Predictive Models of Mortality for Hospitalized COVID-19 Patients: Retrospective Cohort Study

Taiyao Wang, Aris Paschalidis, Quanying Liu, Yingxia Liu, Ye Yuan, Ioannis Ch Paschalidis

Submitted to: JMIR Medical Informatics
on: June 25, 2020

Table of Contents

Predictive Models of Mortality for Hospitalized COVID-19 Patients: Retrospective Cohort Study

Taiyao Wang[1, 2, 3] PhD; Aris Paschalidis[4]; Quanying Liu[5] PhD; Yingxia Liu[6] MD; Ye Yuan[7] PhD; Ioannis Ch Paschalidis[1, 2, 3] PhD

[1]Department of Electrical and Computer Engineering Boston University Boston US
[2]Department of Biomedical Engineering Boston University Boston US
[3]Center for Information and Systems Engineering Boston University Boston US
[4]Brown University Providence US
[5]Department of Biomedical Engineering University of Science and Technology Shenzen CN
[6]Third People's Hospital of Shenzhen Second Hospital Affiliated to Southern University of Science and Technology Shenzen CN
[7]School of Artificial Intelligence and Automation Huazhong University of Science and Technology Wuhan CN

Corresponding Author:
Ioannis Ch Paschalidis PhD
Department of Electrical and Computer Engineering
Boston University
8 Saint Mary's St
Boston
US

Abstract

Background: The novel 2019 coronavirus SARS-CoV-2 and its associated disease, COVID-19, have caused worldwide disruption, leading countries to take drastic measures. As the virus continues to spread, hospitals have struggled to allocate resources to the patients most at risk. In this context, it becomes important to develop models that can accurately predict the severity of infection for each hospitalized patient, helping to guide triage, planning, and resource allocation.

Objective: The aim of this study is to develop accurate models to predict mortality among hospitalized COVID-19 patients, using basic demographics and easily obtainable laboratory data.

Methods: A retrospective study of 375 hospitalized patients in Wuhan, China infected with COVID-19 was undertaken. The patients were randomly split into derivation and validation cohorts. Regularized logistic regression and support vector machine classifiers were trained on the derivation cohort and accuracy metrics (F1-scores) were computed on the validation cohort. Two types of models were developed: i) using laboratory findings from the entire length of stay at the hospital, and ii) using admission laboratory findings obtained no later than 12 hours after admission. The models were further validated on a multicenter external cohort of 542 patients.

Results: Of the 375 patients, 174 (46.4%) succumbed to the infection. The study cohort was composed of 60% (224/375) males and 40% (151/375) females, with a mean age of 58.83 years old. Models developed using patient data from throughout the length of stay had an accuracy as high as 97%, whereas models with admission laboratory variables had accuracy of up to 93%. The latter models developed using admission patient data predicted patient outcomes an average of 11.5 days in advance. Key variables such as lactate dehydrogenase, high-sensitivity C-reactive Protein, and the percent of lymphocytes in the blood were indicated by the models. In line with previous studies, age was also found to be an important variable in predicting mortality. In particular, the mean age of patients that survived COVID-19 infection (50.23 years) was significantly smaller than the mean age of patients (68.75 years) that did not survive the infection (P<.001).

Conclusions: Machine learning models can be successfully employed to accurately predict COVID-19 patient outcomes. Models achieve high accuracies and predict outcomes more than a week in advance, a promising result that can greatly aid hospitals in resource allocation.

(JMIR Preprints 25/06/2020:21788)
DOI: https://doi.org/10.2196/preprints.21788

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Predictive Models of Mortality for Hospitalized COVID-19 Patients: Retrospective Cohort Study

Taiyao Wang,[1] Ph.D., Aris Paschalidis,[2] Quanying Liu, [3] Ph.D., Yingxia Liu,[4] M.D., Ye Yuan, [5] Ph.D., and Ioannis Ch. Paschalidis,[1,*] Ph.D.

[1]Department of Electrical and Computer Engineering, Department of Biomedical Engineering, and Center for Information and Systems Engineering, Boston University, Boston, MA

[2]Brown University, Providence, RI

[3]Department of Biomedical Engineering, Southern University of Science and Technology. Shenzhen, 518000, China

[4]The Third People's Hospital of Shenzhen, Second Hospital Affiliated to Southern University of Science and Technology, Shenzhen, 518000, China

[5]Huazhong University of Science and Technology, Wuhan, China

[*]Corresponding Author
Ioannis Ch. Paschalidis,
Department of Electrical and Computer Engineering,
and Department of Biomedical Engineering,
Boston University
8 Saint Mary's St.,
Boston, MA 02215
USA
e-mail: yannisp@bu.edu
http://sites.bu.edu/paschalidis
Tel: 617-353-0434
Fax: 617-353-0190

Abstract

Background: The novel 2019 coronavirus SARS-CoV-2 and its associated disease, COVID-19, have caused worldwide disruption, leading countries to take drastic measures. As the virus continues to spread, hospitals have struggled to allocate resources to the patients most at risk. In this context, it becomes important to develop models that can accurately predict the severity of infection for each hospitalized patient, helping to guide triage, planning, and resource allocation.

Objective: The aim of this study is to develop accurate models to predict mortality among hospitalized COVID-19 patients, using basic demographics and easily obtainable laboratory data.

Methods: A retrospective study of 375 hospitalized patients in Wuhan, China infected with COVID-19 was undertaken. The patients were randomly split into derivation and validation cohorts. Regularized logistic regression and support vector machine classifiers were trained on the derivation cohort and accuracy metrics (F1-scores) were computed on the validation cohort. Two types of models were developed: i) using laboratory findings from the entire length of stay at the hospital, and ii) using admission laboratory findings obtained no later than 12 hours after admission. The models were further validated on a multicenter external cohort of 542 patients.

Results: Of the 375 patients, 174 (46.4%) succumbed to the infection. The study cohort was composed of 60% (224/375) males and 40% (151/375) females, with a mean age of 58.83 years old. Models developed using patient data from throughout the length of stay had an accuracy as high as 97%, whereas models with admission laboratory variables had accuracy of up to 93%. The latter models developed using admission patient data predicted patient outcomes an average of 11.5 days in advance. Key variables such as lactate dehydrogenase, high-sensitivity C-reactive Protein, and the percent of lymphocytes in the blood were indicated by the models. In line with previous studies, age was also found to be an important variable in predicting mortality. In particular, the mean age of patients that survived COVID-19 infection (50.23 years) was significantly smaller than the mean age of patients (68.75 years) that did not survive the infection ($P<.001$).

Conclusions: Machine learning models can be successfully employed to accurately predict COVID-19 patient outcomes. Models achieve high accuracies and predict outcomes more than a week in advance, a promising result that can greatly aid hospitals in resource allocation.

Keywords: Coronavirus; COVID-19; Mortality; Wuhan; China; Machine Learning; Logistic Regression; Support Vector Machine; Predictive Modeling.

Introduction

The ongoing pandemic due to the novel 2019 coronavirus SARS-CoV-2 has caused worldwide disruption; nations have imposed drastic measures, and the global economy has suffered [1]. The virus causes a disease called COVID-19, and elicits symptoms such as coughs, fever, chills, and a range of respiratory symptoms [2]. As of the end of July, 2020, the total number of confirmed cases have surpassed 15 million and the total number of deaths is approaching 650,000 [3,4].

As the virus continues to proliferate, governments, institutions, and hospitals have struggled to allocate resources such as tests, hospital beds, intensive care unit (ICU) beds, and ventilators. A significant amount of work has already been done to predict and track the spread of the virus [3–8]. Recent and ongoing efforts sought to understand the biomarkers and comorbidities associated with severe disease [9–12]. This work has been important in aiding hospitals to classify patients in terms of their risk. However, the infrastructure to predict hospitalization, mortality, or other patient outcomes is lacking. Predicting such outcomes is essential as it allows clinicians to make informed decisions regarding patients at risk. For example, clinicians can ensure that the proper resources are allocated for patients who are more likely to require critical care and the use of ventilators.

Using blood samples from patients from the Tongji Hospital in Wuhan, China, we utilized supervised machine learning methods to predict mortality following hospitalization. Such machine learning models have been used frequently in the literature for a variety of applications. Some examples include predicting death of patients with sepsis [13,14], identifying patients at high risk of emergency hospital admissions [15], predicting hospitalization due to heart disease [16,17], and predicting diabetes complications [18,19].

The aim of this retrospective cohort study was to develop accurate models to predict mortality among hospitalized COVID-19 patients, using basic demographics and easily obtainable laboratory data.

Methods

Data Collection

Data were collected between January 10, 2020 and February 18, 2020 from patients admitted to the Tongji Hospital in Wuhan, China. Data collection was approved by the Tongji Hospital Ethics Committee. Records included epidemiological, demographic, clinical, and laboratory results, as well as mortality following infection with COVID-19. Data originating from pregnant and breast-feeding women, patients younger than 18 years of age, and records with more than 20% of missing data, were excluded from the analysis [20].

Pre-processing

Prior to any model development, several preprocessing measures were undertaken. Variables were standardized by subtracting the mean and dividing by the standard deviation. Variable elimination was performed in hopes of reducing the complexity of the resulting model, improving out-of-sample performance, and enhancing interpretability. Redundant variables and variables with more than 30% data missing were removed. In addition, we computed pairwise Spearman correlations between variables and removed one of the variables if the absolute correlation coefficient was greater than 0.8. Furthermore, the missing data of the remaining variables were imputed using the median value of the respective variables. This allowed us to include as many patients as possible in our analysis and is a well-documented and popular method of inferring missing values.

Model Development

A total of 375 patients were utilized to develop the models. These patients were split into two groups, a training and validation set. The training set was used to train and develop the models and the validation set was used to determine the accuracy of each model. Unless otherwise noted, 70% of the data was reserved for the training set and the other 30% was reserved for the validation set. Data were first split into training and validation sets, and then feature selection was performed to remove several variables. Models were learned using the training set and tested on the validation set. This process was repeated five times and the average performance (and its standard deviation) was calculated.

Feature selection was performed using ℓ_1-norm regularization and recursive feature elimination with cross-validation. Specifically, we run ℓ_1-regularized logistic regression (LR) and obtained the coefficients of the model. We then eliminated the variable with the smallest absolute coefficient and re-run LR to obtain a new model. We kept iterating in this fashion, to select a model that maximizes a metric equal to the mean performance minus its standard deviation in a validation dataset.

Model Selection

Two different types of regularized models were utilized in this analysis: an ℓ_1-regularized logistic regression (L1LR) model and an ℓ_1-regularized support vector machine (L1SVM) model. Models were initially fit to patient data that was collected at any time during the length of stay at the hospital. However, as there was a possibility that some laboratory tests were done close to the outcome (death or survival), models were also fit to patient data from at most 12 hours after admission. By doing this, we could ensure that predictions were made about patients' outcomes as far in advance as possible.

Logistic regression, in addition to a prediction, provides the likelihood associated with the predicted outcome, which can be used as a confidence measure in decision making.

Model Performance

The performance of models was evaluated by calculating the weighted F1 score on the validation set. The weighted F1 score is defined as the weighted mean of the F1 score of the positive and negative classes, where the F1 score is defined as the harmonic mean of the precision and the recall. The precision (or Positive Predictive Value – PPV) can be expressed as the ratio of the true positives over the sum of the true positives and false positives. The recall is the true positive rate (i.e., the ratio of the true positives over the sum of the true positives and false negatives). The weighted F1 score,

unlike the F1 score, considers all the possible outcomes (in this case survival or death). This can combat potential class imbalance issues and evaluates whether the model accurately predicts mortality and survival, both of which are important in our context. In particular, whereas identifying those with higher mortality risk can help direct more resources and attention to those patients, identifying who is not at risk is also helpful and can free up resources and time spent on those lower-risk patients. In addition to the weighted F1 score, we also determined the Positive Predictive Value (PPV) and the Negative Predictive Value (NPV) – the latter defined as the ratio of the true negatives over the predicted negatives (or precision of the negative class).

Furthermore, and to gain additional insight into the role of specific variables, we developed a "binarized" counterpart to our sparse LR model. Specifically, we defined a threshold for each variable (using the normal range of the variable) and devised a model with each variable being either 0 (normal) or 1 (abnormal). For this model, we computed the Odds Ratio (OR) for each variable, which quantifies how the odds of mortality get scaled by the variable being normal vs. abnormal, while controlling for the remaining variables.

Statistical Power and External Validation

To assess whether our study cohort size was sufficiently large for the models we derived, we conducted a multiple logistic regression power analysis [21]. This tests the null hypothesis that a specific variable has an LR coefficient equal to zero vs. the coefficient value obtained by the model. We set the Type I error probability to 0.05 and the Type II error probability to 0.2 (statistical power of 0.8), from which we obtained a minimal sample size for the variable.

Further, to demonstrate that our models are generalizable, we validated our models on a multicenter external dataset. This dataset contains 432 patients from Shenzhen, China and 110 from Wuhan, China. The dataset contained very limited information encompassing three patient lab tests, the time of the lab tests, the discharge time, and the outcome. Given this limited information, we were only able to validate our best-performing L1SVM model which uses these three lab values.

Results

Patient Demographics and Laboratory Tests

Multimedia Appendix Table 1 details patient demographics in addition to various laboratory values for the full patient population. The average age of patients was 58.83 years. The mean age of patients who survived COVID-19 infection (50.23 years) was significantly smaller than the mean age of patients (68.75 years) who did not survive the infection (P<.001). The proportion of males (224/375, 60%) and females (151/375, 40%) in the study cohort was similar. However, more male patients succumbed to infection (126/174, 72%, P<.001).

Several laboratory tests were found to have statistically different values among patients who survived and succumbed to COVID-19 infection. Patients that succumbed to infection had lactate dehydrogenase (LDH) values roughly 4 times larger than patients who survived (755.58 compared to 215.77, P<.001). Patients who died also had a significantly smaller percent of lymphocytes and eosinophils in the blood (P<.001). Furthermore, the mean level of hypersensitive C-reactive protein (hs-CRP) in patients who died was significantly higher than in patients who survived (P<.001).

As detailed in the Methods section, two different approaches were utilized to model the data. The first approach was to use blood tests of patients throughout their length of stay at the hospital.

Although this ensured there were few missing data points, some of the blood samples were tested close to the outcome (death or discharge from the hospital). In order to predict a patient's outcome in advance, a second approach was to use laboratory tests that were obtained at most 12 hours from admission to the hospital.

Models Using All Laboratory Tests

We first present the results of our predictive models using all laboratory tests. These models were developed as noted in the Methods section. Of the total 375 patients, 24 (6.4%) patients had incomplete measurements and were omitted, leaving a total of 351 (93.6%) patients for model development. Table 1 highlights the accuracy of the best performing models using all patient laboratory tests on the validation set (30% of the 351 patients) and on the external test set described in Methods. A complete list of the models and their accuracies is provided in the Multimedia Appendix.

Table 1: Performance of the best models.

Performance	L1LR 4[a]	L1SVM 3[b]
Validation set weighted F1-score (%), mean (SD)	96.98 (0.93)	97.36 (1.10)
External test set weighted F1-score (%)		94.55

[a]L1LR 4: ℓ_1-regularized logistic regression model utilizing 4 variables selected by recursive feature selection.

[b]L1SVM 3: ℓ_1-regularized support vector machine model utilizing 3 variables selected by recursive feature selection.

Both the logistic regression model (L1LR) and the support vector machine (L1SVM) model presented in Table 1 performed very well with accuracies greater than 95% and a small standard deviation. The L1LR 4 model had an average validation PPV of 97.61% and an average validation NPV of 96.31%. The L1SVM 3 model had similarly high average PPV and NPV of 98.27% and 96.71%, respectively. On the multicenter external test set, the L1SVM model's accuracy remained high, with an accuracy of 94.55%. Furthermore, both models used a small number of variables in their predictions. The variables each model used, and the corresponding weight of each variable, is reported in Table 2. The logistic regression model utilized four variables: LDH, an enzyme found in most living cells and typically released when there is tissue damage; the percent of lymphocytes, a class of immune molecules found in the body; hs-CRP, a protein that is often used as an indication of heart disease and increases with inflammation and infection; and albumin, one of the main proteins found in the blood that is important in regulating the pressure of red blood cells, as well as transporting nutrients, proteins, and other molecules. The L1SVM model used the same variables but did not include albumin.

Table 2: Variables and coefficients in the best models.

L1LR 4[a]		L1SVM 3[b]	
Variable	Coefficient	Variable	Coefficient
LDH	1.35	LDH	1.44
Percent lymphocyte	-0.86	Percent lymphocyte	-0.47

Hs-CRP	0.74	Hs-CRP	0.34
Albumin	-0.64		

[a]L1LR 4: ℓ_1-regularized logistic regression model utilizing 4 variables selected by recursive feature selection.
[b]L1SVM 3: ℓ_1-regularized support vector machine model utilizing 3 variables selected by recursive feature selection.

The coefficients obtained by both methods are comparable since the variables are standardized. Hence, a larger absolute coefficient indicates that the corresponding variable is a more significant predictor. Positive (negative) coefficients imply positive (negative) correlation with the outcome. Of the variables selected by our models, LDH was considered to be the most important (binarized L1LR 4 odds ratio [OR] 55.62, 95% CI 11.41-270.97). The next most important variables were the percent of lymphocytes (binarized L1LR 4 odds ratio [OR] 32.17, 95% CI 5.99-172.90) and hs-CRP (binarized L1LR 4 odds ratio [OR] 13.12, 95% CI 3.65-47.23). Lastly, the L1LR model found that albumin was important in predicting mortality (binarized L1LR 4 odds ratio [OR] 4.08, 95% CI 1.45-11.48). In order to calculate these ORs, and as detailed under Methods, we used a binarized model with the following thresholds: LDH values ≥ 250 were set to 1, 0 otherwise; percent lymphocyte values < 20 were set to 1, 0 otherwise; hs-CRP values ≥ 10 were set to 1, 0 otherwise; albumin values < 34 were set to 1, 0 otherwise.

As outlined under Methods, a power analysis was performed for the L1LR 4-variable model, indicating that our sample size of 351 patients was sufficient. Specifically, this power analysis indicated that 21 patients are sufficient to find the LR coefficient for LDH, 63 patients are sufficient for hs-CRP, 61 patients suffice for percent of lymphocytes, and 162 patients suffice for albumin.

In addition to the models reported previously, we also trained models with several important variables removed. More specifically, we removed LDH, albumin, and D-D dimer, a protein that is produced by the degradation of a blood clot. The accuracy of these models was slightly lower than models that included these factors. Furthermore, the more variables we removed, the worse the accuracy. The validation accuracy of the L1LR model with LDH removed was 94.27% (SD 2.44%), the validation accuracy of the L1LR model with LDH and albumin removed was 93.81% (SD 2.52%), and the validation accuracy of the L1LR model with LDH, albumin, and D-D dimer removed was 93.46% (SD 2.89%) (cf. Multimedia Appendix Table 2). The models highlighted several other important factors not previously indicated as important such as prothrombin activity, a protein used in blood clot formation, the platelet count, the count of the main protein that makes up blood clots, and age, to name a few. After these variables were removed, the two most important factors were hs-CRP and the percent of lymphocytes. When fitting a model to the data using only these two factors, the validation accuracy of the model was 94.87% (SD 1.76%).

Models Using Tests at Most 12 Hours From Admission

In order to predict the outcome of a patient soon after admission to the hospital, we ran several L1SVM models using laboratory tests with patients at most 12 hours from admission. More specifically, we first ran an ℓ_1-regularized logistic regression in order to perform feature selection and then fed the selected features into an ℓ_1-regularized support vector machine model. The average time between admission and the time the laboratory test was conducted was 8.4 hours (SD 2.6 hours). Furthermore, the average time between the time of the laboratory test and the outcome was 11.5 days (SD 7.5 days).

Table 3 details the average F1 score and standard deviation for a select number of the models developed on data collected at most 12 hours from admission. Multimedia Appendix Table 4 reports the variables selected by these models. For all models, the L1SVM was performed five times and optimized using a validation set. Of the 375 total patients, 114 (30.4%) patients had missing data and were excluded, leaving 261 (69.6%) patients for analysis. 90% of the 261 patients were kept for training and 10% of the patients were kept as a validation set. As before, models were fit using all the variables, a limited number of variables, and all variables other than LDH, albumin, and D-D dimer.

Table 3: Performance of select models.

Model	Validation set weighted F1-score (%), mean (SD)
L1SVM all[a]	90.39 (3.25)
L1SVM 7[b]	94.08 (1.81)
L1SVM no LDH, albumin[c]	89.65 (4.30)
L1SVM no LDH, albumin, D-D dimer[d]	89.64 (4.89)

[a]L1SVM all: ℓ_1-regularized SVM model developed using all the variables in the data set.
[b]L1SVM 7: ℓ_1-regularized SVM model utilizing 7 variables.
[c]L1SVM no LDH, albumin: ℓ_1-regularized SVM model developed using all variables but LDH and albumin.
[d]L1SVM no LDH, albumin, D-D dimer: ℓ_1-regularized SVM model developed using all variables but LDH, albumin, and D-D dimer.

All models performed well with accuracies above 87% and standard deviations less than 6%. The number of variables each model used varied greatly. The L1SVM All model utilized 18 of the variables provided in the dataset, the L1SVM 7 model utilized 7 variables, the L1SVM model without LDH and albumin utilized 10 variables, and the L1SVM model with no LDH, no albumin, and no D-D dimer utilized 12 variables. Of these models, the model that included 7 variables including LDH, albumin, and D-D dimer performed the best, with an accuracy of 93.45 % (SD 1.95%). When LDH and albumin were removed from the model, the accuracy dropped roughly 4% and when D-D dimer was removed, the accuracy dropped another 0.4%.

These L1SVM models highlighted several key variables that were not indicated in the models that included all laboratory tests. LDH and hs-CRP, as in the models that utilized all variables were consistently two of the most important markers. However, the percent of lymphocytes found in the blood, did not consistently appear to be important. Interestingly, the number of neutrophils, a different class of immune molecule, in the blood was deemed an important variable.

Discussion

Principal Findings

Our developed L1LR and L1SVM models were able to accurately predict a patient's outcome reaching validation weighted F1 scores as high as 97%. In general, models that utilized laboratory tests from the duration of patients' visit were more accurate than models that were restricted to laboratory tests at most 12 hours after admission. However, even when data were restricted, our models reached accuracies as high as 94%. Such models are more useful because they make predictions upon admission of the patient and, thus, provide enough lead time for making staffing

and resource allocation decisions. As the length of stay of most patients was more than a week, our models can therefore predict, with accuracy exceeding 90%, the outcome of a patient more than a week in advance.

Our patient cohort represented in many ways a typical cohort of hospitalized COVID-19 patients. In particular, the individuals who succumbed to infection tended to be older and male [22–25]. However, the rate of mortality in our study cohort was larger; close to 50% (174/375) of the patients admitted to the hospital passed away. This is likely due to the fact that Tongji Hospital admitted a higher rate of severe and critical cases in Wuhan, China.

The performance of the L1SVM model utilizing all patient lab tests on an external multicenter dataset suggests that our models are generalizable. The performance of the model dropped less than 3% when tested on the external dataset compared to the validation set. This indicates that our model could be used by other hospitals around the globe in an effort to better understand the risk associated with each COVID-19 patient.

Particularly important was the models' ability to perform well with a small number of predictors. Moreover, when certain key predictors such as LDH, albumin, and D-D dimer were removed due to the variables' tendency to exhibit abnormalities at a very late stage of the disease, when the outcome is inevitable, the models still performed well. The ability of the models to perform well, even with few variables, can prove particularly useful as it allows for easy interpretation. Furthermore, it ensures that predictions can be made even when the outcome is not apparent to a sufficiently experienced physician.

A recent study aimed to develop a predictive model based on a few key variables [20]. This study utilized different machine learning methods and opted to create a decision tree. The authors found that LDH, the percent of lymphocytes, and hs-CRP were important predictors of mortality, three variables that we also found were important. The study's models were very accurate with F1-scores around 95%. The key difference between our study is that we utilize laboratory tests 12 hours after admission and test the robustness of the models to the absence of several key variables. This allows us to be confident that our models can accurately predict patient outcomes well in advance, in the absence of key variables, and even when the outcome may not be obvious to a trained medical doctor.

Limitations

One of the main limitations of this study was the relatively targeted study cohort used to derive the models. These patients lived in Wuhan, China, the original epicenter of the novel 2019 coronavirus SARS-CoV-2. Still one of our models was validated on an external multicenter cohort of patients from Wuhan and Shenzen, which suggests that this model could generalize to other patient cohorts, especially in China. It is less clear how well the models generalize to cohorts in other countries where patient characteristics and care practices may differ.

Conclusions

We developed two state-of-the-art supervised machine learning models in an effort to predict the outcome of infection with the novel 2019 coronavirus SARS-CoV-2. We were able to accurately predict mortality with greater than 90% accuracy and identified several important predictors.

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 thank doctors at Tongji Hospital in Wuhan, China, and Dr. George Velmahos at the Massachusetts General Hospital for useful discussions.

Conflicts of Interest

None declared

Abbreviations

Hs-CRP: hypersensitive C-reactive protein
LDH: lactate dehydrogenase
LR: Logistic Regression
SVM: Support Vector Machine
L1LR: ℓ_1-regularized logistic regression
L1SVM: ℓ_1-regularized support vector machine
PPV: Positive Predictive Value
NPV: Negative Predictive Value

Multimedia Appendix

Multimedia Appendix Table 1: Select patient demographics and laboratory tests.

Multimedia Appendix Table 2: Performance of all logistic regression models evaluated on all laboratory tests.

Multimedia Appendix Table 3: Performance of all SVM models evaluated on all laboratory tests.

Multimedia Appendix Table 4: Variables and coefficients of select models evaluated using laboratory tests within 12 hours of admission.

References

1. Coronavirus Disease (COVID-19) - events as they happen [Internet]. 2020 [cited 2020 May 29]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen

2. CDC. Coronavirus Disease 2019 (COVID-19) – Symptoms [Internet]. Centers for Disease Control and Prevention. [cited 2020 May 29]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

3. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases Elsevier; 2020 May 1;20(5):533–534. PMID:32087114

4. COVID-19 Map [Internet]. Johns Hopkins Coronavirus Resource Center. [cited 2020 Jul 25]. Available from: https://coronavirus.jhu.edu/map.html

5. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, Eggo RM, Sun F, Jit M, Munday JD, Davies N, Gimma A, van Zandvoort K, Gibbs H, Hellewell J, Jarvis CI, Clifford S, Quilty BJ, Bosse NI, Abbott S, Klepac P, Flasche S. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. The Lancet Infectious Diseases 2020 May 1;20(5):553–558. [doi: 10.1016/S1473-3099(20)30144-4]

6. Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W. Using Reports of Symptoms and Diagnoses on Social Media to Predict COVID-19 Case Counts in Mainland China: Observational Infoveillance Study. Journal of Medical Internet Research 2020;22(5):e19421. PMID:32452804

7. Gong M, Liu L, Sun X, Yang Y, Wang S, Zhu H. Cloud-Based System for Effective Surveillance and Control of COVID-19: Useful Experiences From Hubei, China. Journal of Medical Internet Research 2020;22(4):e18948. PMID:32287040

8. Yasaka TM, Lehrich BM, Sahyouni R. Peer-to-Peer Contact Tracing: Development of a Privacy-Preserving Smartphone App. JMIR mHealth and uHealth 2020;8(4):e18936. PMID:32240973

9. Guo W, Li M, Dong Y, Zhou H, Zhang Z, Tian C, Qin R, Wang H, Shen Y, Du K, Zhao L, Fan H, Luo S, Hu D. Diabetes is a risk factor for the progression and prognosis of COVID-19. Diabetes Metab Res Rev 2020 Mar 31;e3319. PMID:32233013

10. Frater JL, Zini G, d'Onofrio G, Rogers HJ. COVID-19 and the clinical hematology laboratory. Int J Lab Hematol 2020 Apr 20; PMID:32311826

11. Lippi G, Plebani M, Henry BM. Thrombocytopenia is associated with severe coronavirus disease 2019 (COVID-19) infections: A meta-analysis. Clinica Chimica Acta 2020 Jul 1;506:145–148. PMID:32178975

12. Qin C, Zhou L, Hu Z, Zhang S, Yang S, Tao Y, Xie C, Ma K, Shang K, Wang W, Tian D-S. Dysregulation of Immune Response in Patients With Coronavirus 2019 (COVID-19) in Wuhan, China. Clin Infect Dis [Internet] [cited 2020 May 14]; PMID:32161940

13. Jaimes F, Farbiarz J, Alvarez D, Martínez C. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. Crit Care 2005 Feb 17;9(2):R150. PMID:15774048

14. Vieira SM, Mendonça LF, Farinha GJ, Sousa JMC. Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. Applied Soft Computing 2013 Aug 1;13(8):3494–3504. [doi: 10.1016/j.asoc.2013.03.021]

15. Bottle A, Aylin P, Majeed A. Identifying Patients at High Risk of Emergency Hospital Admissions: A Logistic Regression Analysis. J R Soc Med SAGE Publications; 2006 Aug 1;99(8):406–414. PMID:16893941

16. Dai W, Brisimi TS, Adams WG, Mela T, Saligrama V, Paschalidis ICh. Prediction of hospitalization due to heart diseases by supervised learning methods. International Journal of Medical Informatics 2015 Mar 1;84(3):189–197. PMID:25497295

17. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, Andreini D, Budoff

MJ, Cademartiri F, Callister TQ, Chang H-J, Chinnaiyan K, Chow BJW, Cury RC, Delago A, Gomez M, Gransar H, Hadamitzky M, Hausleiter J, Hindoyan N, Feuchtner G, Kaufmann PA, Kim Y-J, Leipsic J, Lin FY, Maffei E, Marques H, Pontone G, Raff G, Rubinshtein R, Shaw LJ, Stehli J, Villines TC, Dunning A, Min JK, Slomka PJ. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. Eur Heart J Oxford Academic; 2017 Feb 14;38(7):500–507. PMID:27252451

18. Brisimi TS, Xu T, Wang T, Dai W, Paschalidis IC. Predicting diabetes-related hospitalizations based on electronic health records. Stat Methods Med Res SAGE Publications Ltd STM; 2019 Dec 1;28(12):3667–3682. PMID:30474497

19. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, De Cata P, Chiovato L, Bellazzi R. Machine Learning Methods to Predict Diabetes Complications. J Diabetes Sci Technol SAGE Publications Inc; 2018 Mar 1;12(2):295–302. PMID:28494618

20. Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, Sun C, Tang X, Jing L, Zhang M, Huang X, Xiao Y, Cao H, Chen Y, Ren T, Wang F, Xiao Y, Huang S, Tan X, Huang N, Jiao B, Cheng C, Zhang Y, Luo A, Mombaerts L, Jin J, Cao Z, Li S, Xu H, Yuan Y. An interpretable mortality prediction model for COVID-19 patients. Nature Machine Intelligence Nature Publishing Group; 2020 May 14;1–6. [doi: 10.1038/s42256-020-0180-7]

21. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Stat Med 1998 Jul 30;17(14):1623–1634. PMID:9699234

22. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, Barnaby DP, Becker LB, Chelico JD, Cohen SL, Cookingham J, Coppa K, Diefenbach MA, Dominello AJ, Duer-Hefele J, Falzon L, Gitlin J, Hajizadeh N, Harvin TG, Hirschwerk DA, Kim EJ, Kozel ZM, Marrast LM, Mogavero JN, Osorio GA, Qiu M, Zanos TP. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA American Medical Association; 2020 May 26;323(20):2052–2059. PMID:32320003

23. Onder G, Rezza G, Brusaferro S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. JAMA American Medical Association; 2020 May 12;323(18):1775–1776. PMID:32203977

24. Jordan RE, Adab P, Cheng KK. Covid-19: risk factors for severe disease and death. BMJ [Internet] British Medical Journal Publishing Group; 2020 Mar 26 [cited 2020 May 29];368. PMID:32217618

25. Smith-Ray R, Roberts EE, Littleton DE, Singh T, Sandberg T, Taitel M. Distribution of Patients at Risk for Complications Related to COVID-19 in the United States: Model Development Study. JMIR Public Health and Surveillance 2020;6(2):e19606. PMID:32511100

Supplementary Files

Untitled.
URL: https://asset.jmir.pub/assets/2b9420b686b3510a2639f9ca73cab3f0.docx

Untitled.
URL: https://asset.jmir.pub/assets/600c28d5e2e08277dc4083edb4a336fa.docx

Untitled.
URL: https://asset.jmir.pub/assets/b1eed9a82333aca0f4597883bb93f6fa.pdf

Multimedia Appendixes

Untitled.
URL: https://asset.jmir.pub/assets/f184cd47d1a0e820c09168c1d152d64a.docx