ARTICLE

Development of a prognostic model for mortality in COVID-19 infection using machine learning

Adam L. Booth [ID][1] · Elizabeth Abels[1] · Peter McCaffrey [ID][1]

Abstract

Coronavirus disease 2019 (COVID-19) is a novel disease resulting from infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which has quickly risen since the beginning of 2020 to become a global pandemic. As a result of the rapid growth of COVID-19, hospitals are tasked with managing an increasing volume of these cases with neither a known effective therapy, an existing vaccine, nor well-established guidelines for clinical management. The need for actionable knowledge amidst the COVID-19 pandemic is dire and yet, given the urgency of this illness and the speed with which the healthcare workforce must devise useful policies for its management, there is insufficient time to await the conclusions of detailed, controlled, prospective clinical research. Thus, we present a retrospective study evaluating laboratory data and mortality from patients with positive RT-PCR assay results for SARS-CoV-2. The objective of this study is to identify prognostic serum biomarkers in patients at greatest risk of mortality. To this end, we develop a machine learning model using five serum chemistry laboratory parameters (c-reactive protein, blood urea nitrogen, serum calcium, serum albumin, and lactic acid) from 398 patients (43 expired and 355 non-expired) for the prediction of death up to 48 h prior to patient expiration. The resulting support vector machine model achieved 91% sensitivity and 91% specificity (AUC 0.93) for predicting patient expiration status on held-out testing data. Finally, we examine the impact of each feature and feature combination in light of different model predictions, highlighting important patterns of laboratory values that impact outcomes in SARS-CoV-2 infection.

Introduction

Coronavirus disease 2019 (COVID-19) is a novel disease resulting from infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that as of June 5, 2020 has resulted in 394,887 deaths worldwide since its emergence in late 2019 [1]. Currently, there is no vaccine or highly effective therapy against SARS-CoV-2 and there is a dearth of prognostic tools to identify which patients are at increased risk of death. Thus, the effective management of COVID-19 requires the identification of patient mortality risk and the ability to surface such an identification

algorithm from among the new and rapidly accumulating data now available regarding this disease. Previous studies have evaluated laboratory, radiological, and observational findings in COVID-19 patients but have had limited success in determining which patients will have a poor outcome [2]. More specifically, laboratory data such as coagulation factors, serum proteins, serum electrolytes, and cytokines have been studied [2–5] and some of these, such as ferritin and c-reactive protein (CRP), may offer early clinical signs of severe disease onset [2–4]. For example, Zhou et al. [5] observed significantly greater increases in serum ferritin, procalcitonin, and CRP in very severe disease, however, they suggest this could be due to a secondary bacterial infection. Wang [2] found that levels of CRP were positively correlated with the size of lung lesions on imaging and severity of disease presentation. Similarly, Tan et al. [3] demonstrated that early stage disease with significant CRP increase was a predictor of early, severe COVID-19. That being said, there has not been an integrative model that can capture the combined effect of multiple such biomarkers and the interactions between them.

✉ Peter McCaffrey
   pemccaff@UTMB.EDU

[1]   University of Texas Medical Branch, Galveston, TX, USA

The use of machine learning (ML) in medicine is not a new concept, particularly in the area of pathology where rapid growth in the synergistic domains of laboratory information systems (LIS) and digital slide scanning has made large collections of valuable healthcare data available for statistical modeling. In a study by Luo et al. [6], authors demonstrated that using patient demographics and other laboratory test results ML could accurately predict normal or abnormal serum ferritin results. Huang and Wu [7] used deep convolutional neural networks to classify bacterial colonies by morphology prior to employing advanced identification techniques (automated systems and mass spectrometry) saving laboratory technician time and expertise. Furthermore, in the realm of anatomic pathology, studies have shown that ML can be used with significant success in the diagnosis of prostate adenocarcinoma, breast carcinoma, skin cancers, and typing colorectal polyps among other malignant diagnoses [8–10].

In the midst of a rapidly evolving scenario such as the COVID-19 pandemic, ML constitutes a particularly powerful method to surface insights directly from newly generated testing and patient data where guidelines have yet to be established. ML has the ability to improve diagnostic performance relative to hand-selected biomarkers by selecting groups of relevant biomarkers and more consistently capturing both their relative importance to prediction and their interactions among one another [11]. Moreover, ML is amenable to model inspection and interpretability (depending upon the technique used), further allowing models to be evaluated and inspected by experts as a means to fuel and guide subsequent clinical decisions. Lastly, ML performs well in circumstances where structured numerical data are readily available. Our institution performs a considerable amount of SARS-CoV-2 testing presently exceeding 80,000 samples processed between March 13, 2020 and June 5, 2020, resulting in one of the largest data sets of SARS-CoV-2 positive patients in our state. With our institution serving as a testing hub, we generate significant quantities of data that will uniquely allow our institution to identify otherwise unforeseen relationships between test results, laboratory values, and patient outcome.

Materials and methods

Data and outcomes selection

Approval for the study was obtained from the University of Texas Medical Branch Institutional Review Board (IRB# 20-0125). A retrospective query of the LIS was performed for patients with positive testing for SARS-CoV-2 using any of the following platforms: Abbott ID Now, Abbott M2000, Hologic Panther Fusion, Cepheid Gene Xpert, or our own

Table 1 Laboratory search parameters queried in the laboratory information system.

Laboratory search parameters	
Acute care PO_2	D-dimer
Alanine aminotransferase	Estimated glomerular filtration rate
Albumin	Estimated glomerular filtration rate, African American
Alkaline phosphatase	Ferritin
Anion gap	Glucose
Aspartate aminotransferase	Hemolysis index
Bilirubin, total	International normalized ratio, protime
Blood urea nitrogen	Lactate dehydrogenase
C-reactive protein	Lactic acid
Calcium	Potassium
Chloride	Protein, total
CO_2, total	Protime
Creatinine	Sodium

RT-PCR Laboratory Developed Test using SQL [12]. In addition, patient mortality was obtained by querying our Enterprise Data Warehouse for "Deceased" status among SARS-CoV-2 positive patients. Twenty-six serum chemistry and blood gas laboratory parameters (Table 1) were collected and assessed for sparsity, preserving only laboratory tests for which at least 25% of patients had measured values within 14 days following a positive SARS-CoV-2 test and excluding any laboratory values captured within 48 h of death. All patients with positive SARS-CoV-2 test results which were admitted to our hospital for care were included. Part of the rationale for doing this is that patients are more likely to be comparable on latent variables such as comorbid illnesses that we did not explicitly capture via laboratory results. Additionally, filtering on admitted patients increases the density of lab results across our tests of interest, thereby making us less reliant on imputation than if we had not filtered patients in this way. Tests within 48 h of death were excluded to maximize clinical prognostication while preserving sensitivity, thus allowing clinical action to thwart potential mortality. Furthermore, for patients who have had multiple laboratory measurements within the selection window, the earliest laboratory result is used for analysis. All analyses were performed using Python 3.7 and the scikit-learn, pandas, and shap packages.

Model development

From among the 26 laboratory values, multivariate feature imputation was performed using scikit-learn's IterativeImputer method to replace absent laboratory values with probabilistic numerical results (Supplementary fig., Missingness Plot). This method models each feature with missing values as a function of other features and uses the resulting best fit function to estimate missing values. This is

done in a round-robin fashion such that, at each step, a feature column is designated as a prediction output (y), while the other features are collectively considered a feature matrix X and used for a regression of the form $\hat{y} = f(X)$. After initializing missing values to their feature-wise mean, we used Bayesian ridge regression as our multivariate model followed by ten cycles of imputation over the entire feature table. We retained y (expired versus non-expired) as a feature during imputation. We then performed 1000 bootstrap samplings of our imputed data set. Including the dependent variable when imputing missing values among independent variables is motivated by the fact that there is presumed to be a relationship between these variables and excluding the dependent variable from imputation can produce biased estimates, which suppress imputed independent variables toward the null hypothesis. Thus, inclusion of the dependent variable in imputation works to preserve correlations by allowing imputation to model them based upon a full view of the data captured and not including all relevant variables—dependent or otherwise—would artificially bias imputation by blinding it to such correlations. This topic is discussed in detail in both Graham and Enders [13, 14].

Following multiple imputation, data were shuffled and divided into training (80%) and testing (20%) subsets and a logistic regression classifier was trained to predict expiration status based upon these laboratory values. Since most SARS-CoV-2 positive patients were not deceased, class weighting was applied to increase model penalty for failing to correctly identify patients who would expire. Using the model trained on all 26 laboratory values, we then examined regression coefficients as a measure of feature importance to understand the relative influence of each input laboratory value on the model's final prediction. We then selected the subset of five laboratory values to which the model assigned the highest weights after which the model was retrained. Five values were selected in an attempt to provide a simple and parsimonious set of common laboratory tests. Theorizing that these laboratory values may have nonlinear interactions, we then trained a support vector machine (SVM) using a radial basis function kernel from this same set of five laboratory values.

Machine learning

The SVM we trained for this task is a nonlinear model and, as such, it does not lend itself as transparently as a linear model does to interpretation. To better understand how the trained SVM model behaves, we implemented Shapley additive explanations (SHAP) using Python's SHAP package [15]. SHAP is an ML interpretability technique growing in popularity for its ability to capture the marginal contribution of each feature to a model's ultimate output even as those contributions may differ in the context of different specific predictions and in the context of values of other features for those predictions. SHAP is based upon the game-theoretic notion of Shapley values and it considers each feature as a "player" on a "team" of features that works to influence a trained model's prediction. More specifically, a model's baseline output is determined by averaging over all predictions of a given model. Then, each specific prediction is considered as a function of feature influence resulting in some deviation of the model from a baseline prediction. This notion of the "strength" of a positive or negative prediction is then repeatedly tested using different feature "teams" comprising different combinations of features. In doing this, the SHAP approach can empirically determine the influence of each feature for each prediction by comparing how important that feature is to model output both in the presence and absence of combinations of other features. A linear model, such as a logistic regression, takes the following form:

$$\hat{f}(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n,$$

where x in this case is a specific patient's collection of lab values about which we wish to make a prediction, x_n is a specific value within that collection (e.g., CRP), and β_n is a learned weight applied to that lab value. In such a model, the importance of each feature such as CRP, denoted ϕ_{CRP}, can be straightforwardly determined:

$$\phi_{CRP}(\hat{f}) = \beta_{CRP} x_{CRP} - E(\beta_{CRP} X_{CRP}) = \beta_{CRP} x_{CRP} - \beta_{CRP} E(X_{CRP}),$$

where $\beta_{CRP} x_{CRP}$ is the weight of CRP multiplied by the value of CRP for a specific example and $E(\beta_{CRP} X_{CRP})$ is the average of the weight of CRP multiplied by the CRP values for each item in the data set. In simple terms, the importance of a specific feature for a specific prediction outcome in the case of a linear model is the difference between the overall feature weight and the value for that feature for a specific prediction versus the overall feature weight and average value for all predictions. Note that this also does not need to consider other features.

By contrast, a nonlinear model such as our SVM can achieve better performance through capturing nonlinear interactions between features and outcomes at the expense of yielding itself to as simple of an interpretation technique. The Shapley value of a feature, by contrast, is its influence as a member of the feature "team", denoted S, weighted and summed over all combinations of feature values:

$$\phi_{CRP}(\text{val}) = \sum_{S \subseteq \{x_1, \ldots, x_n\} \setminus \{x_{CRP}\}} \frac{|S|!(p - |S| - 1)!}{p!} \times (\text{val}(S \cup \{x_{CRP}\}) - \text{val}(S)),$$

where S is a subset of features, n is the total number of features, and x is the collection of feature values used for a given prediction. The advantage of this approach is that it exhaustively assesses each prediction with all possible combinations of features, annotating the actual influence of each feature for each prediction. Other explainability approaches such as local interpretable model-agnostic explanations approximate feature importance by sparsely modeling the impact of each feature upon predictions in the data set and are especially valuable when exhaustive assessment is not feasible, such as a data set of millions of patients. For data sets of the size we have in this study, we can efficiently perform an exhaustive feature attribution technique and gain richer insight into model behavior. The output of SHAP analysis produces a "force plot", which can be viewed either for a single example or for a whole test population. Because the relationship between features and their corresponding SHAP values is nonlinear and influenced by the values of other features, we profiled the feature-wise relationship between feature and SHAP values in aggregate across all test predictions. To further elucidate such relationships, we also plotted the pairwise relationship between each laboratory value for each test prediction, its corresponding SHAP value, and the magnitude of other members of the five selected laboratory parameters. Imputed values were used in SHAP importance analysis and imputation would impact—at least to some degree—resulting Shapley values. The most effective way in which we currently mitigate this impact is by dropping very sparse measurements leaving us with a fairly dense if not entirely complete initial data set (as seen in the missingness plot). Moreover, as the Shapley value analysis considers the importance of a feature as a function of its relative contribution to outcome, it will favor features that have conspicuously high values especially when other features for the same entity have values closer to their class-wise mean. Taking this into account, although Shapley value analysis is not free from imputation bias, imputed features are themselves less likely to be at the upper or lower bounds of feature variation since they are themselves the product of a regression used during imputation. Finally, the situation in which imputed features would be expected to be anomalous would likely also be a situation where other feature values used for that imputation were also anomalous and the co-occurrence of multiple anomalous values for a patient example would still result in Shapley values which are distributed more evenly among them (since they are all members of the same feature set for a given record) rather than assigning undue importance only to the imputed feature.

Results

We identified 398 patients (43 expired and 355 non-expired) that met criteria for inclusion. Our initial trained
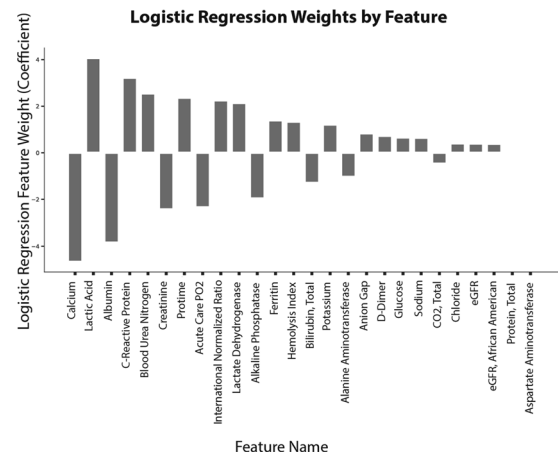

Fig. 1 Learned regression coefficients for each of 26 laboratory values provided to a logistic regression model. Features are each present on the x-axis with their corresponding regression coefficient on the y-axis. Features are ranked according to the absolute value of their corresponding coefficient. The last-ranked features protein, total and aspartate aminotransferase have regression coefficients of zero.

model using all 26 laboratory parameters resulted in a model with 80% sensitivity and 77% specificity for identifying patients who would expire when scored against a held-out test set. Logistic regression permitted the evaluation of feature importance on the model (Fig. 1). The five laboratory values to which the model assigned the highest weights were then selected: CRP, blood urea nitrogen (BUN), serum calcium, serum albumin, and lactic acid. The linear model was then retrained, achieving 90% sensitivity and 77% specificity for expiration status.

Our SVM with class weighting achieved 91% sensitivity and 91% specificity with an AUC of 0.93 and AUPRC of 0.76 (Fig. 2). Additionally, the confusion matrix depicts the individual tallies for our true patient labels (alive and expired) and predicted labels (alive and expired) based upon which the calculated negative predictive value is 98.4%, and positive predictive value is 62.5% for predicting risk of mortality in SARS-CoV-2 positive patients at least 48 h prior to death.

Shapley values were calculated for each feature and each test prediction as a way to profile the relative influence of these laboratory values in model prediction. Figure 3 depicts a force diagram for the model's highest-confidence correct prediction of mortality, the model's single false-negative prediction of survival, and an additive force diagram for the model's mortality prediction across the entire test set. Each diagram represents the influence of each laboratory feature on a single example with a true-positive and false-negative prediction shown. For the single true-positive prediction shown, CRP was most influential in making a prediction of death followed by albumin, which was not decreased and thus limited the model's confidence. Alternatively, for the single false-negative prediction, while

Fig. 2 Receiver operating characteristic (ROC) curve depicting the performance of a trained support vector machine classifier using the top 5 highest-weighted laboratory values. Normalized confusion matrix depicting the support vector machine's prediction of patient expiration versus a patient's true expired status.
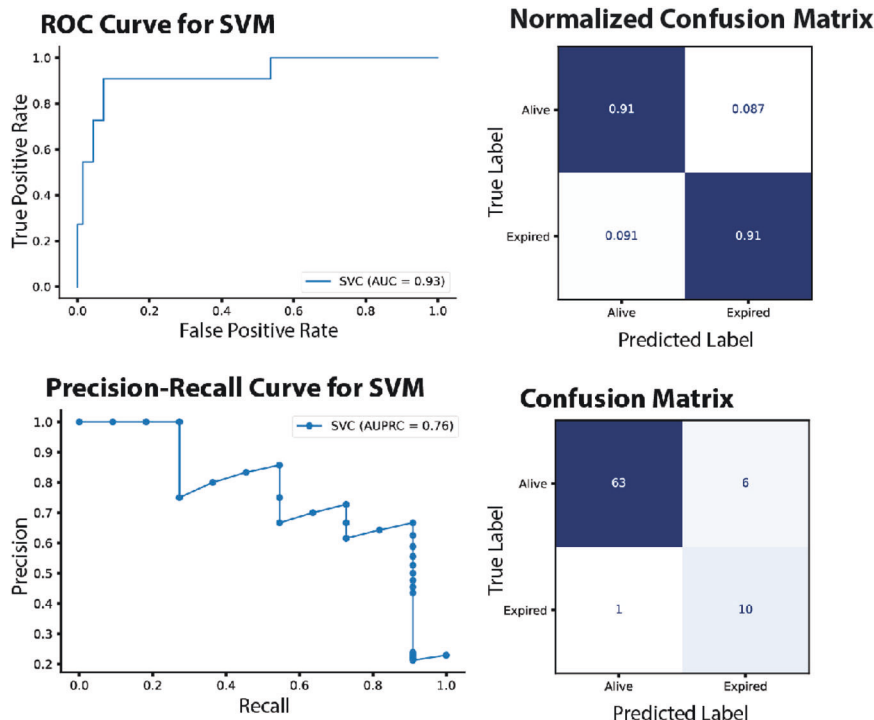
decreased albumin, decreased calcium, and elevated lactic acid contributed toward a prediction for death, the lack of an elevated CRP exerts a strong negative influence on this prediction and induces the model to predict that the patient will not expire. Figure 3 also depicts a comprehensive assessment of these force diagrams across the entire data set wherein each force plot is one slice of the aggregate figure. The x-axis of the aggregate figure denotes which particular slice is either a false-positive, true-positive, false-negative, or false-positive model prediction. As shown in Fig. 3, CRP and calcium are most influential for most predictions but not for all predictions.

CRP, lactic acid, and serum calcium contribute the most to mortality prediction and have the strongest influence on model output when considered over the entire data set as illustrated in Fig. 4. These three features having the highest and lowest SHAP values. As shown in the force diagrams in Fig. 3, certain predictions preferentially use serum BUN and albumin. The relationship between the magnitude of BUN and its corresponding influence on model prediction (i.e., SHAP value) is most pronounced in circumstances where CRP and lactic acid are not elevated and calcium is not decreased. In other words, BUN becomes more relevant in a subpopulation of patients in whom CRP, lactic acid, and calcium are not significantly aberrant. This may indicate that there are two separate phenotypes that can contribute to predicting mortality: one driven by CRP, lactic acid, and calcium, and the other by albumin and BUN. Equally intriguing are the pronounced nonlinear relationships between CRP and calcium and between albumin and lactic

acid both of which are highlighted in Fig. 4. In the case of CRP and calcium, the interaction plot shows that CRP has a strong influence on model outcome only when calcium is elevated. Alternatively, in the case of albumin and lactic acid, albumin has a strong influence on model output primarily when lactic acid is elevated. When CRP and calcium contribute both strong positive and negative influence on prediction, albumin is typically only influential to prediction when decreased. Furthermore, for albumin and CRP there is a nonlinear relationship between feature and Shapley values. Increasingly aberrant values of CRP exert more influence on model predictions except for some samples with high BUN, for which CRP's influence on prediction is less. For albumin, there is a collection of samples over which decreased albumin is increasingly influential for a negative model prediction except for samples where lactic acid is decreased, in which cases changes in albumin have little effect on prediction.

Discussion

The rapid spread of the novel SARS-CoV-2 virus and our limited knowledge of its clinical course has caused a unique shift in our healthcare system. What is usually a slow methodical process to develop and validate clinical tools to aid in the management of diseases has become a rapid frenzy of discovery. Much progress has been made in diagnostic testing, identifying many patients infected with the virus such that there is now a wealth of laboratory data
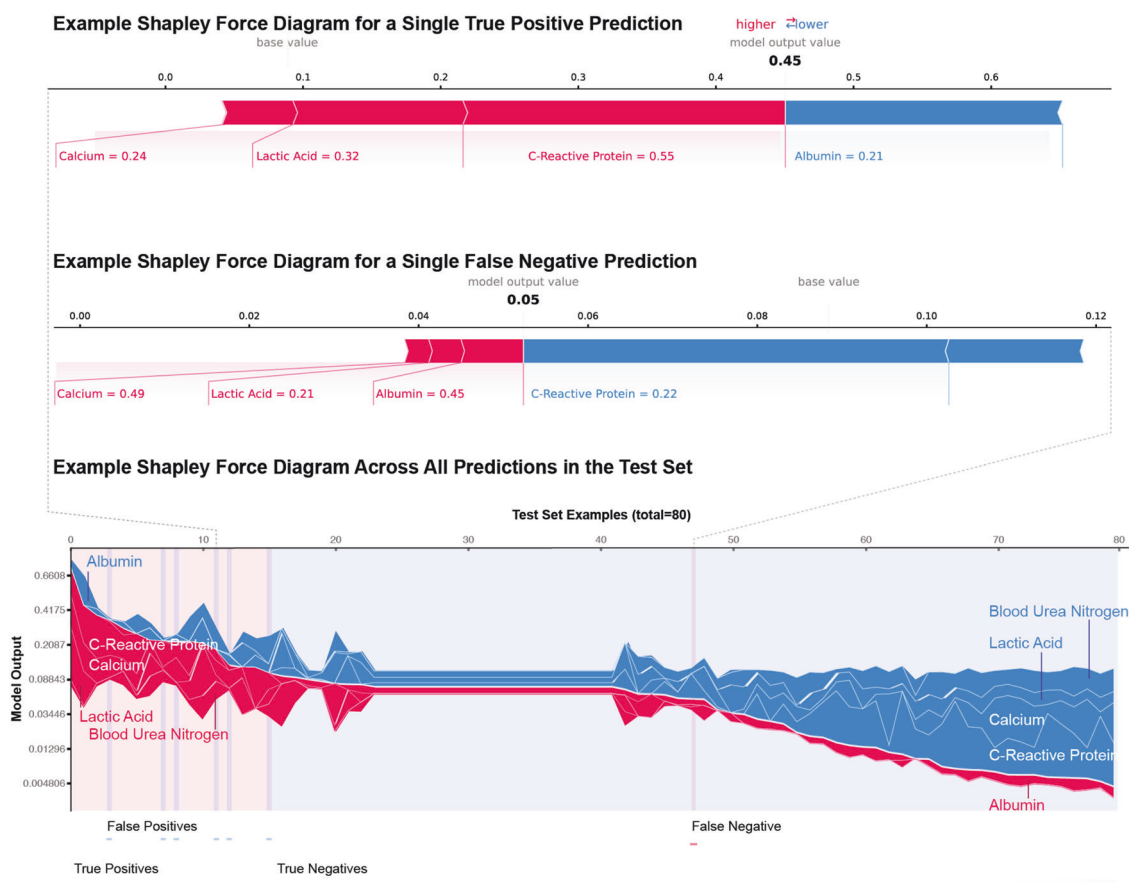
Fig. 3 Feature importance for each of five core laboratory parameters for our trained SVM model. As the influence of laboratory parameters will be different in the setting of different patients, this additive force diagram depicts the changing influence of these parameters throughout the test set. Large blue and pink background areas represent true-positive and true-negative predictions with false-positive and false-negative predictions overlaid.

capturing both SARS-CoV-2 status and concomitant laboratory values. Moreover, this repository continues to grow as hospitals manage rapidly increasing numbers of SARS-CoV-2 cases. Currently, there are multiple studies which report one or two laboratory values associated with prognosis in COVID-19 patients. However, the complex interactions among numerous laboratory values are difficult to illustrate and, moreover, are not well described. To our knowledge, this is the first study using ML to predict mortality in SARS-CoV2 positive patients, which relies exclusively on a multiplex of serum biomarkers.

ML prediction models for COVID-19 have been rapidly entering the medical literature, however, most rely either partially or entirely on subjective clinical data which may vary heavily between observers [16] and institutions. These tools rely either wholly or in part on symptomatic findings and have AUCs ranging from 0.83 to 0.91. This is generally consistent with our findings but many of these models do not include a more detailed assessment of model behavior and sensitivity and specificity for predicting death in the setting of COVID-19 [17, 18]. Interestingly, established prognostic models such as APACHE II, SOFA, and

CURB65 scores have been shown to be useful in predicting death in the setting of COVID-19 with AUCs of 0.84–0.96. Among these, APACHE II demonstrated the highest performance with an AUC of 0.96 and a sensitivity and specificity of 96% and 86%, respectively, although the study demonstrating this exclusively considered ICU patients [19]. Additionally, by utilizing subjective data, such models are likely only to be generalizable within institutions that share the same definitions and approaches to recording the specified data. Pathology is uniquely situated to offer quantitative insight into the progression of COVID-related illness as laboratories can access and process large amounts of objective laboratory data, which are directly uploaded to the LIS. Our ML algorithm took advantage of this structure and obtained all data directly from the LIS without the need for manual entry of any individual values. This not only saves time and resources but ensures that new data can be easily accessed to further train and test this model.

Early identification of at-risk patients has several foreseeable benefits including improved allocation of critical care supplies and staff, ability to contact individuals not currently admitted to the hospital and ensure they are
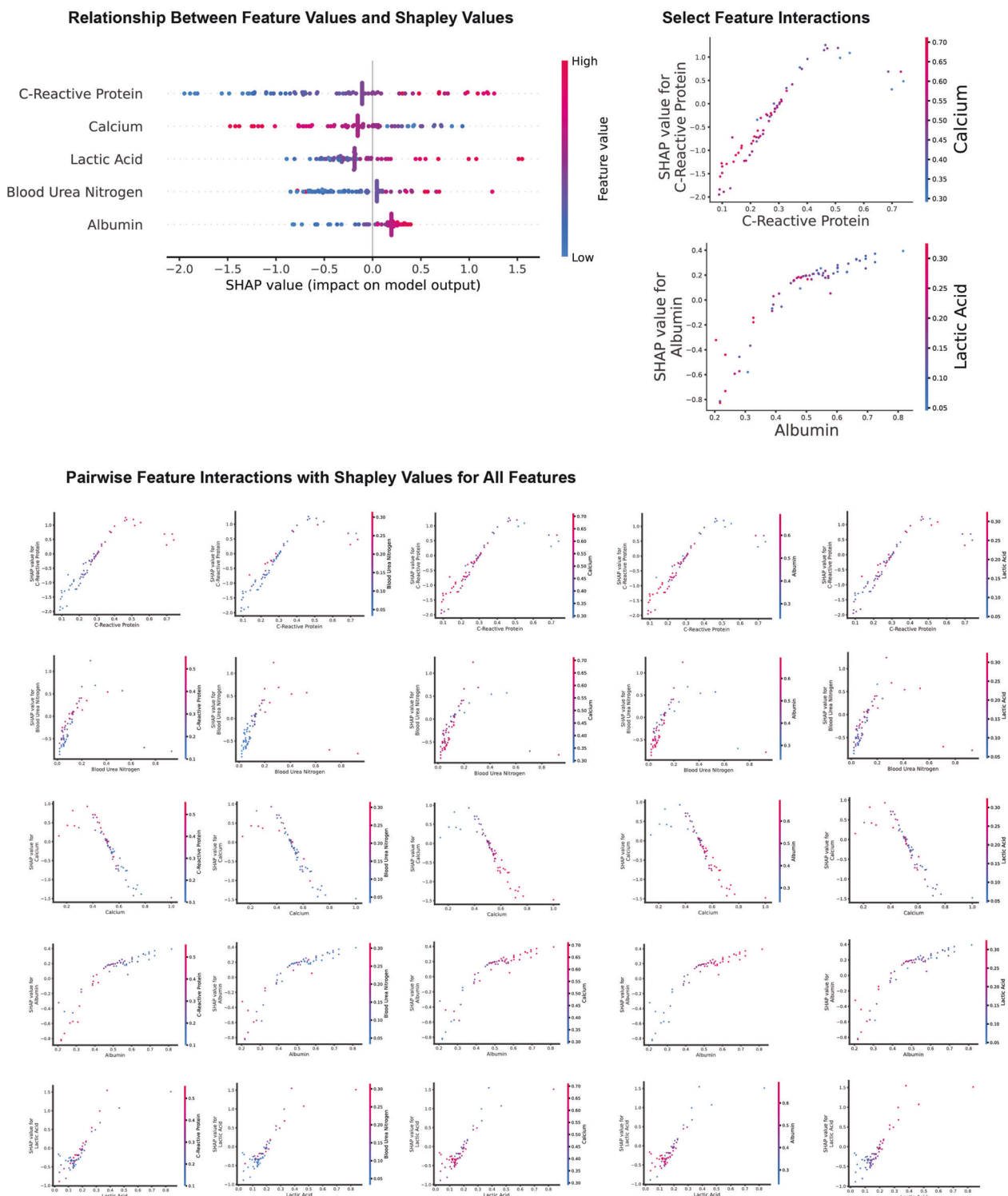
Fig. 4 Relationships between feature and Shapley values. The relationship between the value (color) and Shapley value (x-axis) is plotted for each of five laboratory parameters used to train the SVM model.

evaluated thoroughly, and the ability to relocate hospitalized patients to a center capable of delivering a higher level of care. One ML platform currently being used in over 100 US hospitals to predict mortality and guide hospital

practices is the Epic Deterioration Index (EDI). EDI is a proprietary ML algorithm that assesses multiple clinical and laboratory factors to estimate a hospitalized patient's risk of progressing to severe disease. A recent study through the

University of Michigan evaluated the accuracy and impact of the EDI on an SARS-CoV-2 positive population at their institution showing the EDI to have a relatively high positive predictive value of 80% for patients labeled as high risk, however, the sensitivity was only 39% [20].

Our algorithm aimed to first train for high sensitivity and then optimize for the highest specificity possible with minimal sacrifice to the sensitivity. We initially allowed a simple linear model to work with all 26 laboratory parameters, but sought to simplify the model via feature exclusion, resulting in a parsimonious set of five laboratory parameters: CRP, BUN, serum calcium, serum albumin, and lactic acid. These laboratory parameters comprise an intuitive and clinically relevant subset of commonly tested parameters with which clinicians are already comfortable. These markers can be obtained quickly from most clinical chemistry laboratories with specimens collected using a serum separator or red-top-tube and a heparinized syringe. Lastly, the model's weighting of these five laboratory values unifies several separate observations that appear elsewhere in the literature with regard to severe COVID-19 disease progression or death and highlights potential nonlinear interactions among these parameters that generate additional pathophysiologic hypotheses [2, 3, 21]. This approach is, of course, not without limitations. In using an SVM, we deliberately sought a model that was not confined to axis-parallel decision boundaries as we suspected and intended to capture interactions between features. In comparison to a decision tree, for example, this comes with a trade off in terms of clinical interpretability. Nonlinear models such as this are powerful but they are more limited in their direct application by clinicians in the care of patients. It is our hope that future work in the field of model explainability and decision support will close this gap. Second, SVMs do not easily lend themselves to a probabilistic interpretation of the form "this patient has an x percent chance of death" which is often a useful interpretive paradigm although there are approaches that can allow for such an interpretation [22] in the standard case of hinge loss. Lastly, this model is developed using the data of one medical center, therefore embedding within that data the ordering practices of one medical center as well. Even if the overall distribution of laboratory values for patients is generalizable, it may be that differing protocols result in more sparse features thus requiring more imputation.

CRP and lactic acid are used regularly as serum markers of inflammation. High levels of CRP and lactic acid have been identified as strong predictors of COVID-19 disease severity and elevated CRP has been positively correlated with the size of lung lesions by computed tomography [2, 3, 21]. Interleukin-6, an inflammatory cytokine that induces the production of both CRP and lactic acid, has been shown to be present in high quantities in COVID-19 patients [23]. Furthermore, studies have demonstrated lactic acid elevation in sepsis and circulatory shock, whereas albumin is well known to act as a negative acute inflammatory reactant [24]. Our algorithm placed a strong weight on hypoalbuminemia as a negative model predictor except in the setting of hyperlactatemia. Severe sepsis causes a high anion-gap metabolic acidosis due hyperlactatemia, while hypoalbuminemia is known to lower an anion gap [25]. In such a severe inflammatory state resulting in such high levels of lactic acid, albumin becomes less contributory to predicting mortality.

Multiple studies have identified kidney injury as a sequela frequently present in COVID-19 patients with severe disease, many of whom expired [26–28]. In addition, nephropathies associated with other viruses such as HIV, CMV, and HCV are well known [29, 30]. A review of kidney renal histology in 26 COVID-19 autopsies demonstrated diffuse proximal tubular injury and electron microscopic examination revealed coronavirus-like particles with spikes in the tubular epithelium and podocytes. Positive immunohistochemical staining for the SARS-CoV-2 nucleoprotein has been observed in tubular epithelial cells, which also demonstrated the upregulation of angiotensin converting enzyme 2 (ACE2) [26] that serves as the receptor binding domain of SARS-CoV-2 spike protein facilitating entry into the cell [31]. Of the five biomarkers, our algorithm weighted BUN, albumin, and calcium highly, which are all associated with acute kidney injury (AKI) [32]. Low serum albumin has been reported as a common finding in non-survivors as demonstrated by Huang et al. [33] whereby 25 of 36 had low albumin. Furthermore, AKI is a well-documented cause of hypoalbuminemia, which can exacerbate hypocalcemia [29, 34]. In a multicenter, retrospective study, Li et. al., reported that patients with severe disease including death had an elevated BUN that was statistically significant compared with survivors and patients with non-severe disease. Authors further demonstrated a mortality risk in COVID-19 patients with AKI with an estimated hazard ratio 5.3 times those without AKI [27]. In light of severe disease associated with kidney injury in COVID-19 patients, it is unsurprising to see a pattern of hypoalbuminemia, hypocalcemia, and azotemia identified by our algorithm.

In patients with COVID-19, cardiac disease is a well-documented risk factor for increased risk of death [28, 35, 36]. Autopsy findings have detected the virus in heart tissue by RT-PCR and myocarditis in association with elevated cardiac biomarkers [37]. Like tubular epithelial cells, ACE2 is expressed on myocytes and vascular endothelial cells [26, 31]. Additionally, cardiac arrhythmias have been documented in COVID-19 patients [37]. Myocarditis and hypocalcemia are both independently associated with arrhythmias [24] offering some explanation of the

mechanisms whereby serum calcium is a useful predictive variable and increased risk of death in patients with existing cardiac disease. This is consistent with our model's high weight on low calcium levels to predict mortality. More interestingly, calcium has been hypothesized to support an immune response to COVID-19 based upon reports that calcium channel blockers are associated with reduced mortality in COVID-19 [38]. Although we do not provide a mechanism as to why, our work supports hypocalcemia as a useful feature in predicting mortality in COVID-19.

Our study is not without limitations. These data represent a single institution (University of Texas Medical Branch) and are subject to institutional biases, which may be present in clinical test selection and implementation. More relevant to this model itself, the sample size is unbalanced with a relative minority of SARS-CoV-2 positive mortalities although this is likely to be a limitation in any setting. Given the recency of SARS-CoV-2, such a paucity of mortalities with existing laboratory data is unavoidable without larger multi-institutional data sets.

Finally, this study does not perform subset analysis among other clinical factors such as patient diagnoses. This is partly to preserve parsimony in the model itself as a simple, general tool for predicting risk of mortality. However, it may always be the case that ongoing case accrual will power larger studies that may discover other covariates outside of laboratory testing data. Lastly, as this model is aiming to predict the outcomes of timeseries events, many of which are still ongoing, it is always possible that any patients who currently are alive may become case fatalities at any point. For these reasons, we emphasize that the utility of such a model is likely more in its ability to integrate disparately cited laboratory values as interactive predictors of mortality and to guide discussion on why such features end up being relevant.

An important strength of this study is that we include all SARS-CoV-2 positive patients who are deceased even if those patients did not die while admitted to the hospital. As such patients are present in model training and testing subgroups, this allows our predictive model to portend risk of death for both in-hospital and post-discharge mortality, further enhancing its applicability and its value in sensitively flagging patients at risk for such adverse outcome, especially as COVID-19 presently lacks well-defined management guidelines. Furthermore, this ML approach is easy to deploy, train, and retrain, meaning that, as more data become available, this algorithm will improve with regard to predictive performance. Furthermore, we hope to follow development of such an algorithm with additional studies that quantify its potential benefit to patient survival by prospectively following algorithmically flagged patients.

ML algorithms can simultaneously evaluate the cumulative effects of numerous biomarkers to discover high-order interactions. This intricacy of data interpretation demonstrates the powerful opportunity of using ML in clinical pathology.

Compliance with ethical standards

References

1. Johns Hopkins University & Medicine. Johns Hopkins Coronavirus Resource Center. 2020. https://coronavirus.jhu.edu/.
2. Wang L. C-reactive protein levels in the early stage of COVID-19. Med Mal Infect. 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7146693/.
3. Tan C, Huang Y, Shi F, Tan K, Ma Q, Chen Y, et al. C-reactive protein correlates with computed tomographic findings and predicts severe COVID-19 early. J Med Virol. 2020;92:856–62.
4. Shoenfeld Y. Corona (COVID-19) time musings: our involvement in COVID-19 pathogenesis, diagnosis, treatment and vaccine planning. Autoimmun Rev. 2020;19:102538.
5. Zhou B, She J, Wang Y, Ma X. Utility of ferritin, procalcitonin, and c-reactive protein in severe patients with 2019 novel coronavirus disease. 2020. https://www.researchsquare.com/article/rs-18079/v1.
6. Luo Y, Szolovits P, Dighe AS, Baron JM. Using machine learning to predict laboratory test results. Am J Clin Pathol. 2016;145:778–88.
7. Huang L, Wu T. Novel neural network application for bacterial colony classification. Theor Biol Med Model. 2018;15:22.
8. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25:1301–9.
9. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol. 2020;21:222–32.
10. Korbar B, Olofson AM, Miraflor AP, Nicka CM, Suriawinata MA, Torresani L, et al. Deep learning for classification of colorectal polyps on whole-slide images. J Pathol Inf. 2017;8:30.
11. Ko J, Baldassano SN, Loh P-L, Kording K, Litt B, Issadore D. Machine learning to detect signatures of disease in liquid biopsies—a user's guide. Lab Chip. 2018;18:395–405.
12. McCaffrey P. An introduction to healthcare informatics: building data-driven tools. Academic Press, Cambridge, MA; 2020.
13. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009;60:549–76.
14. Enders CK. Applied missing data analysis. New York: Guilford Press; 2010.
15. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Advances in neural information processing

systems. 2017. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

16. Wynants L, Van Calster B, Bonten MMJ, Collins GS, Debray TPA, De Vos M, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ. 2020;369:m1328.

17. Zhao Z, Chen A, Hou W, Graham JM, Li H, Richman PS, et al. Prediction model and risk scores of ICU admission and mortality in COVID-19. PLoS ONE. 2020;15:e0236618.

18. Yadaw A, Li Y, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. Lancet Digit Health. 2020;2:e516–25.

19. Zou X, Li S, Fang M, Hu M, Bian Y, Ling J, et al. Acute physiology and chronic health evaluation II score as a predictor of hospital mortality in patients of coronavirus disease 2019. Crit Care Med. 2020;48:e657–65.

20. Singh K, Valley TS, Tang S, Li BY, Kamran F, Sjoding MW, et al. Validating a Widely Implemented Deterioration Index model among hospitalized COVID-19 patients. Health Inform. 2020. http://medrxiv.org/lookup/doi/10.1101/2020.04.24.20079012.

21. Peng YD, Meng K, Guan HQ, Leng L, Zhu RR, Wang BY, et al. Clinical characteristics and outcomes of 112 cardiovascular disease patients infected by 2019-nCoV. Zhonghua Xin Xue Guan Bing Za Zhi. 2020;48:E004.

22. Grandvalet Y, Mariethoz J, Bengio S. A probabilistic interpretation of SVMs with an application to unbalanced classification. NIPS, Cambridge, MA; 2006.

23. Chen G, Wu D, Guo W, Cao Y, Huang D, Wang H, et al. Clinical and immunological features of severe and moderate coronavirus disease 2019. J Clin Invest. 2020;130:2620–9.

24. Lichtenauer M, Wernly B, Ohnewein B, Franz M, Kabisch B, Muessig J, et al. The lactate/albumin ratio: a valuable tool for risk stratification in septic patients admitted to ICU. Int J Mol Sci. 2017;18:1893.

25. Colombo J. A commentary on albumin in acidosis. Int J Crit Illn Inj Sci. 2017;7:12–3.

26. Su H, Yang M, Wan C, Yi LX, Tang F, Zhu HY, et al. Renal histopathological analysis of 26 postmortem findings of patients with COVID-19 in China. Kidney Int. 2020. https://www.kidney-international.org/article/S0085-2538(20)30369-0/fulltext.

27. Li Z, Wu M, Yao J, Guo J, Liao X, Song S, et al. Caution on kidney dysfunctions of COVID-19 patients. 2020. https://www.medrxiv.org/content/10.1101/2020.02.08.20021212v2.

28. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet. 2020;395:1054–62.

29. Shankland SJ. The podocyte's response to injury: role in proteinuria and glomerulosclerosis. Kidney Int. 2006;69:2131–47.

30. Prasad N, Patel MR. Infection-induced kidney diseases. Front Med. 2018;5:327.

31. Pillay TS. Gene of the month: the 2019-nCoV/SARS-CoV-2 novel coronavirus spike protein. J Clin Pathol. 2020. https://doi.org/10.1136/jclinpath-2020-206658.

32. Kumar V, Abbas AK, Aster JC. Robbins and Cotran pathologic basis of disease. Ninth edition. Philadelphia, PA: Elsevier/Saunders; 2015.

33. Huang Y, Yang R, Xu Y, Gong P. Clinical characteristics of 36 non-survivors with COVID-19 in Wuhan, China. 2020. https://www.medrxiv.org/content/10.1101/2020.02.27.20029009v2.

34. Wiedermann CJ, Wiedermann W, Joannidis M. Causal relationship between hypoalbuminemia and acute kidney injury. World J Nephrol. 2017;6:176–87.

35. Chen L, Li X, Chen M, Feng Y, Xiong C. The ACE2 expression in human heart indicates new potential mechanism of heart injury among patients infected with SARS-CoV-2. Cardiovasc Res. 2020;116:1097–100.

36. Tian S, Xiong Y, Liu H, Niu L, Guo J, Liao M, et al. Pathological study of the 2019 novel coronavirus disease (COVID-19) through postmortem core biopsies. Mod Pathol. 2020. https://doi.org/10.1038/s41379-020-0536-x.

37. Driggin E, Madhavan MV, Bikdeli B, Chuich T, Laracy J, Biondi-Zoccai G, et al. Cardiovascular considerations for patients, health care workers, and health systems during the COVID-19 pandemic. J Am Coll Cardiol. 2020;75:2352–71.

38. Zhang L, Sun Y, Zeng HL, Peng Y, Jiang X, Shang WJ, et al. Calcium channel blocker amlodipine besylate is associated with reduced case fatality rate of COVID-19 patients with hypertension. 2020. https://www.medrxiv.org/content/10.1101/2020.04.08.20047134v1.