, the negative percentage of ITE values for males varies from 13.88% to 15.69% while for females the values showing small variation vary from 8.47% to 9.4%. However, for Mod 3 (Figure 10), corticotherapy has a more positive effect (ITE = -1) with a slight difference on males than on females. The negative percentage of ITE values for females varies from 8.47% to 12.71% while for males the values vary from 11.27% to 12.5%. For Mod 4 (Figure 11), it can be observed that corticotherapy has a more positive effect (ITE = -1) on males than on females. For Mod 4, for all configurations, the negative percentage of ITE values for males rages from 8.42 to 14.74%; while on females ITE values ranging from 7.94 to 13.75%.

4. Discussions

In this section, we aim to interpret and analyze the results obtained in this study, highlighting the main findings, and addressing their implications.

4.1. Experiment 1

The performance of machine learning models can be improved by combining feature selection and hyperparameter tuning [48]. This approach helps identify the most important variables that have the greatest impact on the outcome and optimizes the model's performance by selecting the best set of parameters for a given cohort. In addition, balancing the cohort using the SMOTE technique can increase the diversity of the data and provide a more representative sample of the sepsis population. Several recent studies have shown the effectiveness of feature selection, hyperparameter tuning, and SMOTE in improving the performance of machine learning models in healthcare applications [49]. Logistic Regression was found to be the best model among all the models evaluated in the study, achieving the highest values in accuracy, precision, recall, F1 metric, and AUC. Therefore, it was selected as the "BestMod" for the rest of the experimental protocol. Moreover, Logistic Regression accurately distinguished sepsis responders and non-responders to corticosteroid treatment with an AUC of 72%. This result is consistent with recent research in the field of machine learning for healthcare, where Logistic Regression has been shown to be effective in predicting outcomes in a variety of clinical settings.

4.2. Experiment 2

To address the research question of whether patient features affect the accuracy of the obtained results **RQ2**, we carried out Experiment 2. By analyzing the findings obtained from the

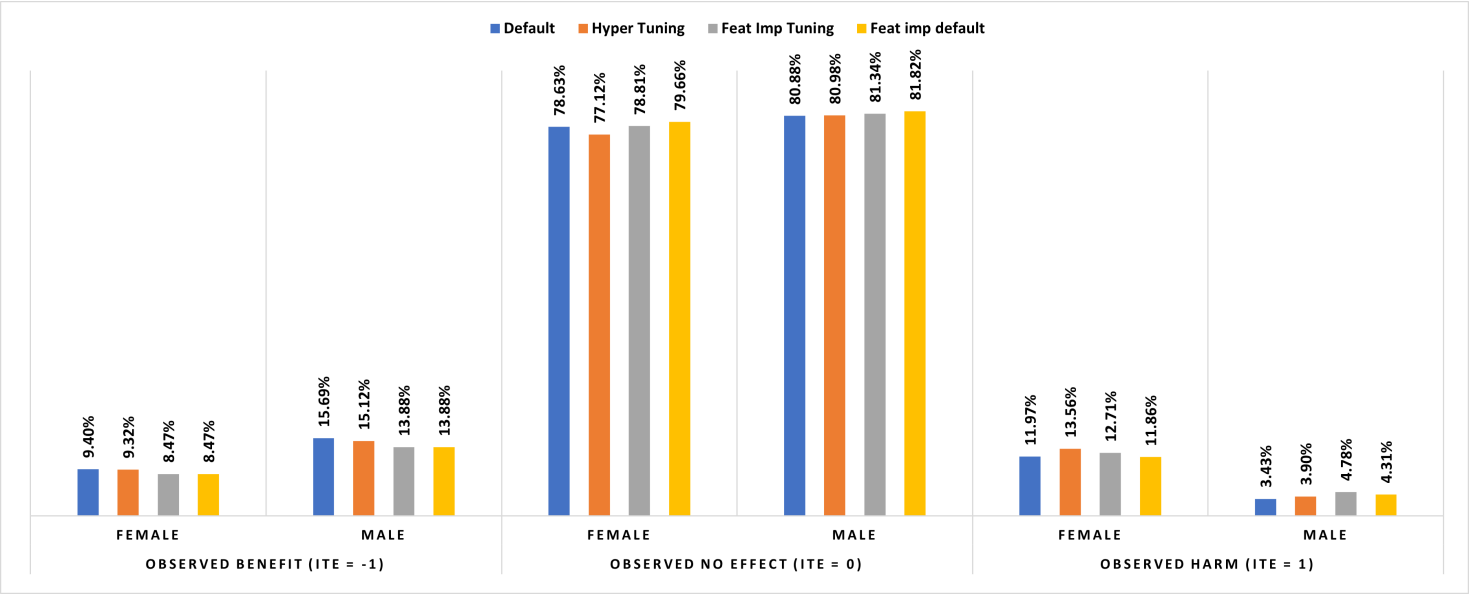

Figure 9: Percentage CATE male and female for Mod 2 with APROCCHSS

model's variants, regarding only patients who got the corticosteroid treatment, we found that when more information is added to the model (Day 0, Day 1, Day 2, and the difference between Day 2 and Day 1), the prediction results are improved. Our best-performing model, which is Mod 3 achieved an AUC score of 76.1%. Moreover, to answer the fourth research question of whether the learned model can be generalized to different sepsis cohorts **RQ4**, we trained the model on APROCCHSS and tested it on the RECORDS cohort (Table 3). By interpreting the results obtained from Mod 5 (AUC score 55% with Feat_Imp_Tuning), we can conclude that the answer is negative. Thus, patient characteristics (such as those affected by COVID-19) can completely reduce the performance of a learned model; hence, the model cannot be generalized to different cohorts.

4.3. Assessment of the effectiveness of the treatment

To complement the analyses conducted in Experiments 1 and 2, evaluating the effectiveness of the treatment will enable us to address **RQ3**, which pertains to the impact of corticotherapy on both individual treated patients and all treated patients as a whole. The reported findings are based on patients who were correctly predicted by the models, ensuring the accuracy and reliability of the results. The analysis of the results reveals that the ITE measure may vary significantly depending on the configuration used for the Logistic Regression. Additionally, the ATE interpretations demonstrate that the longer a model uses patients' monitoring data, the less benefit it may have, with Mod 1 and Mod 2 achieving negative ATE values while Mod 3 registered positive values for default and Hyper_Tuning configurations. Analysis of the results within the RECORDS cohort (Mod 4) indicates that the ITE measure's performance is highly dependent on both the predictive models and the specific cohort used. These results consistently demonstrate the treatment's overall benefit for all patients. These findings suggest that additional data is necessary for a more confident analysis and to refine the approach. Overall, this study provides valuable insights into the potential of using machine learning models to predict treatment effects and highlights the importance of careful evaluation and interpretation of results [50].

In order to conduct a more thorough examination of the efficacy of the treatments, our study utilized the CATE measure. The results show that the effectiveness of corticotherapy varies depending on the gender and the specific model configuration used. For male patients, the treatment had a positive impact on individual patients, as indicated by the negative ATE values and positive ARR values. The NNT values also suggest that the treatment is effective in reducing the risk of the event. However, for female patients, the treatment had a negative impact on individual patients, as indicated by the positive ATE values and negative ARR values. The NNT values also suggest that the treatment is not effective in reducing the risk of the event. The analysis also reveals that corticotherapy has a more positive effect on males than females in all model configurations, but in the model based on RECORDS cohort, it has a more positive effect on females. These findings suggest the need for further investigation to better understand the gender differences in the effectiveness of corticotherapy and to optimize the treatment approach for each gender [51].

To enhance the robustness of our statistical analysis and obtain more precise results, we performed additional calculations using our signature. It's important to note that our signature is based on a Logistic Regression model trained with SMOTE data, and we employed default parameters for feature importance configuration. The primary motivation behind this investigation lies in the active utilization of the signature within the APHP. Offering further insights into treatment effectiveness could greatly assist specialists in early patient intervention.

To achieve this, Figure 12 demonstrates the signature's ability to accurately predict patients who respond to the treatment as CS-sensitive with a 70% success rate. Among these correctly

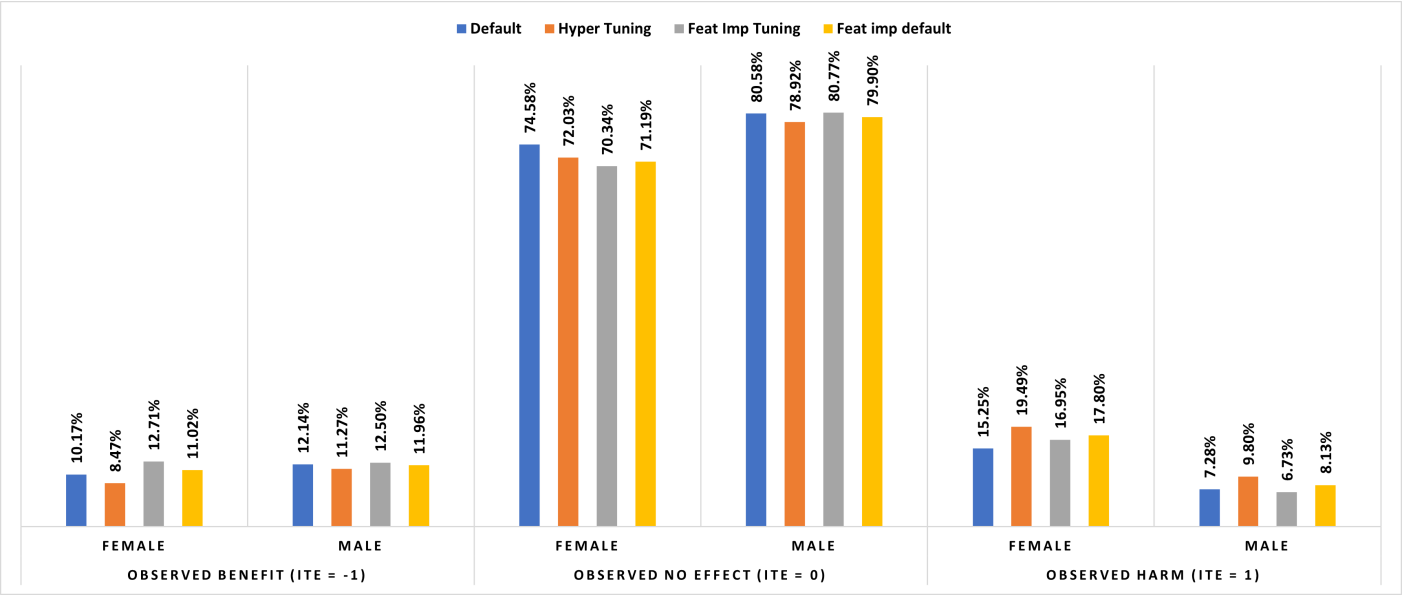

Figure 10: Percentage CATE male and female for Mod 3 with APROCCHSS

predicted patients, 20% exhibited genuine improvement, as indicated by an ITE score of -1. Conversely, the signature also effectively identifies CS-resistant patients, again with a 70% accuracy rate. Among these correctly identified patients, 6% experienced true deterioration, as denoted by an ITE score of 1.

These results not only enhance the precision of our findings but also lay the groundwork for future research avenues. This could include a more in-depth exploration of patient characteristics associated with positive and negative treatment outcomes.

5. Conclusion and future directions

In this paper, we implemented a consistent data mining approach and developed a prediction model, referred to as "signature", which aims at promptly identifying the responsiveness of patients to corticotherapy. We have considered the APROCCHSS and RECORDS cohorts. APROCCHSS results from a randomized controlled trial covering 1241 sepsis patients and the RECORDS cohort includes sepsis patients affected by COVID-19. Data was gathered and collected by the APHP clinicians.

Two experiments were conducted to study the effectiveness of our approach to identify whether a sepsis patient is sensible or resistant to corticosteroid treatment. In the first experiment, we carried out a thorough study of several machine learning models and compared their performance in terms of recognizing CS responsiveness. Additional configurations have been included in order to further study the performance of our models, including hyperparameter tuning and selecting the most important features. We noticed that in the obtained results most of the models demonstrated that when the most relevant features were selected in combination with hyperparameter tuning (i.e., Feat_Imp_Tuning), they achieved the highest level of performance. Logistic Regression is considered to be the best model for our study with an AUC value of 72% with the configuration Feat_Imp_Default; showing good performance in distinguishing sepsis responders and non-responders to corticosteroid treatment.

The second experiment is devoted to exploring the generalizability of the Logistic Regression model by testing it using additional data collected over time or from different patient groups belonging to another cohort: RECORDS. We found that the prediction results are improved when more information about the patient has been added to the model (Mod 3). Specifically, we noticed a statistical significant difference between Mod 1 and Mod 2 where we have added additional features (from Day 0 to Day 0 and Day 1 features). Mod 3 achieved an AUC score of 76% which confirms that training the model with more features increases its performance. On the other hand, when training the model on the APROCCHSS cohort and testing it on the RECORDS cohort, the model was not effective as it recorded an AUC value of 54% with the Default and Hyper_Tuning configurations (i.e., Mod 5). As a consequence, a generalization may not be applied to different sepsis cohorts, especially that the RECORDS testing set includes COVID-19 patients.

In addition, to complement the analysis realized in experiment 1 and experiment 2, we analyzed the effect of corticosteroid treatment on individual treated patients and all treated patients as a whole using treatment effectiveness metrics: ITE, ATE, ARR, and NNT. These measures have been calculated using Logistic Regression (based on SMOTE) using various feature sets (Mod 1 to Mod 4) with the four different configurations: default, Hyper Tuning, Feat_Imp_Tuning, and Feat_Imp_Default. The obtained results showed, in most cases, it is difficult to decide whether the treatment is effective or not with the percentage of patients with an ITE = 0 varying from 73.1% to 84.7% for the different model variants. Moreover, we observed a decreasing benefit when the model uses patient

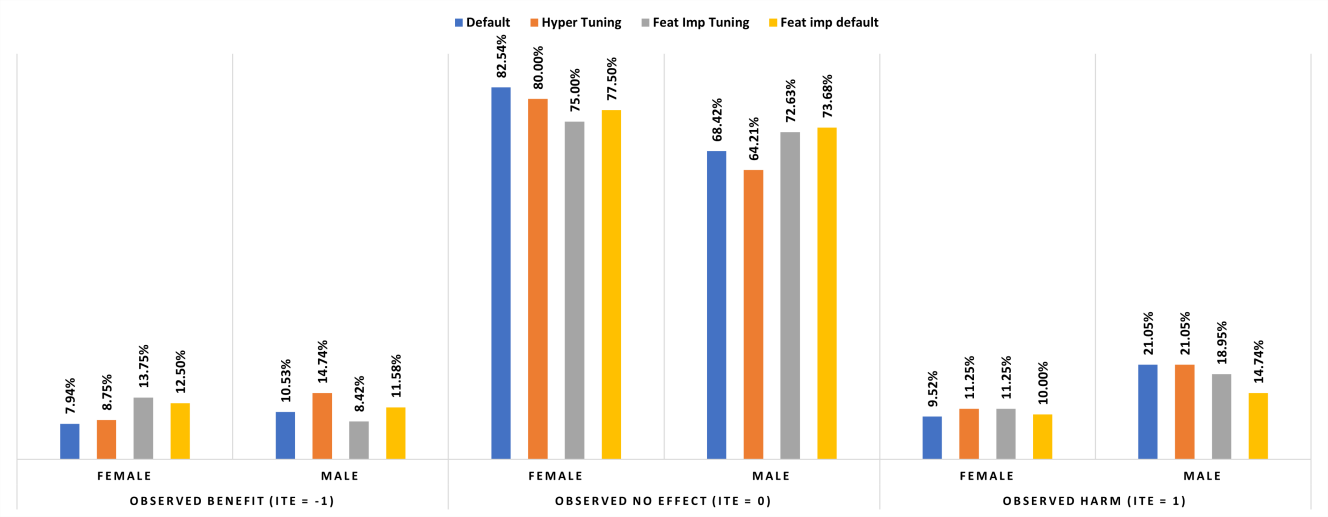

Figure 11: Percentage CATE male and female for Mod 4 with RECORDS

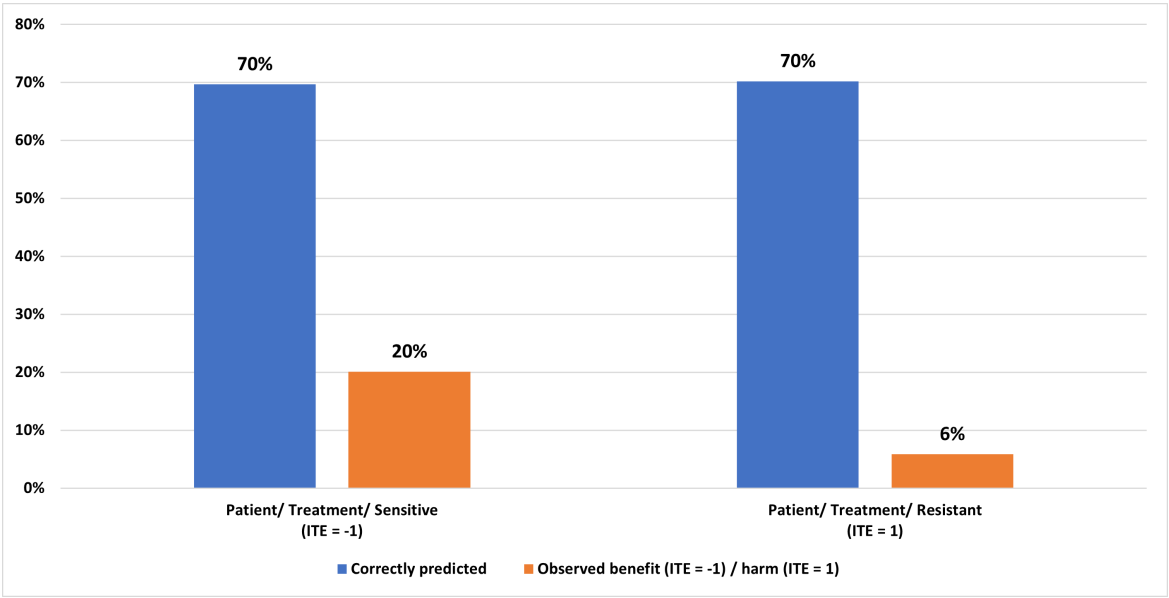

Figure 12: Predictive performance of the signature for treatment response for APROCCHSS

monitoring data for a longer period (Mod 3), which calls for a cautious interpretation and the need for further input data to refine the signature. Also, when investigating the CATE measure taking as a factor the sex of patients, we observed that the treatment had a positive impact on males more than on female.

Limitations of this study stem from its reliance on the initial release of the RECORDS project data, which restricts the depth of analysis. Moreover, it is important to note that the RECORDS observational cohort used in this experiment had a limited observational period, providing only preliminary clinical and biological data. However, working with the full trial version of the cohort is expected to yield more precise and accurate predictions in the future. This expanded cohort will likely enhance the reliability and robustness of the study's findings and predictions. Additionally, the absence of the placebo group in the current research limits the analysis and restricts our findings. Another limitation of this work is the size of the APROCCHSS cohort used for prediction. Generally, the performance of predictive models relies not only on the quality of the input data but also on the size of the dataset. The larger the cohort used for training, the more accurate the model will be.

The RECORDS project is still ongoing, and this signature presents the first release. It will be further updated and refined as new data and variables become available. As future directions, we will add more data and features which will be made at disposal from the RECORDS project such as multi-omics data. The accessibility to multi-omics data has revolutionized the study of biology and medicine by creating pathways for integrated system-level approaches. In addition, combining multi-omics data with clinical information has reached a

high level of significance in order to obtain useful information [52]. Moreover, exploring computer vision techniques in the medical domain may provide us with interesting insights and can improve the detection of CS-sensitivity in sepsis patients by analyzing radiological images which can be more precise than clinical data in terms of locating the infection site as early as possible. Based on the idea of having more input data, and exploiting more advanced methods, sophisticated deep learning-based techniques will be interesting in our future perspectives.

Acknowledgments including declarations

Statements of ethical approval: For the APROCCHSS study, the protocol and qualification of all investigators were approved by the Ethics Committee (Comité de Protection des Personnes, CPP) of Saint-Germain-en-Laye, France, on November 22, 2007. The trial was registered at ClinicalTrials.gov under NCT00625209. For the RECORDS study, the protocol and qualification of all investigators were approved by the Ethics Committee (Comité de Protection des Personnes, CPP) of Dijon, France, on 6 April 2020. The trial was registered at ClinicalTrials.gov under NCT04280497.

Funding: This study has received a public grant through the national program "*Programme d'Investissements d'Avenir (PIA)*" (as part of the France 2030 programme) under the reference ANR-18-RHUS-0004. This work is part of the Federation Hospitalo-Universitaire (FHU) Saclay and Paris Seine Nord Endeavour to PerSonalize Interventions for Sepsis (SEPSIS). This work was also supported by ANR PIA funding: ANR-20-IDEES-0002.

Competing interests: Authors declare that they have no competing interests.

Appendix A. 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.

References

[1] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, et al., The third international consensus definitions for sepsis and septic shock (sepsis-3), Jama 315 (8) (2016) 801–810.

[2] S. Esposito, G. De Simone, G. Boccia, F. De Caro, P. Pagliano, Sepsis and septic shock: New definitions, new diagnostic and therapeutic approaches, Journal of global antimicrobial resistance 10 (2017) 204–212.

[3] A. L. León, N. A. Hoyos, L. I. Barrera, G. De La Rosa, R. Dennis, C. Dueñas, M. Granados, D. Londoño, F. A. Rodríguez, F. J. Molina, et al., Clinical course of sepsis, severe sepsis, and septic shock in a cohort of infected patients from ten colombian hospitals, BMC infectious diseases 13 (1) (2013) 1–9.

[4] K. E. Rudd, S. C. Johnson, K. M. Agesa, K. A. Shackelford, D. Tsoi, D. R. Kievlan, D. V. Colombara, K. S. Ikuta, N. Kissoon, S. 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) (2020) 200–211.

[5] F. Pandolfi, D. Guillemot, L. Watier, C. Brun-Buisson, Trends in bacterial sepsis incidence and mortality in france between 2015 and 2019 based on national health data system (système national des données de santé (snds)): a retrospective observational study, BMJ open 12 (5) (2022) e058205.

[6] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, T. G. Buchman, An interpretable machine learning model for accurate prediction of sepsis in the icu, Critical care medicine 46 (4) (2018) 547.

[7] D. B. Antcliffe, K. L. Burnham, F. Al-Beidh, S. Santhakumaran, S. J. Brett, C. J. Hinds, D. Ashby, J. C. Knight, A. C. Gordon, Transcriptomic signatures in sepsis and a differential response to steroids. from the vanish randomized trial, American journal of respiratory and critical care medicine 199 (8) (2019) 980–986.

[8] J. Fleuriet, N. Heming, F. Meziani, J. Reignier, P.-L. Declerq, E. Mercier, G. Muller, G. Colin, X. Monnet, A. Robine, et al., Rapid recognition of corticosteroid resistant or sensitive sepsis (records): study protocol for a multicentre, placebo-controlled, biomarker-guided, adaptive bayesian design basket trial, BMJ open 13 (3) (2023) e066496.

[9] R. Pirracchio, A. Hubbard, C. L. Sprung, S. Chevret, D. Annane, et al., Assessment of machine learning to estimate the individual treatment effect of corticosteroids in septic shock, JAMA network open 3 (12) (2020) e2029050–e2029050.

[10] J. M. Rague, Effect of hydrocortisone on development of shock among patients with severe sepsis: The hypress randomized clinical trial: Ken d, trips e, marx g, et al. jama. 2016; 316: 1775-1785, Journal of Emergency Medicine 52 (3) (2017) 387–388.

[11] R. Moreno, C. Sprung, D. Annane, S. Chevret, J. Briegel, D. Keh, M. Singer, Y. Weiss, D. Payen, B. Cuthbertson, et al., Time course of organ failure in patients with septic shock treated with hydrocortisone: results of the corticus study, in: Applied physiology in intensive care medicine 1, Springer, 2012, pp. 423–430.

[12] B. Venkatesh, S. Finfer, J. Cohen, D. Rajbhandari, Y. Arabi, R. Bellomo, L. Billot, M. Correa, P. Glass, M. Harward, et al., Adjunctive glucocorticoid therapy in patients with septic shock, New England Journal of Medicine 378 (9) (2018) 797–808.

[13] D. Annane, A. Renault, C. Brun-Buisson, B. Megarbane, J.-P. Quenot, S. Siami, A. Cariou, X. Forceville, C. Schwebel, C. Martin, et al., Hydrocortisone plus fludrocortisone for adults with septic shock, New England Journal of Medicine 378 (9) (2018) 809–818.

[14] H. R. Wong, S. J. Atkinson, N. Z. Cvijanovich, N. Anas, G. L. Allen, N. J. Thomas, M. T. Bigham, S. L. Weiss, J. C. Fitzgerald, P. A. Checchia, et al., Combining prognostic and predictive enrichment strategies to identify children with septic shock responsive to corticosteroids, Critical care medicine 44 (10) (2016) e1000.

[15] L. M. Fleuren, T. L. Klausch, C. L. Zwager, L. J. Schoonmade, T. Guo, L. F. Roggeveen, E. L. Swart, A. R. Girbes, P. Thoral, A. Ercole, et al., Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy, Intensive care medicine 46 (3) (2020) 383–400.

[16] F. Fang, Y. Zhang, J. Tang, L. Lunsford, T. Li, R. Tang, et al., Association of corticosteroid treatment with outcomes in adult patients with sepsis, A systematic review and meta-analysis (2019) 179.

[17] M. M. Islam, T. Nasrin, B. A. Walther, C.-C. Wu, H.-C. Yang, Y.-C. Li, Prediction of sepsis patients using machine learning approach: a meta-analysis, Computer methods and programs in biomedicine 170 (2019) 1–9.

[18] M. Moor, B. Rieck, M. Horn, C. R. Jutzeler, K. Borgwardt, Early prediction of sepsis in the icu using machine learning: a systematic review,

Frontiers in medicine 8 (2021) 607952.
[19] H.-F. Deng, M.-W. Sun, Y. Wang, J. Zeng, T. Yuan, T. Li, D.-H. Li, W. Chen, P. Zhou, Q. Wang, et al., Evaluating machine learning models for sepsis prediction: A systematic review of methodologies, Iscience (2021) 103651.
[20] R. Kamaleswaran, J. Lian, D.-L. Lin, H. Molakapuri, S. Nunna, P. Shah, S. Dua, R. Padman, Predicting volume responsiveness among sepsis patients using clinical data and continuous physiological waveforms, in: AMIA Annual Symposium Proceedings, Vol. 2020, American Medical Informatics Association, 2020, p. 619.
[21] H. Shi, H. Wang, Y. Huang, L. Zhao, C. Qin, C. Liu, A hierarchical method based on weighted extreme gradient boosting in ecg heartbeat classification, Computer methods and programs in biomedicine 171 (2019) 1–10.
[22] R. Chiong, Z. Fan, Z. Hu, F. Chiong, Using an improved relative error support vector machine for body fat prediction, Computer Methods and Programs in Biomedicine 198 (2021) 105749.
[23] M. G. Signorini, N. Pini, A. Malovini, R. Bellazzi, G. Magenes, Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring, Computer Methods and Programs in Biomedicine 185 (2020) 105015.
[24] X. Wang, Z. Wang, J. Weng, C. Wen, H. Chen, X. Wang, A new effective machine learning framework for sepsis diagnosis, IEEE access 6 (2018) 48300–48310.
[25] S. L. Javan, M. M. Sepehri, M. L. Javan, T. Khatibi, An intelligent warning model for early prediction of cardiac arrest in sepsis patients, Computer methods and programs in biomedicine 178 (2019) 47–58.
[26] F. van Wyk, A. Khojandi, A. Mohammed, E. Begoli, R. L. Davis, R. Kamaleswaran, A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier, International journal of medical informatics 122 (2019) 55–62.
[27] M. A. Reyna, C. S. Josef, R. Jeter, S. P. Shashikumar, M. B. Westover, S. Nemati, G. D. Clifford, A. Sharma, Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019, Critical care medicine 48 (2) (2020) 210–217.
[28] A. Raghu, M. Komorowski, L. A. Celi, P. Szolovits, M. Ghassemi, Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach, in: Machine Learning for Healthcare Conference, PMLR, 2017, pp. 147–163.
[29] C. Lam, A. Siefkas, N. S. Zelin, G. Barnes, R. P. Dellinger, J.-L. Vincent, G. Braden, H. Burdick, J. Hoffman, J. Calvert, et al., Machine learning as a precision-medicine approach to prescribing covid-19 pharmacotherapy with remdesivir or corticosteroids, Clinical therapeutics 43 (5) (2021) 871–885.
[30] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, A. A. Faisal, The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care, Nature medicine 24 (11) (2018) 1716–1720.
[31] D. Annane, C. B. Buisson, A. Cariou, C. Martin, B. Misset, A. Renault, B. Lehmann, V. Millul, V. Maxime, E. Bellissant, Design and conduct of the activated protein c and corticosteroids for human septic shock (aprocchss) trial, Annals of intensive care 6 (1) (2016) 1–13.
[32] K. S. Kim, G. J. Suh, K. Kim, W. Y. Kwon, J. Shin, Y. H. Jo, J. H. Lee, H. Lee, Quick sepsis-related organ failure assessment score is not sensitive enough to predict 28-day mortality in emergency department patients with sepsis: a retrospective review, Clinical and Experimental Emergency Medicine 6 (1) (2019) 77.
[33] D. Moher, S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, D. G. Altman, Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials, International journal of surgery 10 (1) (2012) 28–55.
[34] D. Elreedy, A. F. Atiya, A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance, Information Sciences 505 (2019) 32–64.
[35] S. Alabdulwahab, B. Moon, Feature selection methods simultaneously improve the detection accuracy and model building time of machine learning classifiers, Symmetry 12 (9) (2020) 1424.
[36] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, A comparative analysis of gradient boosting algorithms, Artificial Intelligence Review 54 (2021) 1937–1967.
[37] A. Parmar, R. Katariya, V. Patel, A review on random forest: An ensemble classifier, in: International Conference on Intelligent Data Communication Technologies and Internet of Things, Springer, 2018, pp. 758–763.
[38] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez, A comprehensive survey on support vector machine classification: Applications, challenges and trends, Neurocomputing 408 (2020) 189–215.
[39] N. Ketkar, Stochastic gradient descent, in: Deep learning with Python, Springer, 2017, pp. 113–132.
[40] B. Charbuty, A. Abdulazeez, Classification based on decision tree algorithm for machine learning, Journal of Applied Science and Technology Trends 2 (01) (2021) 20–28.
[41] J. Hao, T. K. Ho, Machine learning made easy: a review of scikit-learn package in python programming language, Journal of Educational and Behavioral Statistics 44 (3) (2019) 348–361.
[42] E. Ostertagova, O. Ostertag, J. Kováč, Methodology and application of the kruskal-wallis test, Applied mechanics and materials 611 (2014) 115–120.
[43] J. A. Dorresteijn, F. L. Visseren, P. M. Ridker, A. M. Wassink, N. P. Paynter, E. W. Steyerberg, Y. van der Graaf, N. R. Cook, Estimating treatment effects for individual patients based on the results of randomised clinical trials, Bmj 343 (2011).
[44] R. Newcomb, Absolute risk reduction, Methods and Applications of Statistics in Clinical Trials: Concepts, Principles, Trials, and Design (2014) 1–13.
[45] L. Citrome, T. A. Ketter, When does a difference make a difference? interpretation of number needed to treat, number needed to harm, and likelihood to be helped or harmed, International journal of clinical practice 67 (5) (2013) 407–411.
[46] J. Muschelli III, Roc and auc with a binary predictor: a potentially misleading metric, Journal of classification 37 (3) (2020) 696–708.
[47] S. R. Künzel, J. S. Sekhon, P. J. Bickel, B. Yu, Metalearners for estimating heterogeneous treatment effects using machine learning, Proceedings of the national academy of sciences 116 (10) (2019) 4156–4165.
[48] R. Valarmathi, T. Sheela, Heart disease prediction using hyper parameter optimization (hpo) tuning, Biomedical Signal Processing and Control 70 (2021) 103033.
[49] D. Velusamy, K. Ramasamy, Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset, Computer Methods and Programs in Biomedicine 198 (2021) 105770.
[50] F. Rotolo, X. Paoletti, S. Michiels, surrosurv: An r package for the evaluation of failure time surrogate endpoints in individual patient data meta-analyses of randomized clinical trials, Computer methods and programs in biomedicine 155 (2018) 189–198.
[51] Q. Fan, Y.-C. Hsu, R. P. Lieli, Y. Zhang, Estimation of conditional average treatment effects with high-dimensional data, Journal of Business & Economic Statistics 40 (1) (2022) 313–327.
[52] I. Subramanian, S. Verma, S. Kumar, A. Jere, K. Anamika, Multi-omics data integration, interpretation, and its application, Bioinformatics and biology insights 14 (2020) 1177932219899051.