Application of an adaptable multimodal machine learning approach to predict COVID-19 outcomes for hospital resource management

Silvia Gómez-Zorrilla
Infectious Diseases Department, Hospital del Mar, Infectious Pathology and Antimicrobials Research Group (IPAR)
Inmaculada López-Montesinos
Infectious Diseases Department, Hospital del Mar, Infectious Pathology and Antimicrobials Research Group (IPAR)
Eva Pérez
Epilepsy Unit, Neurology Department, Hospital del Mar
Pau Pomés
Bioenginery-Universitat Pompeu Fabra (UPF)
Maria Luisa Sorli
Infectious Diseases Department, Hospital del Mar, Infectious Pathology and Antimicrobials Research Group (IPAR)
Robert Güerri-Fernández
Infectious Diseases Department, Hospital del Mar, Infectious Pathology and Antimicrobials Research Group (IPAR)
Santiago Grau
Pharmacy Department, Infectious Pathology and Antimicrobials Research Group (IPAR)
Albert Márquez
Information and Comunication System department, Hospital el Mar, Barcelona, Spain.
Jordi Martínez-Roldán
Innovation and Digital Transformation Manager, Hospital del Mar, Barcelona, Spain.
Alessandro Principe (alessandro.principe@upf.edu)
Institut Hospital del Mar d'Investigacions Mèdiques
Juan P. Horcajada
Infectious Diseases Department, Hospital del Mar, Infectious Pathology and Antimicrobials Research Group (IPAR)

Article

Keywords:

Posted Date: April 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-590537/v2

Abstract

Beyond the health consequences, coronavirus disease 2019 (COVID-19) has shown the inadequacies and limitations of healthcare systems all over the world. Predictions of good and bad outcomes could improve lockdown and curfew policies and reduce the COVID-19 health impact. We describe a multimodal machine learning (ML) approach for predicting early hospital discharge and other key clinical outcomes, such as respiratory worsening, intensive care unit admission or the need for rescue therapy. Using this model, our hospital occupancy could have decreased by about 50% during the first wave of the pandemic. Our model used both raw laboratory results (LR) and electronic health records (EHR), which allow better anticipation of worsening through selective reporting of LRs. Our approach could be a useful, customizable, and adaptable hospital resource management tool, especially to assist in interventions targeted at reducing COVID-19 infection rates.

Introduction

Despite the positive results of vaccine trials[1], it will take several months to reach the critical number of vaccinated people in most countries. Assessments of long-term immunity will be crucial to predict whether the vaccination strategy will be able to eradicate, or at least curb the pandemic. Thus, apart from strategies to reduce infection rates, good hospital management is also of paramount importance and could break the vicious cycle that currently affects many economies; indeed, increasing early discharge (ED) and allocating the necessary number of resources for in-room patients could cut hospital occupancy and expenses. Many studies have turned to machine learning (ML) for coronavirus disease-19 (COVID-19[2]), the majority focusing on its diagnosis[3–5] and disease lethality[6–11]. Although the latter is related to clinical outcome, it cannot be translated into better resource allocation. Only a handful of studies in fact have tackled the problem in terms of time[12,13]. Here we describe a multimodal ML approach not only able to predict ED, but also other key clinical variables such as intensive care unit admission (ICU), intensification of oxygen support (OSi) or the need for rescue therapy. The clinical outcome prediction (COP) framework, COP over time (COP-tt), implements a ML algorithm that is already used for mortality risk assessment[11] and clinical outcome prediction[13]. As in the study by Arvind & colleagues[13], COP-tt analyzes time windows, although with a dynamic array of variables, giving more freedom in terms of missing data.

Briefly, for every subject, *states* were recorded. States represented the evolution of patients in time, therefore a subject could present from one to many states, depending on their outcome and available clinical history updates. An electronic health record (EHR) state was a set of variables extracted from a clinical history entry, while a laboratory record (LR) state was a set of blood test results. States were updated after one or more days, when at least one variable or the clinical outcome changed. Each state was associated with an outcome (Fig. 1). To make predictions, models were constructed using the worst outcome. In other words, when the database was queried to train models, the early stages of the severe acute respiratory syndrome (SARS) were associated with the worst one reached by the patient. In this way, the algorithm could evaluate early stages of a severe condition. However, the original state-outcome association was used to calculate how early the algorithm could identify the severity. Models were constructed using a percentage of available subjects (sampling) and predicted the outcome of patients not considered in the training set. Indivual patient

data though were never split, therefore the common pitfall to predict the future from the past was always avoided. Repeated sampling-prediction iterations were made to ensure homogeneous outcome prediction results (see methods for details). For healthcare structures, each iteration could correspond to a week, after which a new set of models would be constructed for the following one, with the advantage that in a real scenario, sampling would not be necessary and new subjects would increase the accuracy of the model.

Results

Subjects and sources

Since LRs and EHRs in our hospital are stored and retrieved separately, the two sources were treated likewise. They are in fact rather different, since variables in the EHRs are entered by the clinician only if deemed necessary. Overall, the data of 4511 subjects with suspected COVID-19 were retrieved between March and August 2020. Of these, 2023 (45%) were confirmed to have been infected by SARS coronavirus 2 (SARS-CoV2), through reverse-transcriptase polymerase chain reaction (RT-PCR) rapid antigen detection or serology. Only subjects with retrievable LR were considered for calculations, so that 1685 in all were processed. Early hospital discharge was considered the main variable affecting hospital resource management. We also assessed other key clinical outcomes for SARS-CoV-2 infection, such as respiratory deterioration, need for rescue therapy or need for ICU admission. To calculate ED, all subjects with suspected COVID-19 were considered, since biological confirmations were not reliable at the beginning of the pandemic. For clinical outcomes, only SARS-CoV-2-positive subjects were considered to avoid bias. See methods for a detailed description of variables and their selection. Over the course of the iterations, the outcomes of 2331 patients (51% of all subjects) were predicted. COP-tt allows prediction and a time to event analysis by moving the classifier along the timeline of the test subjects (Fig. 1). Whenever an event of interest is classified, the time from the real event (e.g. ICU admission) can be calculated.

Prediction of early discharge

LR models allowed better anticipation of ED than EHR (Fig. 2A). For both LR and EHR models, sensitivity was 100%, receiver-operator characteristic area under the curves (ROC-AUCs) were 93.1% and 88.8%, and anticipation was 11 ± 33 days and 5 ± 14 (median ± interquartile-range) respectively. However, we consider that the LR models overestimated anticipation of ED, because they did not account for other causes of prolonged stay. If COP-tt had been implemented, taking both quarantine periods and lockdown restrictions into consideration, our overall hospital occupancy could have been reduced to 49% (based on EHR) or 45% (based on LR) on average. More importantly however, when the pandemic was partially contained and only lockdown restrictions were in place, hospital occupancy could have been lowered to around 40% (Fig. 2B), suggesting that such a system could help keep hospital occupancy to acceptable levels, even reducing other containment measures. Expected readmission rates would have been less than 10% on average (4–14%). ROC-AUCs rose to 98.0% (EHR using all variables, see *Database size, accuracy and prediction related to variable weight*) and 97.6% (LR) when only SARS-CoV2 subjects were considered, but differences in anticipation rates were reduced to 3 ± 10 and 5 ± 20 days respectively, even when adjusted for the current need to initiate a clinical trial (see supplementary results).

Predicting worsening and rescue therapy

For OSi and ICU admission, the sources worked in opposite ways. The EHR models anticipated OSi by 0 ± 2 days and ICU by 4 ± 9 days with ROC-AUCs of 80.6% and 91.2%, respectively. The LR models were less effective and accurate (ROC-AUCs: 73.6% and 82.1%). Interestingly, by using EHRs, the need for rescue therapy (RT) could be anticipated by 2 ± 5 days. The shortage of RT at the beginning of the pandemic was a considerable problem, so that this kind of prediction would have been valuable.

Database size, accuracy and prediction related to variable weight

To measure the impact of variables on accuracy and prediction time, models with a reduced set of variables were trained and evaluated. To generate these models the variable weights were employed. The weight of a variable defines its importance in the training and performance of a model (see Weighting of variables in Methods). At first, the variables representing the 20% of weight were used (2 to 3 variables for EHR, 6 for LR), then variables were added by steps of approximately 20% of weight. Therefore, as expected, the number of variables (database size) rose exponentially (Fig. 3), since the first few variables weighed considerably more than the following (Table 1). ROC-AUCs generally rose from step to step, although the core variables contributed to more than 80% of ROC-AUCs. However, the prediction time trends were the less linear. For both ED by LR and ICU by EHR the prediciton was maximal using all variables, but with an important spike (increase and decrease) at 40% of weight for ICU-EHR. On the other hand, the maximal anticipation of discharge by EHR was achieved using around 85% of variable weight, which was chosen to measure occupancy decrease (Fig. 2). Interestingly, predictions higher than the ones obtained by the full set were achieved also by smaller sets, however, the accuracy lowered to levels that would importantly affect readmission rates, if implemented.

Differences between EHR and LR

To further explore the differences between EHR and LR models, intersection variables were compared, i.e. the blood test variables recorded in both the LR and EHR. As expected, almost all variables were reported less in the EHRs, since they were entered only when necessary and missing numerical fields were automatically set to zero (Fig. 2C). However, the overestimation of C-reactive protein (CRP) confirmed that clinicians selectively reported higher determinations or did not always report decreases. On the other hand, there were fewer or no discrepancies in D-dimer and interleukin-6 (IL6) values since they were necessary for the indication of RT and were entered into the EHRs when available. To better assess the significance of missing values, zeroes were replaced using pooled averages by outcome; for example, if ALT (see Fig. 2C) was missing in a subject's state, it was replaced by the ALT value averaged across all subject states labelled with the same outcome. This was done for ED, using both EHR and LR, and for ICU, using EHR. Interestingly, occupancy curves and anticipation rates did not change either for ED or for ICU. However, in all cases, average values boosted specificity, which reached 94.9% for ED by EHR and 94.1% by LR, up from 86.1 and 86.5% respectively; and 97.4% for ICU, up from 89.5%. This resulted in record high ROC-AUCs of 97.5, 97.7 and 98.7% for ED by EHR, ED by LR, and ICU, respectively. These results suggest that underreported values (for EHR) and completely missing values (for LR) should be avoided when possible. On the other hand, when predicting ED using EHR without the

overreported CRP, both specificity and ROC-AUC fell from 86.1 and 93.3% to 83.1 and 91.5%, respectively. In contrast, excluding the CRP variable to calculate ED with LR did not yield noticeable changes —sensitivity decreased from 100 to 99.9%, specificity from 86.5 to 86.4%, and ROC-AUC from 93.3 to 93.2%.

EHR-LR subset results

Anticipation rates and the ROC-AUCs for EHR models did not change significantly when the results were calculated only on LR subjects, which rules out bias due to unbalanced sampling. For ED, sensitivity was 100% and ROC-AUC was 95%. For OSi, sensitivity was 89.5% and ROC-AUC was 77.8%. Finally, for ICU, sensitivity was 92.4% and ROC-AUC was 88.5%.

Variable change over time and across models

A total of 414 RDF models were calculated for ED: 234 from EHR and 180 from LR, while a total of 252 models were built for worsening of outcome, 120 for intensification of OS (60 from EHR), 132 for ICU admission (60 from EHR). The most important variable over time and for both sources was *Age*, which appears as the leading variable for worsening of outcome. The second most important clinical variable was *Days*, which represents the evolution of COVID-19 in days. Among laboratory variables, CRP is of special interest, since it was the only overestimated EHR variable and the best one for predicting ED from EHR. As expected, *PaO2/FiO2* replaced CRP for predicting worsening of outcome (see Table 1). Variables describing oxygenation and the acid-base balance always ranked high in all models built from LR (see supplementary results). The importance of interleukin-6, in contrast to CRP, progressively increased over time and was therefore the leading variable predicting ICU admission from both sources, EHR and LR. At the same time, D-dimer was also very important and fairly constant between model types and over time. Unexpectedly, early clinical deterioration scores (CURB65[14] and MEWS[15]) were never able to explain more than 3.4% of the predictive power of the model, and then only for ED. See supplementary results for a comparison of EHR and LR.

Table 1
Variable weightings in EHR (electronic health record) models according to the different outcomes evaluated. Abbreviations: OS, oxygen support; ICU, intensive care unit, paO2, partial arterial pressure of oxygen, FiO2, fraction of inspired oxygen; LDH, lactate dehydrogenase; WBC, white blood cell; MEWS, modified early warning score; ALT, alanine aminotransferase; AST, aspartate aminotransferase, BNP, brain natriuretic peptide; COPD, chronic obstructive pulmonary disease.

Early Discharge		**Intensification of OS**		**ICU admission**	
variable	**weight**	**variable**	**weight**	**variable**	**weight**
C-reactive protein	12.8% ±0.8%	PaO2FiO2 ratio	10.7% ±0.7%	PaO2FiO2 ratio	13.9% ±1.0%
Age	8.9% ±0.3%	Age	6.7% ±0.8%	Age	6.5% ±0.3%
PaO_2FiO_2 ratio	7.6% ±0.5%	Procalcitonin	5.8% ±0.6%	LDH	5.4% ±0.4%
Days	5.8% ±0.3%	C-reactive protein	5.7% ±0.4%	Interleukin-6	5.1% ±0.5%
D-Dimer	5.6% ±0.4%	LDH	5.6% ±0.4%	D-Dimer	4.8% ±0.4%
LDH	5.3% ±0.5%	D-Dimer	4.6% ±0.2%	C-reactive protein	4.5% ±0.3%
Creatinine	4.7% ±0.1%	Interleukin-6	4.4% ±0.4%	Procalcitonin	4.5% ±0.5%
WBC	3.8% ±0.1%	Days	4.4% ±0.2%	Days	4.1% ±0.2%
Platelets	3.4% ±0.1%	Creatinine	4.0% ±0.5%	Bacterial co-infection	4.0% ±0.6%
MEWS score	3.4% ±0.3%	WBC	3.8% ±0.3%	WBC	3.9% ±0.3%
CURB-65 score	3.3% ±0.4%	Platelets (10^3/μL)	3.3% ±0.2%	Creatinine	3.6% ±0.2%
Interleukin-6	3.0% ±0.5%	Lymphocytes (10^3/μL)	3.1% ±0.1%	Platelets (10^3/μL)	3.0% ±0.2%
Procalcitonin	3.0% ±0.4%	ALT	3.0% ±0.2%	BNP	2.9% ±0.3%
Lymphocytes (10^3/μL)	2.9% ±0.2%	AST	2.9% ±0.1%	Lymphocytes (10^3/μL)	2.8% ±0.1%
Chronic heart disease	2.7% ±0.3%	MEWS score	2.8% ±0.2%	ALT	2.6% ±0.1%
ALT	2.3% ±0.1%	Bacterial co-infection	2.7% ±0.6%	AST	2.6% ±0.1%
BNP	2.3% ±0.2%	BNP	2.7% ±0.3%	MEWS score	2.4% ±0.2%

Early Discharge		Intensification of OS		ICU admission	
AST	2.2% ±0.1%	Platelets/Lymphocytes	2.5% ±0.3%	Ferritin	2.2% ±0.2%
Ferritin	2.1% ±0.2%	Troponin T	2.4% ±0.2%	Troponin T	2.2% ±0.2%
Platelets/Lymphocytes	1.9% ±0.3%	Ferritin	2.2% ±0.2%	Platelets/Lymphocytes	2.2% ±0.2%
Sex	1.9% ±0.2%	CURB-65 score	2.1% ±0.3%	CURB-65 score	2.0% ±0.3%
Bacterial co-infection	1.9% ±0.3%	Sex	1.9% ±0.2%	Chronic heart disease	2.0% ±0.3%
Troponin T	1.8% ±0.1%	Obesity	1.6% ±0.5%	Obesity	1.6% ±0.2%
Chronic renal disease	1.2% ±0.1%	Chronic heart disease	1.3% ±0.2%	Chronic renal disease	1.3% ±0.3%
Diabetes Mellitus	1.1% ±0.1%	Chronic renal disease	1.1% ±0.2%	Sex	1.1% ±0.1%
COPD	1.0% ±0.1%	Diabetes mellitus	1.0% ±0.2%	Diabetes mellitus	1.0% ±0.1%
Obesity	0.9% ±0.1%	Active cancer	0.8% ±0.1%	COPD	0.8% ±0.2%
Active cancer	0.8% ±0.0%	COPD	0.7% ±0.1%	Active cancer	0.7% ±0.1%
Asthma	0.6% ±0.0%	Asthma	0.6% ±0.1%	Autoimmune disease	0.5% ±0.2%
Autoimmune disease	0.5% ±0.0%	Autoimmune disease	0.5% ±0.1%	Asthma	0.5% ±0.1%
CD4 T lymphocytes T CD4	0.2% ±0.0%	CD4 + T lymphocytes	0.1% ±0.0%	CD4 T lymphocytes	0.2% ±0.1%

Discussion

This study presents the highest ROC-AUC results to date across all COVID-19 clinical outcomes considered so far, at least to our knowledge. It also compares preprocessed human data and raw data using the same ML framework and subjects in a complex clinical setting. In order to do this, a powerful ML tool used especially for categorical classification was adapted for a time series analysis that considered patient states over time. This is of special interest, since most of ML analyses are conceived to classify data in categories, but few work with time series. COP-tt introduces this feature by using subject states instead of collapsing patient data into a single file, and by running its analysis on timelines (Fig. 1). This paradigm could be adapted to work not only with random forest classifiers, but with any ML categorical classifier. Moreover, we determined a minimal number of variables able to obtain ROC-AUCs above 80%, which is around 6 considering the present

data. However, the relation between variables and time to event is complex, as demonstrated by both ED and ICU prediction by EHR (Fig. 3). Future studies will address this relationship, studying metehods to determine optimal variable associations. However, LR results differred from EHR, showing more predictable trends. Thus, the differences between LR and EHR were further analyzed. We found, on the one hand, that underreported values did not alter prediction times but significantly lowered specificity and ROC-AUCs and, on the other, that overreported CRP had a greater impact on the prediction of ED calculated on EHR, suggesting that selective reporting by clinicians acted as a filter that increased the differences between patient states and facilitated ML classification. Further studies should be conducted to assess whether human interpretation of the data enhances ML analysis, a concept that could be applied to different clinical problems, both in diagnosis and follow-up. Although in some healtcare systems LR are integrated into EHR, some categorical variables require and represent a human interpretation, therefore the distinction is still relevant.

Regarding the changes in variables over time and their weight in prediction, *Age* led the group in both sources (EHR and LR), as in other studies, followed by arterial oxygenation measurements. In addition to other studies though, temporal analysis allowed us to see the shifts in importance of different variables for making different predictions, for example, proinflammatory CRP, useful for early discharge and the *PaO2/FiO2* ratio, more suitable for predicting clinical worsening and ICU admission.

To our knowledge, the COP-tt framework is the first one that could be successfully used for both early discharge and prediction of severe COVID-19 outcomes. Throughout the pandemic, the increase in new cases of COVID-19 has been associated with higher hospital occupancy and mortality rates. Despite the vaccination campaigns launched, the number of new cases worldwide continues to rise,[16] and there are reports of hospitalizations of vaccinated patients, as well as deaths [17–19]. At the same time, it is also vital to start easing restrictions, since virtually all healthcare systems are so closely linked to the economic welfare of their respective countries that an economic downturn triggered by the collapse of several industrial sectors would most certainly undermine the functioning of healthcare systems. In this context, the use of the ML prediction approach could be a useful tool to improve healthcare resource allocation and could help avoid the collapse of hospitals during pandemics. Simply by using EHR modelling and selective restrictions, which are interventions aimed at reducing infection rates, our hospital occupancy rates due to COVID19 could have decreased by almost 60% (Fig. 2B). These estimates were calculated on the basis that patients with mild respiratory distress could be discharged and monitored from home. Although COP-tt may be an useful tool for early discharge of patients suffering COVID19 infection, we propose that this strategy should be accompanied by a follow-up after discharge (such as monitoring oxygen saturation through a pulse oximeter and/or daily telephone follow-up) to identify potential clinical worsening or complications at home. In this way, the potential risk of the algorithm may be avoided. Even though this was not possible at the beginning of the pandemic, it is now feasible and the application of our classification framework during the next wave of infection could ease hospital pressure. Thus, we would expect a COP-tt approach to be able to indirectly increase the efficacy of other interventions such as vaccination. Although the final goal of the latter is virus eradication, which is admittedly more complex due to SARS-CoV2 variants, one of the primary objectives of vaccination is precisely to lower infection rates and prevent the collapse of the healthcare system. In addition, the impact of vaccination and different vaccines could be easily monitored using COP-tt by adding a model variable.

The major limitation of the study is that it is based on a single center, although the number of analyzed patients matches, and at times even exceeds, that of multicenter studies[12]. The second major problem is its applicability to other health structures. Nevertheless, our study demonstrates that this framework can work using very different, even contrasting, sources, such as EHR and LR. This suggests that other variables or types of records could be used to achieve the same objective. Furthermore, compared to large multicentric models[20], which aim to generalize their predictive power as much as possible, our study and framework are aimed at local usage, which can be customized according to the hospital infrastructure and laboratory capabilities. The framework is available to clinicians and researchers in the GitHub repository, mentioned in Methods, and a new repository, adapted for a clinical trial, is also available and has been updated.

Summarizing, the presented ML framework is not a model, but a model generator for various COVID19 clinical outcome predictions. It can work under very different circumstances, since it successfully processed 31 (EHR) and 120 (LR) variables, achieving similar results, and is highly robust to missing data. The framework in fact transforms missing data into an advantage rather than a limitation, as demonstrated by the highlighted differences between human-reported laboratory results and objective records. Since the framework is a model generator, we expect it to be adaptable to different settings and objectives, such as assessing the clinical impact of SARS-CoV2 variants[21] or emerging clinical variables[22]. Finally, the early hospital discharge should be associated with a follow-up strategies (e.g., in our institution daily calls were implemented after early discharge) to identify clinical worsening at home and avoid risks related to erroneous outcome classification.

Methods

Materials And Methods

Data sources and preprocessing

The Hospital del Mar database separates electronic health records (EHRs), which are clinical history entries, from laboratory results (LRs). Since all records are in virtually constant use, data retrieval errors can occur, especially during peak data usage times, which became longer and more frequent during the epidemic. Data from EHR were retrieved in a single query using the search word “COVID”, from March until August 2020 inclusive. Since LRs could not be searched in the same way, all data were retrieved using time limits only. All patient identifiers were converted into new random identifiers prior to analysis; the conversion index was kept by JPH and MLS, who were not directly involved in data processing and analysis. EHRs were further processed using keywords and keyword proximity to retrieve numeric or binary variables. Certain clinical conditions, such as cardiovascular diseases, were recorded as binary variables, while numerical variables were age and LR transcriptions. All EHR variables were located using lists of synonyms; proximity (distance from the located keyword) was used to analyze numerical values. Outcomes were retrieved in the same way. To assign outcomes to LRs, EHR data were crosschecked with LR data. Confirmation of SARS-CoV2 infection was determined either by LR or EHR, giving preference to LRs.

Tables were constructed from raw EHR and LR data, where each row corresponded to a patient initialization or update. For each patient therefore one or more rows were compiled. The initialization or update was located in time using the date of entry. Default values were used if any variable could not be retrieved from the entry at a specific time: 0.0 for numerical values; “NO” for binary values. To assess the significance of missing values in EHR, 0.0 were replaced by values averaged by pooling states per outcome class. The special variable, *sex*, was treated as binary, referring only to the biological phenotype. If no assessment of sex could be retrieved, the subject was discarded. Tables were saved as comma-separated value (CSV) files, using “;” as a separator.

Data variables

During the COVID-19 pandemic, physicians from different specialties attended COVID-19 patients in our hospital. In order to homogenize patient management, guidelines were drawn up by the Infectious Diseases Service for data collection in the EHR as well as therapeutic management. EHR variables were selected from the clinical history recording guidelines issued by the COVID-19 task force of our hospital and the clinical experience of the authors involved in COVID-19 shifts, SG-Z, ILM and AP. Demographic, clinical, and epidemiological data collected from the EHR included the following: age, sex, and race, underlying diseases, clinical and outcome data of COVID-19, complementary tests (kidney and liver profile, electrolytes, myocardial enzymes, blood count, coagulation profile, including D-dimer, and inflammatory markers, such as serum interleukin-6, IL6, ferritin and C-reactive protein). On the other hand, LR variables were selected from among the 20% most requested blood tests during the time range specified. For a complete list of the EHR and LR variables, see the variable weight tables in Results and Supplementary Results.

The need for rescue therapy was defined as the addition of IL-6 inhibitors such as *tocilizumab* or *sarilumab* to standard therapy in patients with clinical deterioration (increased pulmonary infiltrates or inflammatory markers and/or changes in the need for oxygen or ventilatory support). With respect to respiratory needs, ventilatory support was defined as the need for non-invasive mechanical ventilation (MV), high-flow nasal cannula or invasive MV. Increased oxygen support was considered during hospital admission when it was necessary to increase oxygen needs (to increase FiO2 using a nasal cannula or mask and/or ventilatory support was needed).

COVID-19 was defined as a SARS-CoV-2 infection confirmed by quantitative reverse transcriptase–polymerase chain reaction, rapid diagnostic tests based on antigen detection in a nasopharyngeal sample, or serology. Those patients with a negative test but who met the clinical diagnostic criteria: respiratory symptoms (dyspnea, cough, sore throat, changes in taste/smell), chest X-ray findings (unilateral or bilateral interstitial infiltrates), and laboratory abnormalities (lymphopenia, increase in C-reactive protein, ferritin, IL-6 or D-dimer), which made COVID-19 the most likely diagnosis in the current epidemiological situation, were considered as probable infection and processed for anticipation of early discharge.

Prediction of clinical outcomes over time (COP-tt)

In the first step, the framework uploads CSV tables to create an internal database, in which subject initialization or updates are stored as states. For the states, variables are either numerical or categorical, with binaries being a subset of categorical variables. When a variable is categorical, COP-tt reads all possible text

entries and automatically assigns them a numerical counterpart, which is used during analysis. Automatic categorical coding was not used in this study. The variable *outcome*, despite being categorical, was not assigned automatically but was defined as shown in Fig. 1 of the main text. Analysis consisted of three steps: data sampling, model training and outcome prediction. For the latter, the outcome of each state was upgraded to the worst outcome recorded for the subject. This upgrade can be disabled to create models for outcome detection rather than prediction (see supplementary results).

For data sampling, the database is split into a training set and a test set. The number of subjects in each set is defined by the training set ratio, which is arbitrary. The ratio in the current study varied, depending on the type of prediction and availability of patients (60 to 90%). Generally speaking, the larger and more varied the training set, the more accurate the prediction. Individual subjects' data, however, was never split, to avoid the common pitfall to predict the past from the future.

Several random decision forest (RDF)[23] models were trained to simulate k-fold cross-validation[24], widely used in machine learning (ML). Before training a specific model, the training set was split into training and test subsets, which were checked to ensure a balanced number of outcomes (50% good). An error threshold was used to avoid unbalanced sampling, since it is not always possible to guarantee well balanced sets. When the training set was unbalanced beyond the arbitrary error threshold, the sampling was repeated. In this study, six RDF models were trained prior to prediction of outcome. The number of RDF trees can be set before analysis.

During the last step, each RDF model classifies all states of a test subject. Since 0 is used for good and 1 for bad outcome, the predicted outcome is defined as bad when the average model prediction is equal to or greater than a threshold (0.5 in this study); hence, if three models predict a bad outcome and three a good one, the final prediction is bad. The prediction is over time because it is calculated at each state, after which the anticipated time before the target outcome is determined. So, for example, for calculating anticipation of intensive care unit (ICU) admission, the final prediction of a bad outcome can be made at a state preceding actual admission to the ICU, which counts as a real prediction; or at the state of admission, which corresponds to detection; or after admission to the ICU, which is failure to detect or anticipate. To calculate the time from prediction, the first state associated with the outcome of interest was used as time reference. Since all states were associated with a date, it was straightforward to calculate the time from the reference state in days. In this particular case, patients not admitted to the ICU did not count for the final statistics for anticipation. Target outcomes and hence outcome discriminators can be set before analysis. The bad versus good discriminators used in this study are shown in Fig. 1 of the main text.

The current version of the framework is written in Python (version 3.8.5), using the Miniconda data science platform (version 4.9.2). The scikit-learn module (version 0.23.2) was used for RDF implementation. COP-tt accepts CSV tables and text files as lists of variable names and types. See the code and instructions in the GitHub repository: joricomico/COVID19_cop: COP-tt test, application to COVID19 (github.com).

Variable comparison between EHR and LR

Because of technical problems in retrieval, LRs represented a smaller and more complete subset of EHRs. Therefore, the Student's t-test or Mann-Whitney U test were used to compare random samples of EHR and LR intersection variables (variables present in both sets), depending on distributions, evaluated with the Shapiro-

Wilk test of normality. One hundred samples of 1000 values were compared for each variable. Average p-values and the percentage of times p was < 0.05 were used to determine statistical significance.

To assess the significance of missing values, automatic zeros were updated using averaged values per outcome. To obtain averaged values, patient states were grouped by respiratory outcome, and values from each state greater than zero were pooled. This was done only for the laboratory tests, either from EHR or LR. Clinical variables were not averaged because they indicated the presence or absence of a condition.

Weighting of variables and reduced set models

RDF models are ensembles of decision trees which fit a random subset of training data. A split variable (a threshold or threshold vector) in a decision tree defines how well a variable can discriminate between two or more outcome groups by setting a threshold that divides such outcomes. The importance of the variable is determined through the Gini index (Gi): $Gi = 1 - \sum_{v=1}^{n} (P_v)^2$, in our case $Gi = [(P_0)^2 + (P_1)^2]$, where $P_{0|1}$ are the probabilities of the good and bad outcomes, respectively. The Gi is weighted by the occurrence of classes, which means that their probability is multiplied by the class ratio. For instance, in case there were 2 good and 3 bad outcomes: $WGi = \left[\frac{2}{5}(P_0)^2 + \frac{3}{5}(P_1)^2\right]$. Using the discriminatory power of each sub-model (a decision tree), variable weights can then be evaluated as the average discriminatory power of each variable to classify target outcomes. To assess the global variable weights, variable weights were calculated for each RDF model and average weights were then calculated across all models.

Variable weights were then used to classify variables by importance, which in turn were used to build reduced set models, without changing the training procedure.

Ethics

Data collection, statistical and machine learning data treatments were performed in accordance with the Declaration of Helsinki. The study was approved by the Clinical Research Ethics Committee of the Parc de Salut Mar (register no 2020/9329/I). Data anonymization was carried out in order to protect the privacy of patients. The need for written informed consent was waived due to the observational nature of the study and the retrospective analysis.

Declarations

Acknowledgments: We would like to thank Janet Dawson for her help in revising the English-language manuscript. The authors would like to thank the Spanish Society of Infectious Diseases and Clinical Microbiology (SEIMC) which supports Dr. Silvia Gómez-Zorrilla's research.

Author contributions:

Conceptualization: AP, SG-Z, ILM, JPH, MLS

Methodology: AP, PP

Investigation: AP, SG-Z, ILM, PP, EP

Visualization: AP

Project administration: JMR, AM, JPH, MLS

Supervision: SG, JPH, MLS

Writing – original draft: AP, SG-Z, ILM

Writing – review & editing: EP, PP, MLS, RGF, SG, AM, JMR, JPH

Competing interests: Santiago Grau has received fees as speaker for Pfizer, Angelini, Kern and MSD and research grants from Astellas Pharma and Pfizer. Juan Pablo Horcajada reports being a speaker with honoraria on advisory boards for MSD, Pfizer, Angelini, and Menarini, and has a research grant from MSD. The authors report no conflicts of interest in this work.

Data and materials availability: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request. All code is available in GitHub (see Supplementary Materials).

References

1. Alderson, J. *et al.* Overview of approved and upcoming vaccines for SARS-CoV-2: a living review. Oxford Open Immunol. **2**, (2021).
2. Lalmuanawma, S., Hussain, J. & Chhakchhuak, L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos, Solitons and Fractals **139**, 110059 (2020).
3. Albahli, S. Efficient gan-based chest radiographs (CXR) augmentation to diagnose coronavirus disease pneumonia. Int. J. Med. Sci. **17**, 1439–1448 (2020).
4. Wang, S. *et al.* A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur. Respir. J. **56**, (2020).
5. Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N. & Mohi Ud Din, M. Machine learning based approaches for detecting COVID-19 using clinical text data. Int. J. Inf. Technol. **12**, 731–739 (2020).
6. Wu, G. *et al.* A prediction model of outcome of SARS-CoV-2 pneumonia based on laboratory findings. Sci. Rep. **10**, 1–9 (2020).
7. Li, X. *et al.* Deep learning prediction of likelihood of ICU admission and mortality in COVID-19 patients using clinical variables. PeerJ **8**, 1–19 (2020).
8. Cheng, F.-Y. *et al.* Using Machine Learning to Predict ICU Transfer in Hospitalized COVID-19 Patients. J. Clin. Med. **9**, 1668 (2020).
9. Wu, G. *et al.* Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: An international multicentre study. Eur. Respir. J. **56**, (2020).
10. Assaf, D. *et al.* Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Intern. Emerg. Med. **15**, 1435–1443 (2020).

11. Das, A. K., Mishra, S. & Gopalan, S. S. Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool. PeerJ **8**, 1–12 (2020).
12. Liang, W. *et al.* Early triage of critically ill COVID-19 patients using deep learning. Nat. Commun. **11**, 1–7 (2020).
13. Arvind, V., Kim, J. S., Cho, B. H., Geng, E. & Cho, S. K. Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19. J. Crit. Care **62**, 25–30 (2021).
14. Lim, W. S. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. Thorax **58**, 377–382 (2003).
15. Subbe, C. P. Validation of a modified Early Warning Score in medical admissions. QJM **94**, 521–526 (2001).
16. Data as received by WHO from national authorities. *Weekly epidemiological update on COVID-19–21 September 2021*. *WHO Epidemilogical Update* vol. 21 Septemb https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19—21-september-2021 (2021).
17. Keehner, J. *et al.* SARS-CoV-2 Infection after Vaccination in Health Care Workers in California. N. Engl. J. Med. **384**, 1774–1775 (2021).
18. Haas, E. J. *et al.* Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. Lancet **397**, 1819–1829 (2021).
19. Juthani, P. V *et al.* Hospitalisation among vaccine breakthrough COVID-19 infections. Lancet Infect. Dis. (2021) doi:10.1016/S1473-3099(21)00558-2.
20. Schwab, P. *et al.* Real-time prediction of COVID-19 related mortality using electronic health records. Nat. Commun. **12**, (2021).
21. Tse, H. *et al.* Emergence of a Severe Acute Respiratory Syndrome Coronavirus 2 virus variant with novel genomic architecture in Hong Kong. Clin. Infect. Dis. (2021) doi:10.1093/cid/ciab198.
22. Marfia, G. *et al.* Decreased serum level of sphingosine-1-phosphate: a novel predictor of clinical severity in COVID-19. EMBO Mol. Med. **13**, (2021).
23. Tin Kam Ho. Random decision forests. in *Proceedings of 3rd International Conference on Document Analysis and Recognition* vol. 1 278–282 (IEEE Comput. Soc. Press).
24. Yan-Shi Dong & Ke-Song Han. A comparison of several ensemble methods for text categorization. in *IEEE International Conference onServices Computing*, 2004. *(SCC 2004). Proceedings. 2004* 419–422 (IEEE). doi:10.1109/SCC.2004.1358033.

Figures

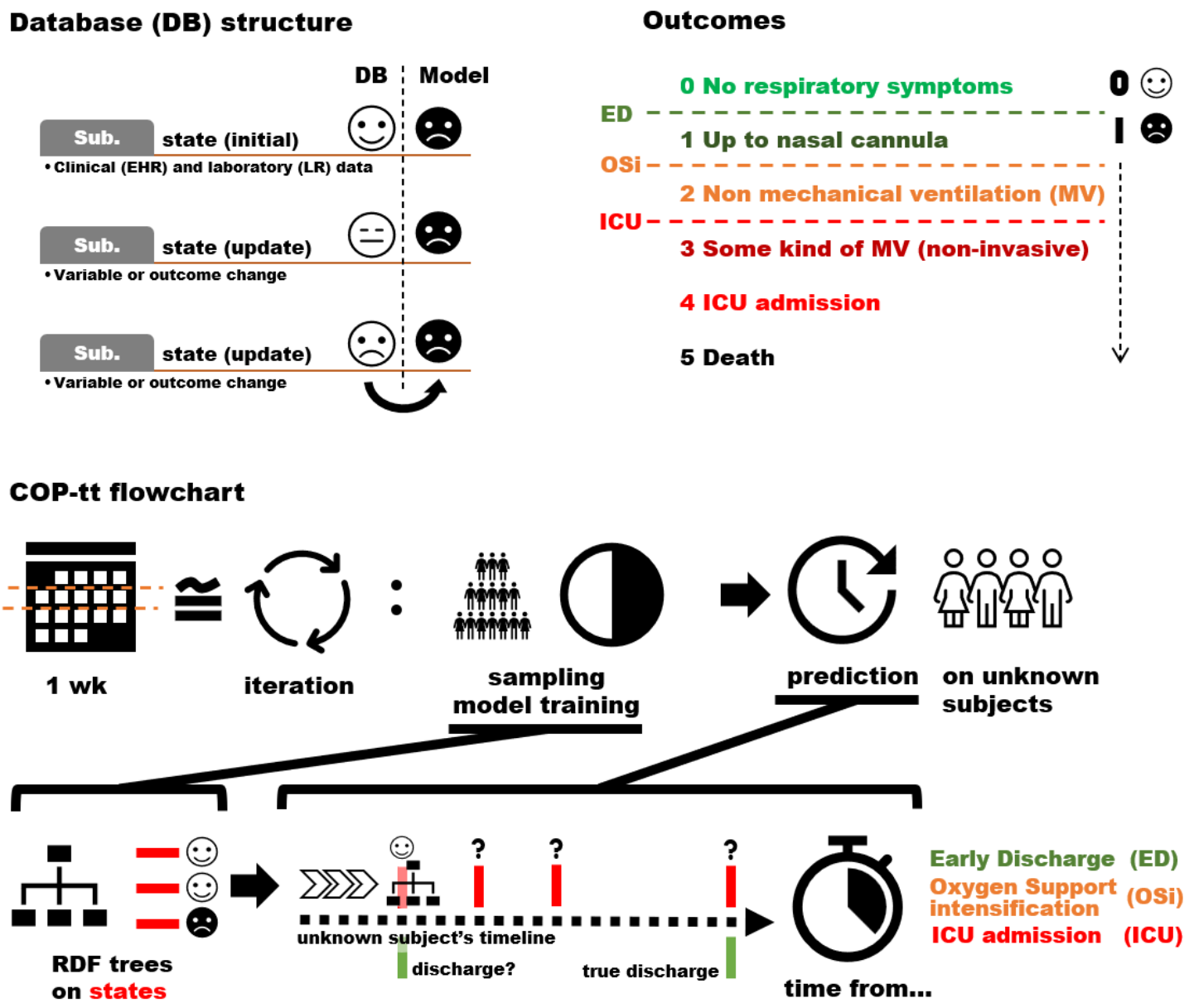

Figure 1

COP-tt database structure, clinical outcomes, and flowchart. Database structure: each subject is represented by a sequence of states reflecting disease progression; states are linked to worst outcomes to make predictions. **Outcomes** range from no respiratory symptoms to death. Since predictions were made using a binary classifier, *the binary discriminator between good (0) and bad (1) outcomes was lowered, depending on the clinical prediction goal,* from early discharge to admission to ICU. **COP-tt flowchart**: iterations could be set within any temporal constraints; here we imagine weekly iterations of sampling and predictions, which in a prospective study would be training models from all or a sample of past subjects and make predictions for new subjects in the selected time span. The models are assemblies of *random decision forest (RDF) trees*; six models were used to reduce the risk of overfitting. RDF results were averaged, and the outcome on unknown subjects was determined over time. Whenever an event of interest, e.g., discharge, was classified, the distance between the real and the predicted event was calculated. See methods for a detailed description.

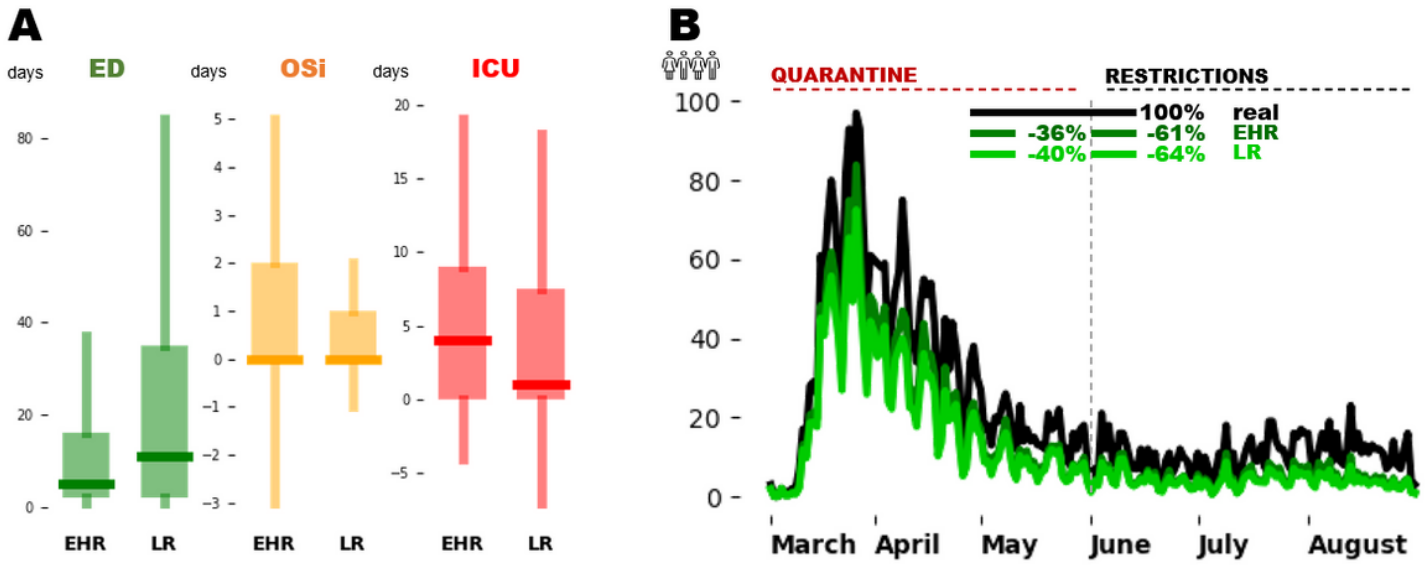

Figure 2

Anticipation, hospital occupancy and model comparison. A. Anticipation (in days) of early discharge (ED), oxygen support intensification (OSi) and intensive care unit admission (ICU). **B.** Occupancy curve adjusted for electronic health record (EHR) or laboratory results (LR) models.

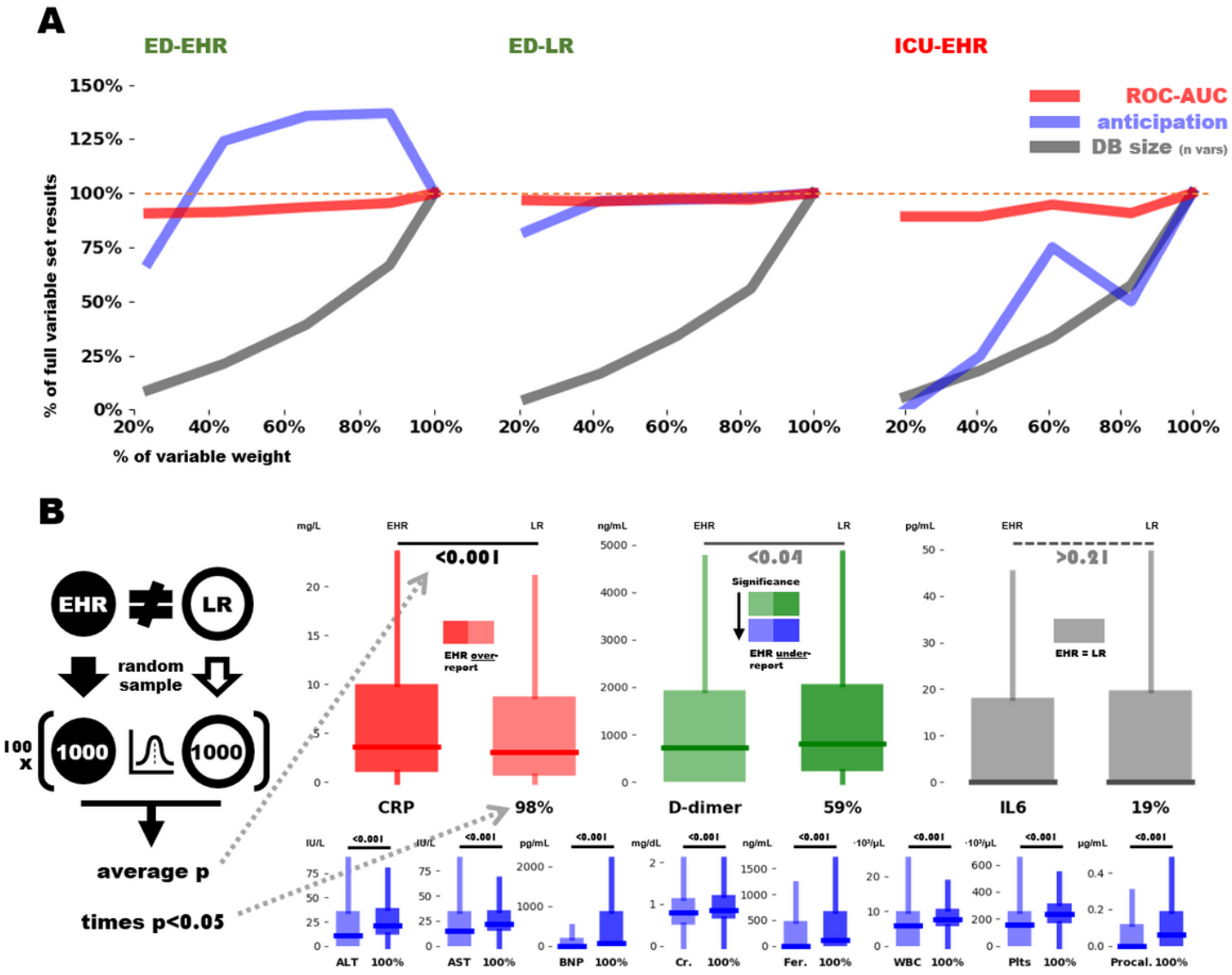

Figure 3

Variable weight related to accuracy and prediction time, comparison between EHR and LR. A. Effect of variable reduction on database size, anticipation time and accuracy (ROC-AUCs): models were reduced using variable weight, which is the average variable weight among decision trees. ROC-AUCs, anticipation and sizes were all compared to the results obtained with the full set of variables. **B.** Comparison of EHR and LR; since the sets did not fully overlap, random sampling was used to determine statistical differences. Alanine (ALT) and Aspartate aminotransferases (AST), brain natriuretic peptide (BNP), creatinine (Cr.), ferritin (Fer.), procalcitonin (Procal.), white blood cells (WBC) and platelets (Plts) were underreported, while proinflammatory CRP was overreported in EHR.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- COVIDopSciRepv2supp2R.docx