Derivation of a Cost-Sensitive COVID-19 Mortality Risk Indicator Using a Multistart Framework

1st Rubén Armañanzas
Data Science Area
Basque Center for Applied Mathematics
Bilbao, Spain
0000-0003-4049-0000

2nd Adrián Díaz
Data Science Area
Basque Center for Applied Mathematics
Bilbao, Spain
0000-0002-2876-6177

3rd Mario Martínez-García
Data Science Area
Basque Center for Applied Mathematics
Bilbao, Spain
0000-0002-6849-6239

4th Santiago Mazuelas
Data Science Area
Basque Center for Applied Mathematics
IKERBASQUE Basque Foundation for Science
Bilbao, Spain
0000-0002-6608-8581

***Abstract*—The overall global death rate for COVID-19 patients has escalated to 2.13% after more than a year of worldwide spread. Despite strong research on the infection pathogenesis, the molecular mechanisms involved in a fatal course are still poorly understood. Machine learning constitutes a perfect tool to develop algorithms for predicting a patient's hospitalization outcome at triage. This paper presents a probabilistic model, referred to as a mortality risk indicator, able to assess the risk of a fatal outcome for new patients. The derivation of the model was done over a database of 2,547 patients from the first COVID-19 wave in Spain. Model learning was tackled through a five multistart configuration that guaranteed good generalization power and low variance error estimators. The training algorithm made use of a class weighting correction to account for the mortality class imbalance and two regularization learners, logistic and lasso regressors. Outcome probabilities were adjusted to obtain cost-sensitive predictions by minimizing the type II error. Our mortality indicator returns both a binary outcome and a three-stage mortality risk level. The estimated AUC across multistarts reaches an average of 0.907. At the optimal cutoff for the binary outcome, the model attains an average sensitivity of 0.898, with a 0.745 specificity. An independent set of 121 patients later released from the same consortium attained perfect sensitivity (1), with a 0.759 specificity when predicted by our model. Best performance for the indicator is achieved when the prediction's time horizon is within two weeks since admission to hospital. In addition to a strong predictive performance, the set of selected features highlights the relevance of several underrated molecules in COVID-19 research, such as blood eosinophils, bilirubin, and urea levels.**

***Index Terms*—COVID-19, Cost-Sensitive Prognosis, Clinical Indicator, Multistart Estimation**

I. Introduction

The capacity to early and accurately identify patients at risk of death has become an urgent yet challenging necessity for the COVID-19 triage at hospital admission. Artificial intelligence through machine learning (ML) constitutes the perfect tool to tackle this problem. ML allows the construction of algorithms for predicting the hospitalization outcome at the point of care based on traditional and widely spread clinical tests and epidemiological inputs.

The World Health Organization's weekly report on August 3^{rd}, 2021, presented an accumulated total of more than 197 Million worldwide cases of COVID-19, with over 4.2 Million associated deaths [1]. After a year since the pandemic started spreading across the globe, the estimated overall death rate accounts therefore for 2.13% of COVID-19 patients. Initial death figures were however much more elevated, with an estimated 13.8-19.1% death rate during the initial outbreak at Wuhan, China (192.6 per 100,000 on the general population) [2]. In the United Kingdom, death rates recorded during the same period by the National Health Service rose to 32.2% among infected patients [3]. It became clear very soon that the seek for potent biomarkers to timely predict COVID-19 infection outcome was and still is an essential field of research. Since then, several papers so far have reported such biomarkers of COVID-19 outcome, both for disease prognosis, and for risk evaluation [4]. In all cases, these biomarker molecules are assessed individually and never mixed in a multivariate fashion.

Another viewpoint to predict the infection's severity or risk of death is to infer machine learning models using general epidemiological descriptors of the infected patients [5], [6]. This approach ignores any routine biochemical test or biomarker of interest. It neither provides clinical guidelines to personalize the risk of the infection for new patients at time of admission. Although being a valuable asset to understand the dynamics of COVID-19 spread, their potential for clinical adoption at the point of care is limited. Other more sophisticated mathematical methods including epidemiological and clinical features to predict COVID-19 outcome are regularly surveyed every six months by Wynants et al. [7]. Out of this papers corpus, two works stand out in the prognosis of COVID-19 patients,

This work was funded by the AXA Research Fund project XXX, by the Basque Government special funding on Mathematical Modelling Applied to Health, and by the Ikerbasque Basque Foundatino for Science.

namely Yan et al. [8] and Knight et al. [3].

Yan et al. used hospital health records from Tongji Hospital in Wuhan collected between January and February, 2020. The database includes biochemical test results for 485 infected patients and aims at identify crucial predictive biomarkers of mortality risk [8]. Most patients had multiple blood samples taken throughout their stay in hospital. However, the model training and testing use only data records from the last day in hospital. This modelization cannot be considered an actual prognosis at triage because the model is learned using patients data captured just one day before discharge. The authors try to mitigate the issue by testing the resulting model on all available patients from the prior 10 days.

The work by Knight et al. [3] constitutes the largest effort to build a prognostic indicator based on routinely adopted clinical tests. The corpus of data is split into 35,463 patients for training and 22,361 for validating purposes. It uses a lasso regression model after a filtering process using a generalized additive model to filter irrelevant features. The final machine learning model is translated into a classical clinical additive index by imposing weights to different cutoffs in a set of 8 predictor variables. This index behaves well when ruling out mortality (92.5% sensitivity), however, it has to pay the price of being largely unspecific when ruling in mortality risk (38.6% specificity).

Current and past major works on COVID-19 mortality or deterioration through machine learning techniques use the traditional validation scheme in which one train & test split is used to estimate the performance of their models [3], [8], [9]. Although extensively used in biomedical domains, this estimator, also known as single hold-out, produces biased estimations of the true performance [10]. This bias comes both from the effect of a fixed divide between train & test sets, and also from a single random split not being representative of the original data distribution.

In contrast, this paper presents a COVID-19 mortality risk indicator that avoids the aforementioned limitations. Specifically, our algorithm was induced over routinely performed blood tests and clinical data available at patient's triage. We developed a training pipeline that uses five different multistart configurations from the original dataset to achieve a high degree of generalization. Two different L1-penalized regression methods helped find a consensus subset of relevant features across multistarts. The model training used class imbalance corrections, and the cutoff outcome probability was adjusted in a cost-sensitive manner to achieve both high sensitivity and specificity. A temporal analysis showed a time horizon of two weeks for the most effective indication of risk. Lastly, we compared the performance of our indicator versus those from Knight et al. [3] and Yan et al. [8], and validated their predictive algorithms in our own data.

The paper is organized as follows. Section II presents our working dataset and the curation process to produce a classical ML dataset. Next, Section III covers the development of the prediction model and our scale for mortality risk. Results and conclusions are finally discussed in Sections IV and V. An interactive COVID-19 risk calculator using our model will be available online upon acceptance of this manuscript.

II. Clinical Dataset

The initial bulk of clinical data was released from the HM hospitals in Spain on July 20^{th} (2020) through the Covid Data Save Lives project [11]. Access to the data is controlled by the source institution. The dataset is comprised by information retrieved from the electronic health records of 2,547 COVID-19 confirmed patients treated during the first wave of the pandemic (December 26^{th} (2019) to June 10^{th} (2020)). Besides the clinical and epidemiological information, the dataset includes results for each biochemical test a patient underwent during their inhospital stay. Only values within a week from admission were used to derive our dataset since our main goal was to develop an early prognosis algorithm.

In a preliminary inspection, we found a total of 530 distinct spellings for cellular and chemical tests. Out of these 530 tests, only 200 included systematic values in the database, suggesting the difference was due to human annotation or non-widely approved individual tests. Original records in the database were unstructured mixing text and numerical values, e.g. entries like 'Positive (>3.0)'. Due to the impossibility to properly filter these cases, they were marked as missing values. Finally, after filtering by missing values rates, only 36 blood tests with a percentage of presence beyond 70% were retained as possible model features. For comorbidities the following relevant conditions were identified for each patient: Chronic Cardiac Disease, Chronic Respiratory Disease (including asthma), Chronic Renal Disease, Mild to Sever Liver Disease, Dementia, Chronic Neurological Conditions, Connective Tissue Disease, HIV or AIDS, Cancer, Obesity. For symptoms, we checked for Dyspnea, Fatigue, Lost of Consciousness, Myalgia, Sputum, Anosmia, Fever, Diarrhea, Vomiting, and Cough.

TABLE I
DEMOGRAPHIC CHARACTERISTICS FOR THE FINALIZED DATASET TO PREDICT IN HOSPITAL MORTALITY. TIME SPAN COVERS FROM DECEMBER 26^{th} (2019) TO JUNE 10^{th} (2020). VALUES FOR THE INTERQUARTILE RANGE (IQR) SHOW THE 25^{th} AND 75^{th} DATA PERCENTILES.

Characteristic	No. of patients (%)	Total No.
Inpatient mortality	276 (15.35%)	1798
Ventilator use	1035 (57.56%)	1798
Female patients	707 (39.32%)	1798
Male patients	1091 (60.68%)	1798
ICU admission	167 (9.29%)	1798
	Mean $\pm$ Std.	Median [IQR]
Age (years)	67.79 $\pm$ 15.67	69 [57,80]
No. of selected comorbidities	0.49 $\pm$ 0.77	0 [0,1]
ICU stay (days)	8.72 $\pm$ 10.50	5 [1,12]
Oxygen saturation	94.67 $\pm$ 4.81	95 [94,97]
Heart rate (bpm)	79.28 $\pm$ 14.75	78 [70,88]

After a detailed data inspection, it was noted that the distribution of missing values across patients followed a block structure of repeated misses. Some patients showed as much as 17 missing tests individually. We mitigated this loss of information by removing those patients in which more than

3 tests were lost. The remaining lost values were afterward imputed by unsupervised similarity (5 neighbors) [12]. A total of 1798 patients were finally retained (see Table I for population descriptors). A data dictionary on the initial 36 features and patient identifiers for each multistart are provided as supplementary material.

A second dataset including 1949 new patients was later released on April 19^{th} (2021) with admission dates up to the end of February, 2021. Most of these patients showed empty test records and we recovered only 121 fully formed instances. This set of 121 unseen patients was used as an independent hold-out group to validate the performance estimations retrieved during model training.

III. Derivation of the Mortality Indicator

A. Multistart Configuration

A high variance in performance estimation leads to unstable, not reproducible results and it may lead to inconsistent conclusions when comparing to other models. Our approach to overcome such inaccuracies was to set a multistart scenario for all the numerical experiments [13]. The multistart scenario is even more powerful when preserving the class proportions of the whole dataset into each one of the subsets. This stratified multistart helps reduce dataset shift and improves estimation stability, reducing the overall variance in the process [10]. We therefore produced 5 multistart datasets by stratified random sampling the initial data.

Each multistart configuration was afterwards split into a train & test sets, with 80% and 20% of each multistart data. These percentages were chosen to mimic a single split of a 5-fold cross-validation procedure. Internal estimations of hyperparameters were carried out solely within each train set using a 5-fold inner cross-validation [14]. The externally unseen test sets were eventually used to estimate performance figures over each multistart configuration.

B. Feature Subset Selection

The first stage of our preprocessing pipeline comprised the selection of relevant blood tests features using the multivariate recursive feature elimination [15] method over each of our five multistart configurations. This process ensures the selection of highly predictive and non-redundant feature subsets. The feature selection was computed using the values of the 36 blood test kept after data cleaning.

A majority voting policy returned that out of the 36 initial tests, 34 of them were concurrently selected in the optimal subset of each multistart. In addition, six clinical descriptors were included in further analyses: age, sex, heart rate, oxygen saturation, number of comorbidities, and number of symptoms. Comorbidities and symptoms were encoded as ordinal four-valued scales, where 0 maps total absence, and 3 maps three or more occurrences.

To determine the final set of relevant features to retain out of the 40 intermediate features, we relied on a consensus over the regression coefficients of logistic and lasso regressions computed over each of the multistart training sets. We kept all those features for which the regression coefficients were above a regularization cutoff of 10^{-3} in at least four of the five multistarts for either regressions. This estimation was carried out on an inner 5-fold cross-validations on each of the multistart configurations. A total of 19 features were eventually withheld.

C. Supervised Learning

Derivation of the mortality indicator followed a binary supervised classification framework in which a function ϕ is induced from the data: $\phi : (x_1, \ldots, x_n) \rightarrow \{0, 1\}$, where $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{R}^n$ conforms the observation and $\{0, 1\}$ are the possible values for the supervised variable C.

Logistic regression was chosen as the learning technique to derive our mortality indicator. Briefly, a logistic model returns the probability p that $C = 1$ as

$$p_{C=1} = S(\beta_0 + \sum_{i=1}^{n} \beta_i x_i), \tag{1}$$

where S is the sigmoid function. β_0 is commonly referred to as the intercept of the model, whereas each β_i are known as the predictors' coefficients.

Values for each coefficient were estimated from a database as a regression for the logarithm of the odds, which maps the $logit$ of the probability for class 1. Regression estimators were computed via coordinate descent with an L1-norm penalization. The intercept was nullify. In addition, we compensated for the class imbalance by inserting class weights proportional to the relative frequencies of each class within the training set. The final set of features and coefficients learned from data constitute our consensus model.

D. Cost-Sensitive Calibration

The calibration procedure refers to the determination of an optimal probability cutoff in order to tamper class membership and minimize the number of false negatives [16]. In general, membership to a class is assigned as the class that maximizes the output probability returned by a trained model ϕ. For a binary outcome with $C \in \{0, 1\}$, this membership assignment is calculated based on a 0.5 probability cutoff.

Given the nature of the medical problem at hand, it is desirable to minimize the number of cases in which a patient with bad outcome (mortality) is predicted as likely to survive. This minimization of the false negative cases comes with a price, i.e., an increase in false positives due to the reduction in prediction specificity. In an effort to set a general and cost-sensitive cutoff, this process was done using the posterior probabilities of the outcomes for each of our multistart configurations.

IV. Results

A. Mortality Prognosis

The set of variables included in the mortality predictor are detailed in Figure 1. It displays the SHAP plot [17] when the consensus model predicts the mortality outcome of all cases in

TABLE II

PERFORMANCE ESTIMATION OF THE PROGNOSTIC BINARY OUTCOME BASED ON A 5 MULTISTART TRAIN (80%) & TEST (20%) VALIDATIONS. CONFUSION MATRICES AND FIGURES OF MERIT ARE COMPUTED USING A PROBABILITY OF 0.4 AS CLASSIFICATION THRESHOLD FOR ASSIGNING A POSITIVE OUTCOME.

Prediction	Death	Surv.	Death	Surv.	Death	Surv.	Death	Surv.	Death	Surv.
Positive	51	81	49	73	49	63	49	86	49	85
Negative	4	224	6	232	6	242	6	219	6	220
AUC	0.922		0.916		0.915		0.895		0.886	
Accuracy	0.764		0.780		0.808		0.744		0.747	
Sensitivity	0.927		0.891		0.891		0.891		0.891	
Specificity	0.734		0.761		0.793		0.718		0.721	
F_1 Score	0.545		0.553		0.587		0.516		0.519	

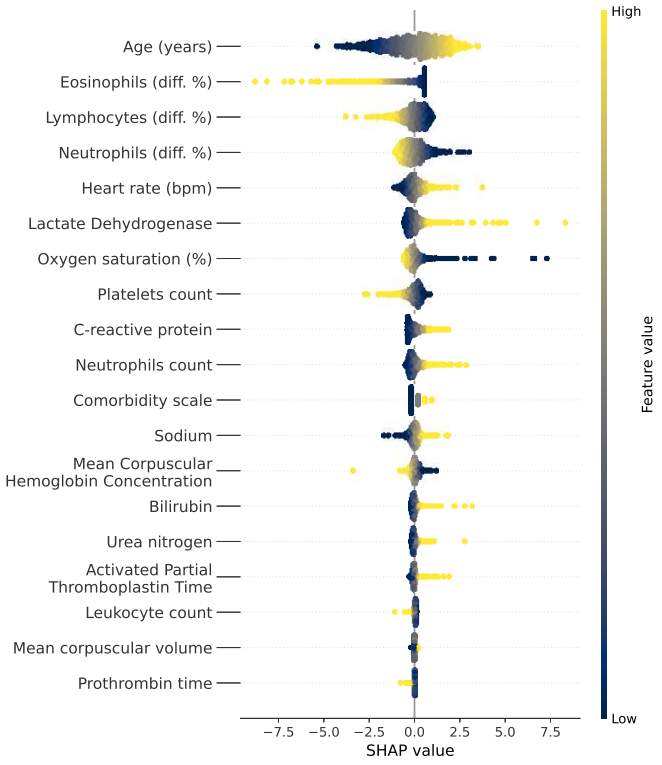


Fig. 1. Final set of 19 features included in the prediction model. SHAP values represent each feature relevance to change the model's output, i.e., the impact of high or low value can have in the final outcome.

the dataset. Blue and yellow values in the plot correspond to low and high values for each predictor, respectively. Positive deviation in SHAP values is a measure of strength towards mortality outcome, whereas negative SHAP values indicate strength in favor of survival.

The model calibration was tackled by the analysis of the output probabilities assigned to each test case for the five multistarts. All test cases (displayed in Figure 2) were segmented into the four possible categories, namely true and false, positive and negative predictions, respectively, based on a probability of 0.5 as positive class cutoff. It is clear that true negatives and positives show distinct probability densities with average median values on 0.11 and 0.85. False positives are concentrated around a probability of 0.64, whereas the few false negatives showed a high variance with an average median on 0.34.

Figure 2 provides clues on how the training process pushed the models to minimize false negatives at the cost of more false positives. In clinical terms this is a desired effect: avoid as much as possible a negative prediction when the patient could be at real risk. To even push this tendency forward, we inspect the individual probabilities for all the false negatives and discover that almost 40% of them had a probability in the interval [0.4, 0.5). We then decide to rescue those by lowering the probability cutoff for the positive class to be 0.4 instead of the natural 0.5. The trade-off for gaining 40% of false negatives by sliding down this probability was to err on a 9% of originally labeled true negatives as new false positives.

The quantitative performance for each multistart model based on a 0.4 cutoff are presented in Table II, including areas under the ROC curve and individual confusion matrices. The estimated accuracy of the mortality predictor is of 0.769 ± 0.026, with 0.898 ± 0.016 sensitivity, and 0.745 ± 0.031 specificity. Individual ROC plots are also available as supplementary material.

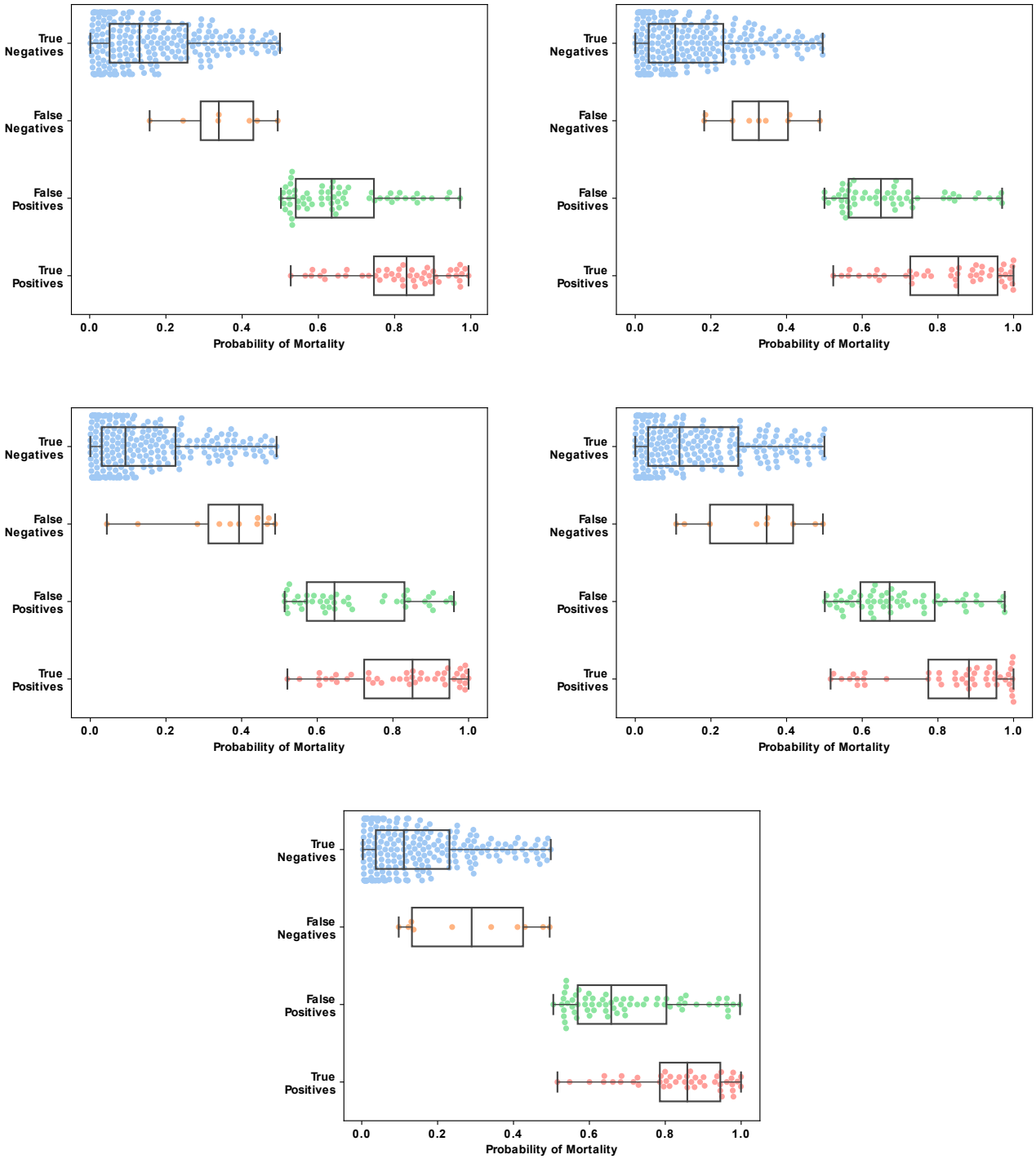
The overall performance estimation obtained by the multistart configuration was validated when using the model to predict the outcome of the 121 patients kept as hold-out group. Confusion matrix and performance figures are presented in Table III. There was no false positive in this case, showing how the calibration process fulfilled the overall aim of not missing any deadly infection. On the contrary side, the specificity almost reached to a 76%, indicating that less than 25% of survivors were warned of a possible serious outcome.

TABLE III

PERFORMANCE OF THE MORTALITY RISK PREDICTOR TESTED ON THE INDEPENDENT HOLD-OUT DATA.

Prediction	Death	Survival
Positive	13	26
Negative	0	82
AUC	0.880	
Accuracy	0.785	
Sensitivity	1.000	
Specificity	0.759	
F_1 Score	0.5	

Fig. 2. Categorical beeswarm boxplots of the predictions' probability for each of the five train & test multistarts. Note that the predictions are segmented based on a probability cutoff of 0.5, i.e., without any cost-sensitive calibration.



B. Mortality Risk Assessment

Output probabilities in Figure 2 for true negative and positive predictions display a clear step wise structure. Such steps suggest that the derivation of a mapping between the mortality predictor's output and a categorical risk scale is possible. To derive such scale, we computed the average 75^{th} percentile probability value for the true negatives, and the opposite 25^{th} percentile value for the true positives. The results were spot on 0.244 and 0.751, respectively. We accordingly created three risk categories for mortality:

1) Low – Outcome probability greater or equal than 0 and lower than 0.25.
2) Medium – Outcome probability greater or equal than 0.25 and lower than 0.75.
3) High – Outcome probability greater or equal than 0.75.

Evaluation of this risk score throughout consecutive weeks is detailed in Table IV. The time points or horizons are defined as one, two, three, or more than three weeks since admission to hospital. Patients for each time window are unique, mapping total stay times between first visit and discharge.

TABLE IV
MORTALITY RISK ASSESSMENT FOR DIFFERENT TIME HORIZONS OF INPATIENT STAY. EACH CONTINGENCY TABLE SHOWS MEANS AND STANDARD DEVIATIONS COMPUTED FROM THE PROGNOSTIC OUTCOMES OVER THE FIVE MULTISTART TEST SETS.

[0,7) days	Death	Survival
Low Risk	0.40 ± 0.37%	64.38 ± 5.28%
Medium Risk	5.00 ± 1.29%	16.19 ± 3.49%
High Risk	11.80 ± 2.02%	2.22 ± 1.34%
[7,14) days	Death	Survival
Low Risk	0.27 ± 0.61%	51.13 ± 3.33%
Medium Risk	3.38 ± 0.92%	33.37 ± 5.10%
High Risk	6.49 ± 1.09%	5.33 ± 2.32%
[14, 21) days	Death	Survival
Low Risk	0.40 ± 0.91%	35.10 ± 5.10%
Medium Risk	5.78 ± 3.68%	38.42 ± 2.99%
High Risk	9.67 ± 3.32%	10.60 ± 4.77%
≥20 days	Death	Survival
Low Risk	4.34 ± 2.52%	23.05 ± 4.20%
Medium Risk	13.22 ± 9.47%	44.35 ± 8.60%
High Risk	9.60 ± 3.37%	5.41 ± 3.48%

Applying the risk assessment to the predictions returned for the hold-out group (see Table III) showed a high precision for the low risk label with 0 deaths overall on that category. For the medium risk label, the model was also very effective during the first three weeks of hospital stay, categorizing as medium only patients who survived the infection. Only on those patients with a hospital stay beyond three weeks, the risk predictor incorrectly tagged as medium risk 5 of them, who eventually passed.

C. Cross-study Model Performance

A number of short communications already shown the under performance of mortality prognostic models on external datasets by using data from other countries different from where the scores were designed [18]–[20]. Two of the usually chosen models, either because of their simplicity, or because the sample size was large, were Yan et al. [8] and the ISARIC 4C Mortality Score [3]. In the case of Yan et al., a decision tree with just three nodes predicted the mortality outcome more than 10 days in advance of the discharge with an accuracy slightly over 90%. The nodes account for three COVID-19 related biomarkers, namely lactic dehydrogenase (LDH), lymphocyte percentage, and high-sensitivity C-reactive protein (hs-CRP). In the case of the 4C mortality score, it was trained over more than 35,000 patients and in-house validated with an extra cohort of over 22,000 extra cases. The 4C score induces a mortality risk score based on 6 clinical variables and 2 blood tests. Authors in [3] report an overall accuracy of 70.75%, with sensitivity and specificity of 92.5 and 38.6, respectively.

TABLE V
PERFORMANCE OF ISARIC 4C MORTALITY SCORE [3] WHEN APPLIED TO OUR CLINICAL DATASET BASED ON THE REPORTED HIGH MORTALITY RISK THRESHOLD OF 9 POINTS.

Prediction	Death	Survival
Positive (4C $\geq$ 9)	272	942
Negative (4C $<$ 9)	4	580
Accuracy	0.474	
Precision	0.224	
Sensitivity	0.985	
Specificity	0.381	
F_1 Score	0.365	

Figure 3 and Table V present the results of both models when predicting mortality on our dataset. Even though both models share little methodologically, they are similarly conservative when predicting, with a sensible imbalance towards the positive class. Numerically, this effect translates into a specificity value of 0.193 and 0.381, respectively. The problem is then the poor performance for the positive class, where the positive predicted value (PPV) go only up to 0.174 and 0.224.

A direct explanation for this behavior is the large mortality rate in each work. For Yan et al., it was of 46.4%, whereas in Knight et al. reached a 32.2% within training sets. Such large positive class balance and the goal of not missing any critical case make the models unspecific when the population in which they were trained changes significantly. A clear example is also the high relevance in the mortality decision for Yan et al. of the LDH values (see Figure 3). Relatively high levels of LDH alone seem to play a crucial role in distinguishing the vast majority of cases that require immediate medical attention in our dataset.

In the case of Knight et al., the tendency to be overprotective and predict a positive outcome over a negative one seems more of a design decision given how authors chose the translation between the statistical coefficients of their lasso model and the linear weights derived to compute the final score. To assess the method's performance, only a single random train & test split was done. It is well documented that this practice can impose population shifts in the training of clinical models. Other recent publications also reported substantial disagreements with the original performance of the 4C score [20], [21].

Conversely, we tried to validate our consensus model over the datasets collected by the two works. Although Yan et al.

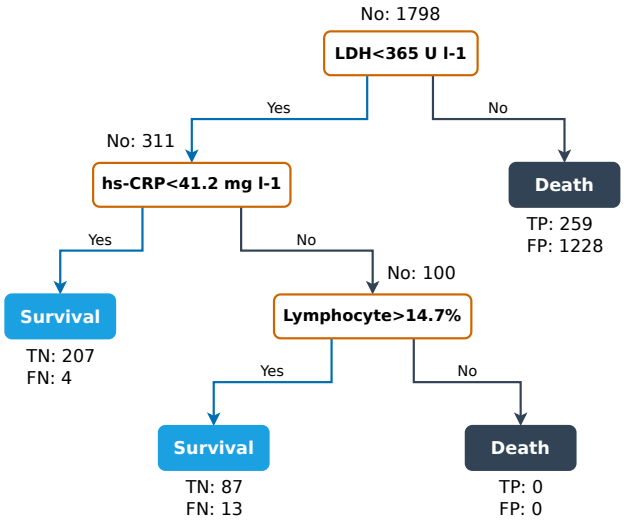


Fig. 3. Performance of Yan et al. prognostic model [8] when applied to our clinical dataset. Predictions reflect true (T) or false (F) cases with positive (deaths, P) or negative (survival, N) outcomes.

provide their database online, not all the variables in our model are included, making impossible to run any prediction. In the case of Knight et al., we were unable to access their data corpus despite our best efforts.

V. Discussion

The set of features in our consensus model (see Figure 1) includes several tests already reported to be highly relevant in COVID-19 pathology: Age, Lymphocytes, Heart rate, Lactate Dehydrogenase (LDH), Oxygen saturation, Platelets, C-reactive protein, Comorbidities, and Leukocytes. More importantly, other tests appear to have a large relevance in the multivariate prediction, pointing out to new findings. Among these, we should highlight the importance of high levels of eosinophils in patients who overcame the infection. It is not the first time that the importance of blood eosinophils is reported for COVID-19 [22], although it is a relatively underrated molecule for the research community. Another significant finding by our model is how large levels of Bilirubin and Urea nitrogen correlate with bad outcomes. We hypothesize that large levels of bilirubin and urea could be early indicators of hepatic and renal dysfunctions, two typical targets for the SARS-CoV-2 virus infection [23], [24].

Another popular biomarker of bad prognosis for the COVID-19 research community is the serum level of D-Dimer. Our feature selection, on the contrary, systematically rejected this molecule as relevant when the multivariate prognosis model is learned. D-Dimer is known to be a key molecule to monitor a possible course of the infection towards a severe coagulopathy, however, its values are not as predictive of mortality at the beginning of the hospital stay. With respect to epidemiological key features, the male biological sex is commonly designated as a risk factor for bad outcome in COVID-19. Although widely accepted, our data and model point out to a more nuance finding: the biological sex adds no significant risk in a patient's prognosis at time of hospitalization.

In terms of numerical performance, our risk indicator reaches AUC and sensitivity values around 0.9, with specificity estimators over 0.75 (see Table II for full details). It should be noted that several of the false negative patients for which the model still errs are cases of patients who either died from non-COVID causes or stay in hospital during extended periods of time. This fact suggests that the model identifies these patients as potentially with good prognosis based on their initial values at triage, although a fatal outcome is eventually registered. In parallel, and due to the cost-sensitive calibration process to minimize false negatives, we expected to have an increase in the number of false positives. Interestingly, most false positive predictions are associated to people with large clinical histories, i.e., several comorbidities and immunosuppression conditions. The rate of respirator usage for these patients is significantly higher than in the rest of patients. The conclusion is that these patients received more active medical interventions to save their lives when identified as high risk by healthcare providers. It is worth noting that the risk indicator was correct in assigning high risk to these patients. It was the medical intervention which fortunately changed the expected natural course of the infection. We envision the integration of all these factors in future evolutions of our mortality indicator.

The time horizon analysis in Section IV-B offers a clear interval of hospitalization for which our risk assessment is more accurate, from day 1 to day 14. Beyond two weeks of stay, the risk assessment at triage loses precision by pooling most patients into the medium risk category. An important recommendation can be therefore stated: For those patients still in hospital two weeks after admission, the mortality risk assessment should be recomputed using updated biochemical test values. Although the recommendation for this re-evaluation is particular to our prognosis model, we believe this should be the general case for any machine learning-enabled method used as a clinical aid in COVID-19 care.

All the aforementioned findings and performance estimations obtained from the multistart inference process were validated by predicting an unseen set of new patients, i.e., a hold-out group of patients later released by the data producer. With a sensitivity of 1 and an AUC of 0.880, our model proved its potential to capture all the patients with an actual severe prognosis while not penalizing the specificity on the negative class (0.759 value). In terms of risk assessment, the model also demonstrated high sensitivity when predicting the mortality risk of patients who stayed hospitalized at least during three weeks. It is also noteworthy that the hold-out group of patients included admissions throughout a year since the pandemic started, up to the end of February 2021. Predictions from our model were reliable for all these patients and proved not to be affected by data biases due to the various outbreaks origins of this cohort.

VI. Conclusions

In this paper we have introduced an explainable, highly sensitive, machine learning model for the prediction of the mortality risk associated to newly admitted COVID-19 patients. The model was inferred from commonly available biochemical tests included in any typical blood panel. A cost-sensitive analysis during the induction process allowed the model to be highly accurate to severe infections, while keep the percentage of false positives under a modest 25%. This is a key feature in order to be deployed in a clinical setting. The trade-off between a deadly prediction and the actual outcome provides reliability to the healthcare provider whose mission is to assign resources to an individual patient. When a prognostic model almost always produce a bad prognosis (e.g. seen in Figure 3 and Table V), such model becomes unusable for the clinical domain where resources are limited, especially during a pandemic.

Generalization was obtained by inferring the model over a multistart framework, which proved to be a successful approach. The model's estimated performance was similar to the performance figures recorded when applying to the hold-out group, even though training encompassed data from the initial surge of COVID-19 and the validation cohort ranged through the first full year of the pandemic.

It is still under scrutiny how the different variants affects the course of the infection and their severity degrees when infecting genetically different population. Therefore and despite the reported good performances at training and validation, our prognostic model should be validated across different populations and countries. These independent validations would have a double-fold aim: to test how it generalizes to a wide range of population; and also to test whether is able to predict disease outcomes produced by different virus variants already existing. To these ends, we make our model available to the community through a risk calculator on a web frontend, and provide the full model specification in a single exportable file for download.

Acknowledgment

R.A. would like to thank Dr. Iñaki Inza and Dr. Borja Calvo for their support and insightful comments during the design and execution of this work.

References

[1] World Health Organization. (2021, Aug.) COVID-19 weekly epidemiological update. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20210803_Weekly_Epi_Update_51.pdf

[2] J. Liu, L. Zhang, Y. Yan, Y. Zhou, P. Yin, J. Qi, L. Wang, J. Pan, J. You, J. Yang *et al.*, "Excess mortality in wuhan city and other parts of china during the three months of the covid-19 outbreak: findings from nationwide mortality registries," *BMJ*, vol. 372, 2021.

[3] S. R. Knight, A. Ho, R. Pius, I. Buchan, G. Carson, T. M. Drake, J. Dunning, C. J. Fairfield, C. Gamble, C. A. Green *et al.*, "Risk stratification of patients admitted to hospital with covid-19 using the isaric who clinical characterisation protocol: development and validation of the 4c mortality score," *BMJ*, vol. 370, p. m3339, 2020.

[4] A. Pourbagheri-Sigaroodi, D. Bashash, F. Fateh, and H. Abolghasemi, "Laboratory findings in covid-19 diagnosis and prognosis," *Clinica Chimica Acta; International Journal of Clinical Chemistry*, vol. 510, p. 475, 2020.

[5] C. An, H. Lim, D.-W. Kim, J. H. Chang, Y. J. Choi, and S. W. Kim, "Machine learning prediction for mortality of patients diagnosed with covid-19: a nationwide korean cohort study," *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.

[6] M. Pourhomayoun and M. Shakibi, "Predicting mortality risk in patients with covid-19 using machine learning to help medical decision-making," *Smart Health*, vol. 20, p. 100178, 2021.

[7] L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, M. M. Bonten, D. L. Dahly, J. A. Damen, T. P. Debray *et al.*, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *BMJ*, vol. 369, 2020.

[8] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang *et al.*, "An interpretable mortality prediction model for covid-19 patients," *Nature machine intelligence*, vol. 2, no. 5, pp. 283–288, 2020.

[9] R. K. Gupta, E. M. Harrison, A. Ho, A. B. Docherty, S. R. Knight, M. van Smeden, I. Abubakar, M. Lipman, M. Quartagno, R. Pius *et al.*, "Development and validation of the isaric 4c deterioration model for adults hospitalised with covid-19: a prospective cohort study," *The Lancet Respiratory Medicine*, 2021.

[10] G. Santafe, I. Inza, and J. A. Lozano, "Dealing with the evaluation of supervised classification algorithms," *Artificial Intelligence Review*, vol. 44, no. 4, pp. 467–508, 2015.

[11] HM Hospitales. (2020) Covid Data Save Lives. [Online]. Available: https://www.hmhospitales.com/coronavirus/covid-data-save-lives

[12] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[13] R. Martí, "Multi-start methods," in *Handbook of metaheuristics*. Springer, 2003, pp. 355–368.

[14] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2005.

[15] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002.

[16] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, vol. 17, no. 1. Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

[17] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv:1705.07874*, 2017.

[18] M. J. Quanjel, T. C. van Holten, P. C. Gunst-van der Vliet, J. Wielaard, B. Karakaya, M. Söhne, H. S. Moeniralam, and J. C. Grutters, "Replication of a mortality prediction model in dutch patients with covid-19," *Nature Machine Intelligence*, vol. 3, no. 1, pp. 23–24, 2021.

[19] C. Dupuis, E. De Montmollin, M. Neuville, B. Mourvillier, S. Ruckly, and J. Timsit, "Limited applicability of a covid-19 specific mortality prediction rule to the intensive care setting," *Nature Machine Intelligence*, vol. 3, no. 1, pp. 20–22, 2021.

[20] M. Covino, G. De Matteis, M. L. Burzo, A. Russo, E. Forte, A. Carnicelli, A. Piccioni, B. Simeoni, A. Gasbarrini, F. Franceschi *et al.*, "Predicting in-hospital mortality in covid-19 older patients with specifically developed scores," *Journal of the American Geriatrics Society*, vol. 69, no. 1, pp. 37–43, 2021.

[21] Z. Wellbelove, C. Walsh, T. Perinpanathan, P. Lillie, and G. Barlow, "Comparing the 4c mortality score for covid-19 to established scores (curb65, crb65, qsofa, news) for respiratory infection patients," *The Journal of Infection*, 2020.

[22] G. Xie, F. Ding, L. Han, D. Yin, H. Lu, and M. Zhang, "The role of peripheral blood eosinophil counts in covid-19 patients," *Allergy*, 2020.

[23] A. Kumar, A. Arora, P. Sharma, S. A. Anikhindi, N. Bansal, V. Singla, S. Khare, and A. Srivastava, "Gastrointestinal and hepatic manifestations of corona virus disease-19 and their relationship to severe clinical course: A systematic review and meta-analysis," *Indian Journal of Gastroenterology*, vol. 39, no. 3, pp. 268–284, 2020.

[24] Y. Cheng, R. Luo, K. Wang, M. Zhang, Z. Wang, L. Dong, J. Li, Y. Yao, S. Ge, and G. Xu, "Kidney disease is associated with in-hospital death of patients with covid-19," *Kidney international*, vol. 97, no. 5, pp. 829–838, 2020.