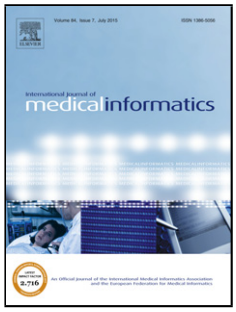
Journal Pre-proof



Personalized Predictive Models for Symptomatic COVID-19 Patients Using Basic Preconditions: Hospitalizations, Mortality, and the Need for an ICU or Ventilator<!–<ForCover>Wollenstein-Betech S, Cassandras CG, Paschalidis IC, Personalized Predictive Models for Symptomatic COVID-19 Patients Using Basic Preconditions: Hospitalizations, Mortality, and the Need for an ICU or Ventilator, *International Journal of Medical Informatics*, doi: 10.1016/j.ijmedinf.2020.104258</ForCover>–>

Salomón Wollenstein-Betech, Christos G. Cassandras, Ioannis Ch. Paschalidis

PII: S1386-5056(20)30616-X

DOI: https://doi.org/10.1016/j.ijmedinf.2020.104258

Reference: IJB 104258

To appear in: *International Journal of Medical Informatics*

Received Date: 3 May 2020

Revised Date: 26 July 2020

Accepted Date: 17 August 2020

Personalized Predictive Models for Symptomatic COVID-19 Patients Using Basic Preconditions: Hospitalizations, Mortality, and the Need for an ICU or Ventilator

Salomón Wollenstein-Betech[1], Christos G. Cassandras[1], and Ioannis Ch. Paschalidis[1,2,*]

[1]Department of Electrical & Computer Engineering, and Division of Systems Engineering, Boston University, 8 Saint Mary's St., Boston, MA 02215,

[2]Department of Biomedical Engineering, Boston University, 44 Cummington Mall, Boston, MA 02215,

*Corresponding author

Mailing Address: Department of Electrical and Computer Engineering, Boston University, 8 Saint Mary's Street, Boston, MA 02215, USA
Tel: (617) 353-0434 Fax: (617) 353-6440
E-mail: yannisp@bu.edu

ABSTRACT

Background: The rapid global spread of the SARS-CoV-2 virus has provoked a spike in demand for hospital care. Hospital systems across the world have been over-extended, including in Northern Italy, Ecuador, and New York City, and many other systems face similar challenges. As a result, decisions on how to best allocate very limited medical resources and design targeted policies for vulnerable subgroups have come to the forefront. Specifically, under consideration are decisions on who to test, who to admit into hospitals, who to treat in an Intensive Care Unit (ICU), and who to support with a ventilator. Given today's ability to gather, share, analyze and process data, personalized predictive models based on demographics and information regarding prior conditions can be used to (1) help decision-makers allocate limited

resources, when needed, (2) advise individuals how to better protect themselves given their risk profile, (3) differentiate social distancing guidelines based on risk, and (4) prioritize vaccinations once a vaccine becomes available.

Objective: To develop personalized models that predict the following events: (1) hospitalization, (2) mortality, (3) need for ICU, and (4) need for a ventilator. To predict hospitalization, it is assumed that one has access to a patient's basic preconditions, which can be easily gathered without the need to be at a hospital and hence serve citizens and policy makers to assess individual risk during a pandemic. For the remaining models, different versions developed include different sets of a patient's features, with some including information on how the disease is progressing (e.g., diagnosis of pneumonia).

Materials and Methods: National data from a publicly available repository, updated daily, containing information from approximately 91,000 patients in Mexico were used. The data for each patient include demographics, prior medical conditions, SARS-CoV-2 test results, hospitalization, mortality and whether a patient has developed pneumonia or not. Several classification methods were applied and compared, including robust versions of logistic regression, and support vector machines, as well as random forests and gradient boosted decision trees.

Results: Interpretable methods (logistic regression and support vector machines) perform just as well as more complex models in terms of accuracy and detection rates, with the additional benefit of elucidating variables on which the predictions are based. Classification accuracies reached 72%, 79%, 89%, and 90% for predicting hospitalization, mortality, need for ICU and need for a ventilator, respectively. The analysis reveals the most important preconditions for making the predictions. For the four models derived, these are: (1) for hospitalization: *age, pregnancy, diabetes, gender, chronic renal insufficiency, and immunosuppression*; (2) for mortality: *age, immunosuppression, chronic renal insufficiency, obesity and diabetes*; (3) for ICU need: *development of pneumonia (if available), age, obesity, diabetes and hypertension*; and (4) for ventilator need: ICU and pneumonia (if available), *age, obesity, and hypertension.*

Key words: Predictive models, COVID-19, coronavirus, SARS-CoV-2, hospitalization, mortality, ICU, ventilator, Electronic Health Records (EHRs).

1 INTRODUCTION

Currently, the world is facing a health and economic crisis due to the spread of the virus SARS-CoV-2 which causes a disease referred to as COVID-19 [1]. By the end of April 2020, the virus has spread to over 3.3 million people worldwide and has killed over 230,000 [2,3]. During this pandemic, governments and hospitals have struggled to allocate scarce resources, including tests, treatment in intensive care units (ICUs) and ventilators [4,5].

As the virus continues to spread, *predicting* hospitalizations, mortality, and other patient outcomes becomes important for several reasons: (i) using risk profiles to inform decisions on

who should be tested (for the virus and/or antibodies) and at which frequency, (ii) providing more accurate estimates of who is more likely to be hospitalized and the type of care they may need, (iii) informing plans for staffing, resources, and prioritizing the level of care in extremely resource-constrained settings. Equally importantly, as societies adapt to the pandemic, predictive models can (i) assess individual risk so that social distancing measures can transition from “blanket" to more targeted (e.g., deciding who can return to work, who is advised to stay at home, who should be tested, etc.) and (ii) direct policy decisions on who should receive priority for vaccination, which will be critical as initial vaccine production may not suffice to vaccinate everybody.

In an attempt to understand better the disease, several predictive models have been developed during this pandemic [6]. One of the limitations of all of these predictors is their high-risk of bias given their small sample sizes. In fact, out of the 66 predictors summarized in [6], the mean and standard deviation of the sample size and test size used are 443.5 ± 560 and 155 ± 276 respectively. In turn, our work provides with a less biased predictor by employing a dataset which is 4.7 times larger than the biggest dataset reported in [6].

To develop predictive models, we leverage supervised machine learning methods that learn from given examples of predictive variables and associated outcomes – the so called *training* set. Performance is then evaluated on a separate *test* set. In the specific application of interest, we will focus on classification, a setting where the outcome is binary, e.g., someone is hospitalized or not.

Many models have been used to predict a patient admission to a hospital, mortality and other health care applications based on comorbidities. Some examples include: predicting morbidity of patients with chronic obstructive pulmonary disease [7], febrile neutropenia [8], as well as classifying the hospitalization of patients with preconditions on diabetes [9], heart disease [10,11], and hospital readmission for patients with mental or substance use disorders [12]. Recent advances in the machine learning literature have suggested that sparse classifiers, those that use few variables (e.g., $l1$-regularized Support Vector Machines), have stronger predictive power and generalize better on out-of-sample data points than very complex classifiers [13]. Related work has shown that regularization is equivalent to robustness, that is, learning models which are robust to the presence of outliers in the training set [14]. Moreover, the benefit of using sparse predictors is the enhanced *interpretability* they provide for both the model and the results.

1.1 Objective

Construct data-driven predictive models using data from patients tested for SARS-CoV-2 to predict if a patient will (1) be hospitalized, (2) succumb to the disease, (3) need treatment in an ICU, and/or (4) need a ventilator. To train and test these classifiers we use a public dataset [15] made available by the Mexican government that contains individual information on: demographics (e.g., location), preconditions (e.g., hypertension) and outcomes (e.g., admission to an ICU) for every person who has been tested for SARS-CoV-2 in Mexico.

1.2 Main Contributions

- We provide descriptive statistics of the distribution of hospitalized and deceased patients given basic information on preconditions and demographics.

- We develop interpretable models that not only predict the outcomes but also quantify the role of various variables in making these predictions.
- The models we develop leverage data from Mexico. This can motivate additional work using the same data, while the models could be applicable to other Latin American countries with similar population characteristics. This adds to existing work using Electronic Health Records which has focused on patients in the US, Europe, or Asia.

The remainder of the paper is organized as follows: In Section 2 we describe the data used accompanied by descriptive statistics and preprocessing procedures. In Section 3 we describe the binary supervised classification models used and the performance evaluation metrics employed. In Section 3, we present the main results. Discussion of the results can be found in Section 4 and Conclusions in Section 5.

2 DATA DESCRIPTION AND PREPROCESSING

2.1 Data

We use a dataset that has been open for the general public by the Mexican Government (and updated daily) [15]. These data include information about every person who has been tested for SARS-CoV-2 in Mexico. They include demographic information such as: *Age, Location, Nationality, the use of an indigenous language*; as well as information on prior medical conditions, including whether the patient has: *diabetes, chronic obstructive pulmonary disease (COPD), asthma, immunosuppression (e.g., due to treatment for cancer or auto-immune conditions* [16]*), hypertension, obesity, pregnancy, chronic renal failure, other prior diseases,* and whether was or is using *tobacco*. In addition, the data report the dates on which the patient first noticed symptoms, the date when the patient arrived to a care unit, and the date when the patient was deceased (if applicable). Finally, it contains fields showing whether the patient was hospitalized, has pneumonia, needed a ventilator, was treated in an ICU, as well as the result of the SARS-CoV-2 test. To confirm a case, the Ministry of Health in Mexico requires that, in addition of being tested positive, the patient presents at least two of: cough, fever or headache, and at least one of: dyspnea, arthralgia, myalgia, odynophagia, rhinorrhea, conjunctivitis or chest pain. More technical details on the surveillance model used are provided in [17].

As of May 1st, 2020, the data contained more than 91,179 observations out of which more than 20,737 account for positive tests, around 15,000 tests were being processed, and the rest are negative test results. Table 2-1 provides a more precise description of the dataset.

Table 2-1: Descriptive statistics of data set as on May 1st, 2020. In parenthesis, we denote the number of observations belonging to the randomly selected test set.

Total number of tests	**91,179**
Positive	20,737 (6,239)
Waiting for Result	15,445 (4,677)
Negative	54,997
Total number of patients hospitalized	**24,099 (3,801)**
Positive	8,221 (1996)
Waiting for Result	4,389 (1,737)

Negative	11,489 (0)
Pneumonia	14,462 (1,737)
Need Ventilator	1,809 (246)
Need ICU	2,059 (258)
Deceased (Positive or Waiting for Result)	3,192 (501)
Number of observations with pre-conditions with non-negative test	
Diabetes	6,042 (1,878)
COPD	825 (231)
Asthma	1,235 (385)
Immunosuppression	632 (190)
Hypertension	7,238 (2,161)
Pregnant	221(64)
Cardiovascular disease	991 (267)
Obesity	6,998 (2,056)
Chronic renal insufficiency	820 (235)
Demographics of patients with non-negative test	
Contact with a positive COVID case	11,355 (3,360)
Speak an indigenous language	466 (128)

2.2 Basic Analytics

We provide plots that help us observe trends in the data. We begin by disaggregating data into age groups. In the lower plot of Figure 1 the number of observations of patients having a positive test or waiting their result per age is shown. In addition, the upper bar plot denotes the percentage of the patients in a certain age range who have been hospitalized. This information is aligned with the current knowledge on COVID-19, which indicates that older people have higher risk of being hospitalized. Also, this plot suggests that the risk of being hospitalized increases linearly from the age of thirty up to seventy-five and then plateaus. We ran an ordinary linear regression (OLS) to calculate the rate at which the percentage of hospitalization increases for every additional year of age. The rate results to be 0.014 with an R^2 equal to 0.99. This suggests that the risk of hospitalization increases by approximately 1.4% for every year of age between 30 and 75 years old.

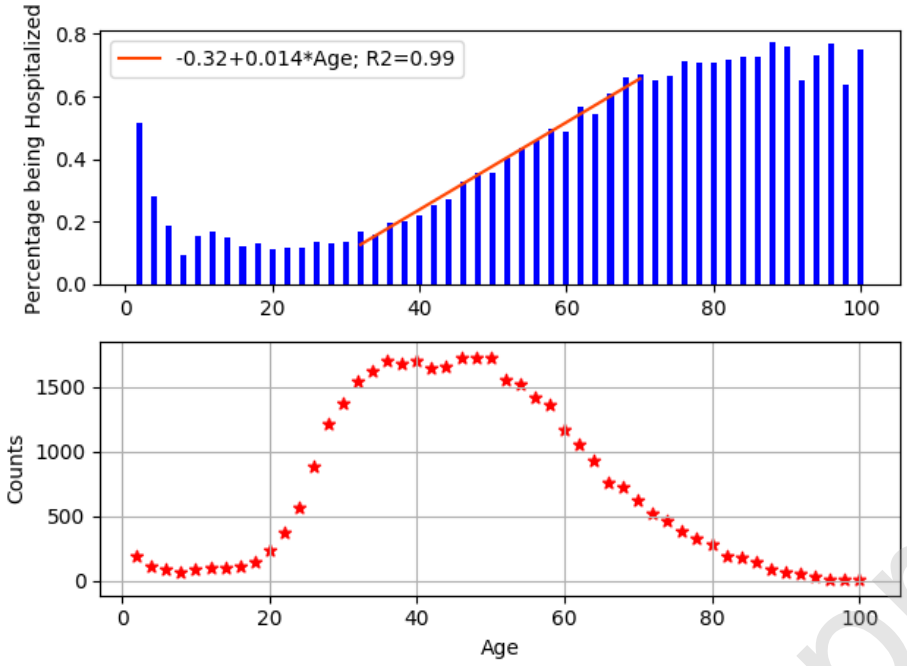


Figure 1: Lower: Number of patients tested positive or waiting for result by age; Upper: Percentage of these patients that have been hospitalized.

Next, in Figure 2 we report the fraction of patients who have been hospitalized, deceased, needed an ICU or a ventilator given a certain precondition, e.g., in the upper-left box we divide the number of hospitalized patients with pneumonia by the total number of patients with pneumonia. We observe that for both hospitalizations and deaths, preconditions such as chronic renal insufficiency, COPD, diabetes, immunosuppression, cardiovascular disease and hypertension are critical. Nevertheless, even though this gives us information about the risk of a precondition, it does not include the sensitivity regarding how age and preconditions affect a patient with COVID-19.

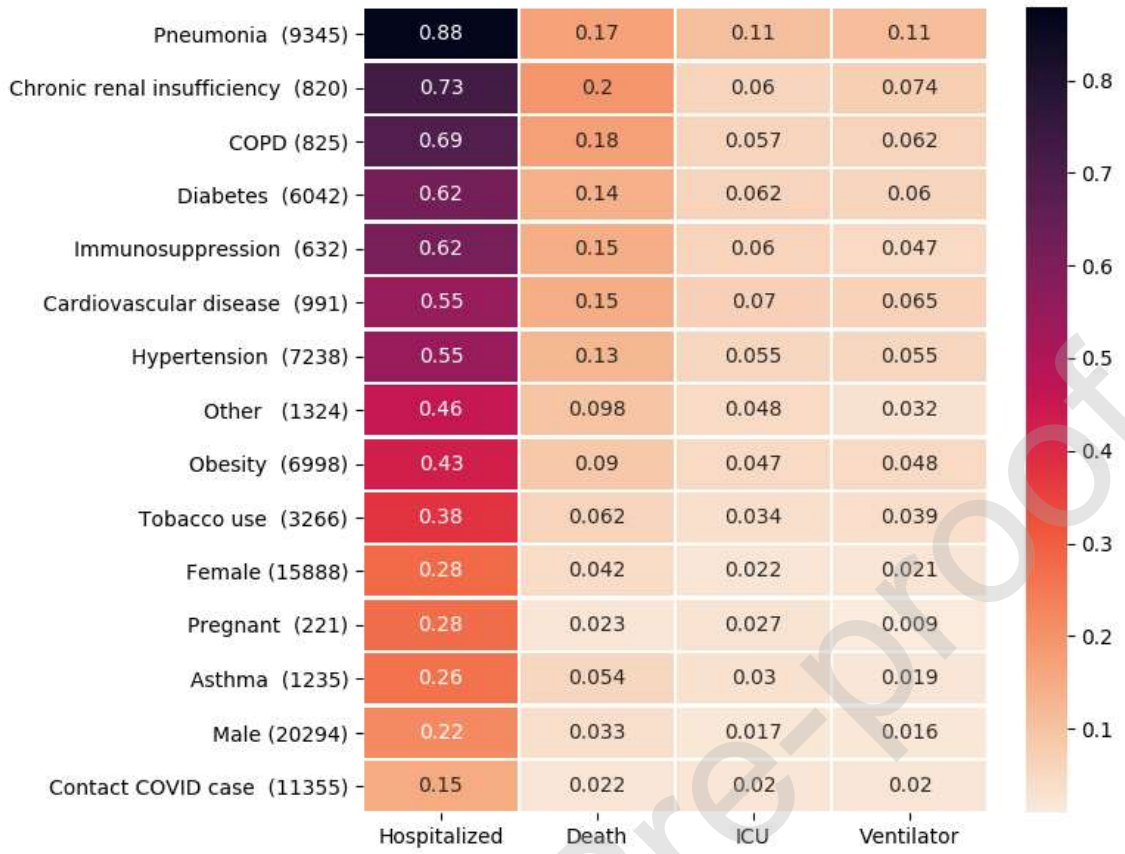


Figure 2: Fraction (%) of patients with a precondition that have been hospitalized, have died or required an ICU or ventilator.

To complement the previous table, we report the percentage of the hospitalized by age group and by existing preconditions in Figure 3. To that end, we create age groups for every five years and report results for groups with at least ten observations, otherwise the bin is left blank. On the top row of the table, we include the statistic for a patient without any preconditions. As an example, the top-left entry reports the ratio of the number of patients between 0-5 years old without preconditions who have been hospitalized divided by the number of patients between 0-5 years without preconditions who may or may not have been hospitalized. We observe that chronic renal insufficiency, diabetes, and immunosuppression are among the preconditions that are associated with a higher hospitalization rate.

Figure 3: Fraction (%) of population per age group being hospitalized given a precondition.

Finally, we present histograms reporting the lag times among various states of the disease for the Mexican population. For this analysis, we separate the data in three groups: individuals with ages between 0-20, 20-50, and patients over 50 years old. In Figure 4 (left), we plot the distribution of the number of days between the onset of symptoms and a subsequent hospitalization. Figure 4 (center) depicts the distribution of time (days) between hospital admission and death. Interestingly, we observe that a large portion of the patients who were hospitalized died the same day they were admitted. This could be explained either by a healthcare system working at capacity in which only seriously-ill patients are admitted or by the abrupt deterioration of a patient's condition [18,19] and should be further investigated. The rest of the distribution behaves like the tail of a Weibull distribution with very few patients being hospitalized for more than three weeks. Finally, Figure 4 (right) shows the distribution of the number of days between the onset of symptoms and death (the mean is 9.8 days).

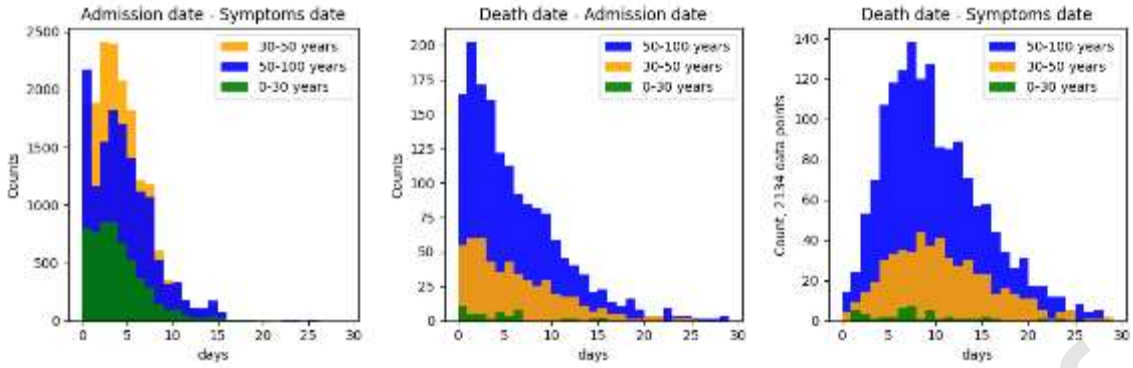


Figure 4: Histograms showing (left) the time between the onset of symptoms and death, (center) the time between hospital admission and death, and (right) the time between the onset of symptoms and death.

2.3 Preprocessing

2.3.1 Removing outliers

We found a few outliers which are easily identified, for example, the pregnancy of male patients, the date of death of a patient being earlier than the day the patient was admitted to the hospital. Such data points were removed from the dataset.

2.3.2 One-hot encoding

The data contain precondition features reported as categorical. Specifically, each of these precondition features takes the value *yes, no, unknown* or *unspecified*. We generate one-hot encoding for all these features. One-hot encoding converts the categorical feature to multiple binary variables by creating auxiliary variables that help distinguish between the different categories of a feature. For the case of our data, one-hot encoding generates three binary variables for each specific precondition; these variables (as opposed to categories) are: *no, unknown* and *unspecified.* Then, for each observation, at most one of these variables will be active, pointing to the correct value for the original feature. If none of the three is active, then the value of the precondition is *yes.*

2.3.3 Removing correlated variables

We find and delete variables that are highly correlated since they, in general, provide similar information. Specifically, we compute pairwise correlations among the variables, and remove one variable from each highly correlated pair (using a threshold of 0.8 for the absolute correlation coefficient). We found that the correlated binary features were the ones corresponding to *unknown* or *unspecified* for preconditions. This is because observations that contain an *unknown* or *unspecified* value, typically have this same value for all preconditions (not just for one), indicating potential issues in data gathering. Hence, we remove all these auxiliary variables denoting unknown or unspecified preconditions.

3 METHODS AND METRICS

In this section, we briefly introduce the methodologies used to build the binary classifiers. For each model, we train the classifier using four different supervised classification

methodologies: sparse Support Vector Machines (SVM), sparse Logistic Regression (LR), Random Forests (RF) and gradient boosted decision trees (XGBoost). For healthcare applications, the first two are preferable due to their interpretability. In turn, the last two are the state-of-the-art classification algorithms today and will serve as a basis to compare the accuracy of the interpretable methods with the non-interpretable benchmark models. Appendix B provides details on these methods, particularly because the robust/sparse LR and SVM formulations are not standard.

3.1 Cross-Validated Recursive Feature Elimination

Classifiers based on few variables are desirable because they have stronger predictive power, generalizing better out-of-sample, and offering enhanced interpretability [20,21]. Aiming to reduce the number of variables, we employ a Recursive Feature Elimination (RFE) procedure [22] to find the variables that optimize a given performance metric. The general framework of this algorithm begins by building a classifier using all the features and computing an importance score for each predictor. In the case of Logistic Regression (or Linear SVM), we use as important score the absolute value (or magnitude) of the linear coefficient β_i of feature i. After this step, the least important feature (the one with the smallest $|\beta_i|$) is deleted from the dataset. We repeat iteratively this process until we are left with one feature. Then, for each of these iterations we report the performance of the model (using cross-validation over the training set) and we pick the set of features that maximize this value. Additionally, at each iteration, we use the same cross-validation process to tune the hyper parameters of the classifier to achieve the best performance. In this work, we use LR to eliminate variables based on their coefficients as described earlier, as it gives a clear and interpretable meaning of the score for each variable. At each iteration we use a stratified ten-fold cross-validation (over the training set) to estimate the AUC performance until we are left with one variable. Finally, we pick the features for which we obtain the model with the maximum AUC value. This subset of variables is then used to train all the predictive models.

3.2 Performance Evaluation

The primary objective of learning a classifier is to maximize the prediction accuracy (or equivalently minimize a loss function), and in our health care setting offer interpretability of the results.

We characterize the prediction accuracy of a classifier using two commonly used metrics: (1) the *false positive (or false alarm) rate* which measures how many patients were predicted to be in the positive class, e.g., hospitalized, while they truly were not, as a fraction of all negative class patients. In the medical literature, the term *specificity* is often used and it equals 1 minus the false positive rate; and (2) the *detection rate* that captures how many patients were predicted to be on the positive class while they truly were, as a fraction of all positive class patients. This term is often referred to as *sensitivity* or *recall.* Another term commonly used is *precision* defined as the ratio of true positives over true and false positives.

A single metric that captures both types of error is the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC). ROC plots the detection rate (or sensitivity or recall) over the false positive rate. A naïve random selection (assigning patients to classes randomly)

has AUC of 0.5 while a perfect classifier an AUC of 1. To complement the AUC metric, we report the *accuracy* that computes the ratio of the number of correct predictions over all predictions within the test set. In addition to the ROC AUC and *accuracy,* and rather than evaluating the *precision* and *recall* for both the positive class and the negative class, we report a single metric, the weighted F1-score. Specifically, the F1-score is the harmonic mean of precision and recall and can be computed for both the positive and negative classes. The weighted F1-score is just an average of the per class F1-scores weighted according to the number of (test) samples in each class. We are interested in this performance metric because it is as important to accurately predict who is *not* likely to have a specific outcome (e.g., hospitalized), in addition to who will. For example, one can ease restrictions on those who are predicted to have lower risk. In fact, having more false positives corresponds to being more conservative with patients by assigning higher-risk profiles, and what is needed is striking the right balance between being conservative vs. having a lot of false positives. The weighted F1-score is one appealing way of quantifying this trade-off.

For individual variables and each different model, we also report the Odds Ratio (OR), which indicates how the odds of observing the outcome are scaled by having that variable take the value 1 (vs. 0), while controlling for all other variables in the model.

We finally emphasize that all metrics we report are computed on a randomly selected test set of patients (i.e., out-of-sample) which corresponds to 30% of the observations and has not been used for training the models. In addition, all metrics were calculated using a discriminant classification threshold which was selected by optimizing the AUC and reported in Table 2 : Summary of results of all models using LR.

4 RESULTS

We build binary classification models to predict hospitalization, mortality and the need for an ICU or ventilator. At a minimum, all models use a set of *base* features composed by: age, gender, diabetes, COPD, asthma, immunosuppression, hypertension, obesity, pregnancy, chronic renal failure, tobacco use, other disease, as well as the SARS-CoV-2 test result, which is either positive or pending (we exclude all negative cases to train our models). In this section, we provide a summary of the results while in Appendix A we provide all results.

4.1 Hospitalizations

Our first model predicts if a patient who has tested positive or is waiting for the test result will be hospitalized given their *base* features. This model has a moderate accuracy for all methodologies employed which accounts for an AUC of 0.74 and an accuracy of classifying 72% of the observations correctly. An interesting observation is that SVM and LR performs better than RF and XGBoost.

The coefficients of the SVM and LR models have the same trend and suggest that the features that contribute the most for predicting the hospitalization of a patient are: age over 80 (OR=3.2), age between 65-80 (OR=2), pregnancy (OR=2.3), diabetes (OR=2.3), chronic renal insufficiency (OR=2.3), immunosuppression (OR=2), COPD (OR=1.5), and gender. The rest of the variables (Obesity, Hypertension, Other, Tobacco Use, Cardiovascular disease and Asthma)

have a much smaller impact. It is however possible that some of these variables have smaller coefficients because the effect is captured by another highly correlated variable (e.g., obesity and diabetes).

4.2 Mortality

We explore two models to predict mortality. The first model assumes we only know the *base* features of a patient whereas the second model includes variables that indicate if the patient has been hospitalized or not, has pneumonia, or has needed an ICU or ventilator. The reason to consider the first model is to have a classifier which identifies which patients are the most vulnerable prior to hospitalization, while the second model predicts the mortality of an individual in the hospital by using information on how the disease is progressing. In order to have a more balanced dataset and to detect better the deceased class, we ran this model only on the observations of patients who have been hospitalized and have been tested positive or are waiting for their test result.

For the model which considers the case that only uses the *base* features of a patient (prior to attending a healthcare facility)**,** we are able to predict with 79% accuracy and with an AUC equal to 0.63 the mortality of a patient. Moreover, when we include more information about the hospitalization, pneumonia, ICU, and ventilation, the classification task achieves a similar accuracy but a higher detection rate of 0.701 (an increase of ~12% in detection).

Both interpretable models, LR and SVM, suggest that the variables that are critical for predicting mortality are the patient's age, gender, immunosuppression (OR=1.68), chronic renal insufficiency (OR=1.46), obesity (OR=1.4) and diabetes (OR=1.32). For the model that has more features, as expected, information about the need for ventilator and ICU are highly relevant when predicting mortality.

4.3 ICU need

Similar to the mortality case, we train two classification models to predict the need for an ICU. The first model predicts the need for an ICU bed using the *base* features and assumes that we don't know if the patient will or will not develop pneumonia. This might serve for planning purposes, as it will help us predict which individuals are more likely to need an ICU in case they contract SARS-CoV-2. This model achieves an accuracy of 89% with an AUC of 0.55 (XGBoost). Additionally, when we include information about the development of pneumonia, the AUC of the model increases by about 10% to 0.64, highlighting the importance of using the most recent information of a patient while predicting its outcome.

In these cases, SVM and LR suggest that information on: age, development of pneumonia (OR=4.13), if available, diabetes (OR=1.23), obesity and hypertension are among the most important variables to predict the need for an ICU.

4.4 Ventilator Need

In the same way as in the mortality and ICU models, we develop two models to predict the need for a mechanical ventilator given that a patient is either a confirmed or suspected COVID-19 case. The first model evaluates the situation prior to knowing if patient has developed pneumonia or needs an ICU**.** The accuracy reached by this model is higher than both the

mortality and the ICU models, achieving an accuracy of 90% and an AUC of 0.58. In addition to this model, the second instance uses information about the development pneumonia and the admission to an ICU**.** As expected, this additional information is relevant for predicting ventilation need. It increases its accuracy to 92% and the AUC to 0.86.

Moreover, both models classifying the need for a ventilator show that information on ICU (OR=15.5) and pneumonia (OR=9.1), if available, age, gender, chronic renal insufficiency (OR=1.5), obesity (OR=1.4), hypertension (OR=1.16) and diabetes (OR=1.12) are the most relevant features for predicting the need for a mechanical ventilator.

To summarize and provide interpretability we report in “Table 2 : Summary of results of all models using LR.” the performance metrics for all the models and in “Table 3 : Odds ratios for all models, considering LR-l1.” the odds ratio for each model variables using LR. We observe that the coefficients of both interpretable models (SVM and LR) are consistent and have an accuracy comparable, or higher than RF and XGBoost.

Table 2 : Summary of results of all models using LR.

	Hospitalization	Mortality	Mortality (advanced)	ICU	ICU (advanced)	Ventilator	Ventilator (advanced)
Discriminant Threshold	0.424	0.36	0.32	0.22	0.22	0.23	0.35
Accuracy	0.718	0.793	0.794	0.894	0.894	0.899	0.917
F1w	0.7	0.716	0.75	0.844	0.844	0.851	0.911
AUC	0.749	0.634	0.701	0.534	0.636	0.578	0.859

Table 3 : Odds ratios for all models, considering LR-l1.

	Hospitalization	Mortality	Mortality (advanced)	ICU	ICU (advanced)	Ventilator	Ventilator (advanced)
Age-80-100	3.180	2.361	3.212	1.000	1.000	1.000	1.002
Pregnant	2.321	1.000	1.245	1.000	1.000	1.000	1.000
Diabetes	2.291	1.324	1.309	1.230	1.197	1.120	1.082
Chronic Renal Insufficiency	2.268	1.458	1.468	0.631	0.627	1.000	1.513
Immunosuppression	2.088	1.684	1.699	0.922	0.958	0.589	1.000
Age-65-80	2.073	1.461	1.744	1.204	1.298	1.294	1.133
COPD	1.536	1.266	1.000	0.963	0.913	0.911	0.641
Other	1.411	1.363	1.317	1.000	1.025	0.729	0.562
Obesity	1.323	1.399	1.232	1.330	1.247	1.441	1.313
Hypertension	1.157	1.315	1.179	1.169	1.151	1.162	1.092
Age-50-65	1.000	1.000	1.000	1.019	1.102	1.116	1.000
Tobacco Use	0.965	0.852	0.871	0.720	0.701	0.872	1.115
Cardiovascular Disease	0.962	1.048	1.200	1.003	1.010	1.000	1.000
Asthma	0.773	1.420	1.737	1.037	1.040	0.748	0.625
Gender	0.549	0.687	0.705	0.780	0.806	0.732	0.806
Age-30-50	0.457	0.618	0.665	0.903	0.979	0.701	0.597
Age-0-30	0.259	0.271	0.269	0.638	0.731	0.733	0.789
Ventilator			4.341				
ICU			1.297				15.534
Pneumonia			1.276		4.125		9.098

5 DISCUSSION

Overall, the models we develop range from moderately to significantly accurate. Predicting hospitalizations appears harder just based on the basic variables at our disposal, particularly considering all patients who have a positive test or with a test pending. Potential additional features are at play including state of health (measured through detailed lab results) and the viral load they were exposed to. Furthermore, a number of hospitalizations are driven by socioeconomic factors, e.g., the living arrangements of a patient and whether he/she can pose infection risk for many others. Still, an AUC of 0.75 is significantly better than random and the results could help tighten estimates on the number of hospitalizations expected.

From an actionable and planning perspective, predicting ICU treatment and ventilator need are quite useful. These models can be quite accurate, achieving accuracies of 89% and 90%, respectively, when information on how the disease is progressing is taken into account (e.g., development of pneumonia). Similarly, the mortality model can achieve an accuracy of 76%.

An interesting observation is that interpretable models (such as LR and SVM), when used in conjunction with robustness/regularization approaches and elaborate feature selection procedures, can lead to performance that is comparable, if not better than more complex and expensive classifiers. The significant advantage of the former models is that they are interpretable and provide information on which variables drive the predictions.

This study has some limitations. It is important to emphasize that the dataset used in this work lacks critical information (such as lab results, vital signs, among others) to be able to provide a clinical understanding of COVID-19. Rather, the focus of this work is to help inform decisions on how to best allocate limited medical resources, and to help design targeted policies for vulnerable subgroups which might not have access to clinical and lab assessments. Interesting patterns can be observed in our results, motivating further research directions in resource allocation during a pandemic. For example, our results suggest that pregnancy is an important variable for predicting hospitalization but not mortality, ICU or ventilation, potentially indicating a bias towards being more conservative and hospitalizing pregnant women when they may not need it. Readers should also be aware that, due to the insufficient testing resources in Mexico, the dataset might be biased toward overestimating deaths. While the dataset may reflect all deaths, it does not include mild-moderate COVID-19 cases as these are never tested. Another limiting factor is that the dataset does not include specific dates at which hospitals discharge patients, which is of high importance to assess the utilization of medical equipment. Finally, to the extent that these risk models can be used to prioritize the use of resources, we understand that medical risk is not the only factor in making such decisions. Nevertheless, in order to quantify medical risk one can leverage the models presented in this work.

6 CONCLUSIONS

We develop models to identify the medical risk of a patient with (or suspected for) COVID-19. We hope this work can help hospitals and policymakers to distribute more effectively their limited resources including tests, ICU beds and ventilators, as well as, to motivate countries and healthcare systems to standardize and share data with the medical informatics community. Moreover, we hope this research spreads the knowledge of the existence of this public dataset

and motivates researchers to work with these data. Finally, we hope that risk models are taken into account to fine-tune social distancing advisories, moving from “blanket” to risk-based, as well as prioritizing vaccine distribution to the more vulnerable and to those who need to interact with the more vulnerable. For the sake of reproducibility and to facilitate the analysis for further research we have made our models and results available on a Github repository [23].

7 ACKNOWLEDGEMENTS

Research partially supported by the NSF under grants IIS-1914792, DMS-1664644, and CNS-1645681, by the ONR under MURI grant N00014-19-1-2571, and by the NIH under grant 1R01GM135930.

The authors would like to thank Diana Sverdlin-Lisker at the Massachusetts Institute of Technology for useful discussions.

8 SUMMARY TABLE

What was already known	What this study added to the knowledge
Due to the fast spread of COVID-19, a lot of attention has been devoted to measuring and predicting the spread. There have also been anecdotal reports on certain prior conditions that appear to lead to more severe disease.	This work is among the first to use data to develop explicit models predicting hospitalization, ICU treatment, ventilator use, and mortality for individual patients.
Most research related to COVID-19 has been done in countries and communities where the virus hit first. These include China, Italy, Spain, US.	Our research focuses on the Mexican population, which has particular characteristics of interest to Latin American countries with similar socio-economic conditions and health care systems that may become more congested due to COVID-19.
Most research related to COVID-19 that employs Machine Learning techniques has been focused on learning from complex data sources such as chest scans [24–28].	We focus on a basic set of preconditions that are known for the vast majority of the population without the need to attend a medical facility. Hence, the risk metrics we develop can be computed for anyone susceptible to COVID-19, helping to prioritize testing, care, and post-surge social distancing and vaccination policies.

Authors’ Contributions

S.W.-B. co-designed and analyzed the methods, co-wrote the manuscript, performed the analysis, and produced results and figures. C.G.C co-led the study, co-designed the methods, and commented on the manuscript. I.C.P. co-led the study, co-designed the methods, and co-wrote the manuscript.

CONFLICT OF INTEREST STATEMENT:

The authors have no financial or personal relationships with other people or organizations that could inappropriately influence (bias) their work.

REFERENCES

[1] WHO announces COVID-19 outbreak a pandemic, (2020).

[2] E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, Lancet Infect. Dis. (2020). https://doi.org/10.1016/S1473-3099(20)30120-1.

[3] COVID-19 Global Cases by Johns Hopkins University, 2020. https://www.gisaid.org/ epiflu-applications/global-cases-covid-19/.

[4] At the Top of the Covid-19 Curve, How Do Hospitals Decide Who Gets Treatment? - The New York Times, (n.d.). https://www.nytimes.com/2020/03/31/us/coronavirus-covid-triage-rationing-ventilators.html (accessed April 29, 2020).

[5] The Hardest Questions Doctors May Face: Who Will Be Saved? Who Won't? - The New York Times, (n.d.). https://www.nytimes.com/2020/03/21/us/coronavirus-medical-rationing.html (accessed April 29, 2020).

[6] L. Wynants, B.V. Calster, G.S. Collins, R.D. Riley, G. Heinze, E. Schuit, M.M.J. Bonten, J.A.A. Damen, T.P.A. Debray, M.D. Vos, P. Dhiman, M.C. Haller, M.O. Harhay, L. Henckaerts, N. Kreuzberger, A. Lohmann, K. Luijken, J. Ma, C.L.A. Navarro, J.B. Reitsma, J.C. Sergeant, C. Shi, N. Skoetz, L.J.M. Smits, K.I.E. Snell, M. Sperrin, R. Spijker, E.W. Steyerberg, T. Takada, S.M.J. van Kuijk, F.S. van Royen, C. Wallisch, L. Hooft, K.G.M. Moons, M. van Smeden, Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal, BMJ. 369 (2020). https://doi.org/10.1136/bmj.m1328.

[7] Z. yong Huang, S. Lin, L. li Long, J. yang Cao, F. Luo, W. cheng Qin, D. ming Sun, H. Gregersen, Predicting the morbidity of chronic obstructive pulmonary disease based on multiple locally weighted linear regression model with K-means clustering, Int. J. Med. Inf. 139 (2020) 104141. https://doi.org/10.1016/j.ijmedinf.2020.104141.

[8] X. Du, J. Min, C.P. Shah, R. Bishnoi, W.R. Hogan, D.J. Lemas, Predicting in-hospital mortality of patients with febrile neutropenia using machine learning models, Int. J. Med. Inf. 139 (2020) 104140. https://doi.org/10.1016/j.ijmedinf.2020.104140.

[9] T.S. Brisimi, T. Xu, T. Wang, W. Dai, I.C. Paschalidis, Predicting diabetes-related hospitalizations based on electronic health records, Stat. Methods Med. Res. 28 (2019) 3667–3682. https://doi.org/10.1177/0962280218810911.

[10] T.S. Brisimi, T. Xu, T. Wang, W. Dai, W.G. Adams, I.C. Paschalidis, Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach, Proc. IEEE. 106 (2018) 690–707. https://doi.org/10.1109/JPROC.2017.2789319.

[11] T.S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I.C. Paschalidis, W. Shi, Federated learning of predictive models from federated Electronic Health Records, Int. J. Med. Inf. 112 (2018) 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007.

[12] D. Morel, K.C. Yu, A. Liu-Ferrara, A.J. Caceres-Suriel, S.G. Kurtz, Y.P. Tabak, Predicting Hospital Readmission in Patients with Mental or Substance Use Disorders: A Machine Learning Approach, Int. J. Med. Inf. 139 (2020) 104136. https://doi.org/10.1016/j.ijmedinf.2020.104136.

[13] A.Y. Ng, Feature selection, L1 vs. L2 regularization, and rotational invariance, in: Proc. Twenty-First Int. Conf. Mach. Learn. ICML 2004, 2004: pp. 615–622. https://doi.org/10.1145/1015330.1015435.

[14] R. Chen, I.C. Paschalidis, A Robust Learning Approach for Regression Models Based on Distributionally Robust Optimization, J. Mach. Learn. Res. 19 (2018) 1–48.

[15] Datos Abiertos - Dirección General de Epidemiología | Secretaría de Salud | Gobierno | gob.mx, (n.d.). https://www.gob.mx/salud/documentos/datos-abiertos-152127 (accessed April 29, 2020).

[16] Calculadora de complicación de salud por COVID -19, (n.d.). http://www.imss.gob.mx/covid-19/calculadora-complicaciones (accessed June 25, 2020).

[17] Lineamiento estandarizado para la vigilancia epidemiológica y por laboratorio de la enfermedad respiratoria viral, (2020). https://www.gob.mx/cms/uploads/attachment/file/552972/Lineamiento_VE_y_Lab_Enf_Viral_20.05.20.pdf.

[18] Clinical progression of patients with COVID-19 in Shanghai, China, (n.d.). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7102530/ (accessed May 2, 2020).

[19] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong, Y. Zhao, Y. Li, X. Wang, Z. Peng, Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel

Coronavirus–Infected Pneumonia in Wuhan, China, JAMA. 323 (2020) 1061–1069. https://doi.org/10.1001/jama.2020.1585.
[20] E. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, Springer-Verlag, New York, 2009. https://doi.org/10.1007/978-0-387-77244-8.
[21] L.N. Sanchez-Pinto, L.R. Venable, J. Fahrenbach, M.M. Churpek, Comparison of variable selection methods for clinical predictive modeling, Int. J. Med. Inf. 116 (2018) 10–17. https://doi.org/10.1016/j.ijmedinf.2018.05.006.
[22] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422. https://doi.org/10.1023/A:1012487302797.
[23] salomonw/covid-predictors-mexico, GitHub. (n.d.). https://github.com/salomonw/covid-predictors-mexico (accessed May 2, 2020).
[24] W.J. Guan, Z.Y. Ni, Y. Hu, W.H. Liang, C.Q. Ou, J.X. He, L. Liu, H. Shan, C.L. Lei, D.S.C. Hui, B. Du, L.J. Li, G. Zeng, K.Y. Yuen, R.C. Chen, C.L. Tang, T. Wang, P.Y. Chen, J. Xiang, S.Y. Li, J.L. Wang, Z.J. Liang, Y.X. Peng, L. Wei, Y. Liu, Y.H. Hu, P. Peng, J.M. Wang, J.Y. Liu, Z. Chen, G. Li, Z.J. Zheng, S.Q. Qiu, J. Luo, C.J. Ye, S.Y. Zhu, N.S. Zhong, Clinical Characteristics of Coronavirus Disease 2019 in China, N. Engl. J. Med. (2020). https://doi.org/10.1056/NEJMoa2002032.
[25] M. Chung, A. Bernheim, X. Mei, N. Zhang, M. Huang, X. Zeng, J. Cui, W. Xu, Y. Yang, Z.A. Fayad, A. Jacobi, K. Li, S. Li, H. Shan, CT imaging features of 2019 novel coronavirus (2019-NCoV), Radiology. 295 (2020) 202–207. https://doi.org/10.1148/radiol.2020200230.
[26] A. Bernheim, X. Mei, M. Huang, Y. Yang, Z.A. Fayad, N. Zhang, K. Diao, B. Lin, X. Zhu, K. Li, S. Li, H. Shan, A. Jacobi, M. Chung, Chest CT Findings in Coronavirus Disease-19 (COVID-19): Relationship to Duration of Infection, Radiology. (2020) 200463. https://doi.org/10.1148/radiol.2020200463.
[27] E. Tartaglione, C.A. Barbano, C. Berzovini, M. Calandri, M. Grangetto, Unveiling COVID-19 from Chest X-ray with deep learning: a hurdles race with small data, (2020).
[28] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, W. Ji, Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR, Radiology. (2020) 200432. https://doi.org/10.1148/radiol.2020200432.
[29] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297. https://doi.org/10.1007/bf00994018.
[30] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
[31] K. Koh, S.-J. Kim, S. Boyd, Y. Lin, An Interior-Point Method for Large-Scale 1-Regularized Logistic Regression, 2007.
[32] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32. https://doi.org/10.1023/A:1010933404324.
[33] L. Breiman, J. Friedman, C. Stone, R. Olshen, Classification and regression trees, CRC Press. (1984).
[34] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., Association for Computing Machinery, New York, New York, USA, 2016: pp. 785–794. https://doi.org/10.1145/2939672.2939785.
[35] Tree Boosting With XGBoost — Why Does XGBoost Win “Every” Machine Learning Competition?, (n.d.). https://medium.com/syncedreview/tree-boosting-with-xgboost-why-does-xgboost-win-every-machine-learning-competition-ca8034c0b283 (accessed April 29, 2020).