Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Using Machine Learning to Unveil Demographic and Clinical Features of COVID-19 Symptomatic and Asymptomatic Patients

Sarah Al Youha
Ministry of Health
Mohammad Alkhamis
Kuwait University https://orcid.org/0000-0001-6451-5247
Sulaiman Al Mazeedi
Ministry of Health
Mohannad Al-Haddad
Ministry of Health
Mohammad H. Jamal
Ministry of Health
Salman K. Al-Sabah (salman.k.alsabah@gmail.com)
Kuwait University

Research article

Keywords: SARS-CoV-2, machine learning, symptomatic, asymptomatic, interactions.

Posted Date: August 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-52330/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Background: Demographic and clinical features of COVID-19 patients are critical components in shaping their symptomatic status. However, the relationship between patients' symptomatic status and their features are typically complicated and nonlinear.

Methods: We explored important features that drive the symptomatic status of COVID-19 patients and reveal their interactions with other relevant factors. We used an extensive multi-algorithm machine learning (ML) pipeline and 68 demographic and clinical features to fit a predictive model to 3,995 patients in the State of Kuwait between February and June 2020. Our ML pipeline comprised five algorithms, including logistic regression (LR), random forest (RF), support vector machine (SVM), gradient boosting (GBM), and extreme gradient boosting (XGM).

Results: SVM outperformed all algorithms (AUC = 0.77 and accuracy = 70.01%), while logistic regression had the lowest predictive power (AUC = 0.65 and accuracy = 66.14%). Our ML model identified C-reactive, respiratory rate, transmission dynamics, and other demographics as the most important predictors of COVID-19 symptomatic patients. While, only demographic features were important predictors for asymptomatic patients. However, our ML model further revealed that the non-linear relationships between impaired renal function, other clinical biomarkers and demographic features were critical in shaping the risk of being symptomatic patient.

Conclusions: We demonstrated remarkable predictive performance of our ML model over traditional statistical methods in identifying important clinical and demographic features of symptomatic vs. asymptomatic. Further application of our ML pipeline in the COVID-19 case definition and guiding pharmaceutical and none-pharmaceutical interventions will help reduce the public health and economic implications of this devastating virus on local and global scales.

Highlights

- C-reactive protein and community type were the top two important predictors for symptomatic patients.

- Only demographic features were important predictors for asymptomatic patients.

- There are strong non-linear relationships between impaired renal function and other features in symptomatic patient.

- Machine learning outperformed classical logistic regression in predicting symptom status.

- Machine learning can help improve case definition and public health surveillance.

Background

The rapid emergence and spread of the novel severe acute respiratory syndrome 2 (SARS-CoV-2) led to the catastrophic pandemic of coronavirus disease 2019 (COVID-19). The complexity of COVID-19 infections and their associated clinical course, are shaped by the substantial heterogeneity in the characteristics, dynamics, and comorbidities of the affected populations [1]. Therefore, untangling the demographical and clinical features of COVID-19 infected individuals is critical for improving currently implemented intervention strategies. Machine learning (ML) algorithms have previously been used to further our understanding of infectious diseases and have recently become more interpretable, due to newly introduced tools [2]. These tools have enabled us to gain deeper mechanistic insights into the drivers of different clinical outcomes [3].

The spectrum of COVID-19 clinical presentations, includes asymptomatic, presymptomatic and symptomatic cases [4]. Recent studies have estimated that asymptomatic COVID-19 carriers account for 40–45% of COVID-19 cases [5]. These individuals have been hypothesized to be unknowing 'super spreaders' of the virus, as they do not exhibit any overt physical signs of being infected [6]. Although several authors have explored the clinical characteristics of asymptomatic COVID-19 carriers, these studies are limited by small patient numbers [7]. As this population does not exhibit outward physical signs of the disease, it is difficult to study them. Consequently, most of the literature related to this important group of patients has not been addressed at a large scale. Also, asymptomatic patients may also become symptomatic after a few days [8]. Of those patients that become symptomatic, the majority have a milder clinical presentation of COVID-19 [9], while approximately 18–33% may require hospitalization and admission into the intensive care unit [10, 11]. The reasons for these widely differing disease manifestations are currently not well understood.

As most current studies, comparing the clinical characteristics of symptomatic and asymptomatic patients, have relied on small samples and traditional statistical methods, their findings may not be representative of a larger, more general, hospitalized population [12, 13]. Further, traditional statistical methods have many underlying assumptions, such as linearity of the relationship between the outcomes and its predictors, which neglect the complexity that underpins COVID-19 infections. Statistical linear models tend to underperform with large datasets and are extremely susceptible to overfitting due to the large correlations between variables [14, 15]. Conversely, ML algorithms can explore and accommodate large datasets with thousands of variables of different varieties and require minimal statistical assumptions [3]. Additonally, ML algorithms have proven to be superior in making individual-level predictions and are able to interrogate multiple interactions between risk factors, rendering them powerful tools to unveil novel insights into the clinical epidemiology of COVID-19 [15, 16].

To the knowledge of the authors, ML algorithms have not yet been widely adopted in exploring the characteristics of COVID-19 infected populations in a clinical setting. Thus, we applied an interpretable multi-algorithms ML ensemble pipeline to a dataset consisting of approximately 4000 COVID-19 patients detected in the State of Kuwait. We explored demographic and clinical factors that were shaping the risk of being asymptomatic or symptomatic. Furthermore, we studied complex interactions between

important risk factors that likely contributed to symptom status in infected patients and assessed our ML model in the context of individual-level clinical diagnosis, to address their advantage and limitations.

Methods

Setting and Data source

Due to unique circumstances, Kuwait is ideally suited to further our understanding of the varying spectrum of COVID-19 patients' clinical presentations. In Kuwait, before the first case of COVID-19 was even reported in the country, repatriation efforts were instigated, and screening for SARS-CoV-2 was performed for all incoming passengers. Any travelers who tested positive for the SARS-CoV-2 virus, irrespective of symptomology, were admitted to Jaber Al-Ahmad Al-Sabah Hospital. In addition, during those early stages of the pandemic, all cases of COVID-19 that were diagnosed in Kuwait, including asymptomatic carriers, were admitted as inpatients at this same hospital. As a result, Jaber Hospital eventually became a hospital that is solely dedicated to treating COVID-19 patients. This presented an opportunity to study the clinical features of COVID-19 patients, including those who were asymptomatic, due to the large proportion of patients that were admitted to Jaber Hospital, after being diagnosed with COVID-19, secondary to the mass random screening and traveler screening efforts. Additionally, asymptomatic and presymptomatic patients all received identical investigations and work-up, to their symptomatic counterparts, and all within the same facility, contributing to the homogeneity of this patient population.

Data comprised 3 995 admitted patients between the 24th of February and the 27th of May 2020, and was an extension of a population described earlier elsewhere [17]. All patients, included in the study, tested positive for the SARS-CoV-2 virus using real-time polymerase chain reaction (RT-PCR). We extracted their demographic and clinical variables (*i.e.,* features) upon admission from the hospital's electronic medical record system (Supplementary Table 1). Patients were classified as being symptomatic or asymptomatic based on their initial hospital presentations. Those who were classified as initially asymptomatic on admission were re-classified as symptomatic if their medical records indicated they developed symptoms later during their clinical course. Thus, the final dataset comprised SARS-CoV-2 symptom status as an outcome and 63 features (Additional file 1). We used the R software environment to conduct all statistical analyses.

Data pre-processing

We used an extensive ML ensemble pipeline [18], which is comprised of five supervised algorithms, including logistic regression (LR), support vector machine (SVM), random forest (RF), gradient boosting (GBM), and extreme gradient boosting (XGB) to build a predictive model for the SARS-CoV-2 symptom status. We used multiple ML algorithms and compared the differences in their predictive performances. Some ML algorithms are less flexible to missing data (*i.e.,* SVM) [18], and therefore we imputed missing data using the RF-based nonparametric routine implemented in 'missForest' R package [19]. Also, because some of our selected algorithms (*i.e.,* LR) cannot accommodate strongly collinear features, we

removed features with the largest mean absolute correlation (> 0.7) [20]. Additionally, we applied the 'Boruta' R package, which contains an RF-based algorithm to only include those that are relevant for prediction [21]. We used the Boruta procedure to increase the efficiency, predictive performance, and accuracy of our ML algorithms [22]. We controlled for the class imbalance between asymptomatic and symptomatic (Supplementary Table 1), using a down-sampling routine, to avoid biased predictions toward the majority outcome class (*i.e.,* symptomatic) [23]. We then partitioned the data randomly into training (80%) and testing sets.

Model training and evaluation

We trained our ML algorithms using the feature set determined by the Boruta procedure. We ran LR, SVM, GBM, and XGB ML algorithms using the 'Caret' R package [24], while the 'random Forest' R package [25] was used to ran the RF algorithm. We used a K-fold cross-validation approach (K = 10) to estimate model performance parameters for each algorithm, including area under the curve through a receiver operator characteristic (ROC), accuracy (Acc), sensitivity (Se), specificity (Sp). We calculated these parameters using the average confusion matrix across all folds of the cross-validation and used the default grid parameter settings in the training process of all algorithms. The K-fold approach was selected, because it can prevent overfitting and artificial inflation of Acc, due to the use of the same data for training and validation [18]. We then selected the best performing algorithm to predict the probability ( ) of COVID-19 symptom status by comparing the estimated validation parameters of each model using the testing dataset.

Model Interpretation

We visualized and examined the best performing model results using the 'iml' R package [26]. Breiman's permutation routine [27] was used to estimate feature importance, which quantifies the expected loss in predictive performance (*i.e.,* how correctly the model classifies symptomatic vs. asymptomatic) for a pair of patients compared to the full model when a specific feature has been switched [27]. Therefore, if the permutation routine does not affect the selected model's predictive performance, the feature is then classified as unimportant. We estimated the global and individual effect of the top six important features using partial dependence (PD) and centered individual condition expectation (cICE) methods, respectively [28]. We used Friedman's H-statistic method [29] to quantify the overall interaction strength of each features. We then reused the same method to quantify the top features' interaction strength (*i.e.,* feature with the highest overall interaction strength) with the other features. This method uses a partial dependency decomposition routine to quantify the portion of the variance explained by each interaction [29]. Finally, we computed Shapley values ( ), following a game theory approach, to quantify individual-level predictions and the contribution of each feature to those predictions for two (symptomatic and asymptomatic) randomly selected patients [30].

Results

Our results indicate that 65% of the infected patients were symptomatic, and 48 of the 63 selected features were significantly associated with the study outcome (Additional file 1). However, the Boruta procedure indicated that 41 features were important for predicting the asymptomatic status of infected patients. Results of the ML pipeline suggest that the SVM algorithm outperformed the other algorithms, across all validation parameters (*i.e.,* AUC, Acc, Sp and Se; Table 1) and was able to correctly classify symptom status of the infected patients 77% of the time (AUC = 0.77; Table 1). In contrast, the LR algorithm was the worst performing algorithm, and was only able to correctly predict symptom status 65% of the time (AUC = 0.65; Table 1).

Table 1. Cross-validation summary results for the machine learning algorithms. Best performing algorithm is boldfaced.

Model	AUC*	Accuracy (%) ± SE*	Specificity (%) ± SE	Sensitivity (%) ± SE
LR*	0.65	66.14 ± 1.07	75.55 ± 0.88	64.90 ± 1.24
RF*	0.73	67.02 ± 0.45	73.69 ± 0.49	68.55 ± 0.65
SVM*	**0.77**	**70.01 ± 0.45**	**79.82 ± 0.73**	**73.76 ± 0.88**
GBM*	0.72	65.32 ± 0.45	74.66 ± 0.73	65.49 ± 0.88
XGB*	0.74	66.95 ± 1.16	72.39 ± 0.35	66.25 ± 1.13

AUC: Area Under the Curve; SE: Standard Error; LR: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine; GBM: Gradient Boosting; XGB: Extreme Gradient Boosting.

Our ML approach revealed that C-reactive protein (CRP), community type, travel history, the presence of diffuse opacification in a chest x-ray (CxR), respiratory rate, and transmission type were the most important features for predicting symptom status (Figure 1). PD plots showed that the risk of being symptomatic increases when the concentration of CRP rises (Figure 2A). There was a slight increase in the risk of being symptomatic for patients belonging to the citizen-resident community (Figure 2B), with no travel history (Figure 3C) and who had presence of diffuse opacification in their CxR (Figure 2D). Our PD plots also show that infected patients with a respiratory rate > 20 breaths/min (Figure 2E) and that were healthcare workers (Figure 2F) were more likely to be symptomatic.

Estimated glomerular filtration rate (eGFR), followed by creatinine levels, had the strongest overall interactions with other features, in shaping the risk of being symptomatic (Figure 3A). Further, the strongest interaction with eGFR was urea, followed by the presence of diffuse opacification in CxR, sex, total protein, alkaline phosphatase (ALP), and sodium (Figure 3B). Patients with low eGFR (< 50 mL/min./1.73^2) and urea (< 20 mmol/L) were most likely to develop symptoms (> 0.75; Figure 4A). The risk of developing symptoms was higher in patients with diffuse opacification and a low eGFR (Figure 4B). Similarly, females with a low eGFR were more likely to develop symptoms than males (Figure 4C).

Patients with high total protein (> 60 g/L), but low ALP (< 200 IU/L) and sodium (< 140 mmol/L) had the highest risk of developing symptoms (Figures 4D-F).

Our Shapely values of all selected features that contributed to the risk of developing symptoms in two randomly selected patients are summarized in Figure 5. A patient with high C-reactive protein (= 99.14 mg/L), low eGFR (= 40 mL/min./1.73^2), low sodium (= 132 mmol/L), and has diffuse opacification on CxR is most likely to develop symptoms (Figure 5A). Conversely, an infected patient with a history of recent travel, with low CRP (= 4 mg/L) and normal white blood count (= 9.1 x 10^9/L) is less likely to develop symptoms (Figure 5B).

Discussion

To our knowledge, this study represents the first attempt to use an integrated and interpretable ML analytical framework on a large series of COVID-19 infected patients, to uncover more in-depth insights, into the risk factors that shaped their symptom status. We identified important demographic and clinical predictors of symptom status and unveiled their complex non-linear relationships. In general, we found that the combination of clinical and demographic features, including targeted inflammatory markers, radiographic findings, laboratory blood tests, and patients' characteristics, were important in predicting the risk of being symptomatic.

Our ML model revealed CRP levels, diffuse opacification in a CXR, and respiratory rate (Fig. 1), were the most important features in predicting which patients infected with COVID-19 will subsequently develop symptoms. These results are consistent with past studies, as CRP levels have been identified as potential biomarkers for symptomatic COVID-19 patients [31]. Moreover, these finding supports the notion, that imaging modalities, such as chest x-rays and computerized topography, are important for diagnosing COVID-19 patients [32].

Community type, travel history, and transmission type were the most important demographic features for predicting the risk of symptomatic status (Fig. 1). Being a citizen-resident of Kuwait is associated with having a higher socioeconomic status compared non-citizen residents. Despite this, our results indicate that citizen-residents were more likely develop symptoms (Fig. 2B). This finding is surprising, especially as Alkhamis *et al.*, inferred that significant spreading and cluster events in migrant workers communities were substantially more severe than residents-citizens due to their densely populated areas and poor living conditions [33]. A potential explanation for this, may be that migrant workers, tend to represent a much younger subset of the population in Kuwait [34]. Our ML model inferred that patients with a recent travel history are less likely to develop symptoms (Fig. 2C). We attributed this finding to the government's extensive intervention measures of testing and forced institutional quarantine of arriving travelers at the beginning of the epidemic in Kuwait [33]. Also, healthcare workers were more likely to develop symptoms despite having access to personal protective equipment during their duties (Fig. 2F). Being in close contact with a large number of COVID-19 patients for a prolonged period of time, whilst performing

various high-risk procedures such as aerosol generated procedures (e.g. intubation) may be contributory risk factors [35].

Past studies have inferred a linear relationships between evidence of kidney injury, on admission, and the clinical course of COVID-19 [36]. Using our ML pipeline, we were able to explore this relationship further, by modeling nonlinear interactions between features (Fig. 3) and found that eGFR has the strongest overall interactions with other variables in shaping the risk of being symptomatic (Fig. 3A). Indeed, COVID-19 patients with low eGFR may be more likely to be severely ill on admission than patients with normal kidney function, as described elsewhere [37]. The mechanistic process that underlies this has been hypothesized to be due to the presence of the angiotensin-converted enzyme-2 (ACE-2) receptor in the kidney, which has been shown to be 100 times greater in the kidney, than in the lung. In addition to its function as a receptor for SARS-CoV-2 entry into the alveolar cells of the lung, the ACE-2 enzyme has also been shown to interact with the virus directly, affecting the renin-angiotensin aldosterone system (RAAS) physiologically. This process might indicate that patients with chronic kidney disease (CKD) may be more susceptible to getting a complicated COVID-19 infection since they have high RAAS activities, resulting in a systemic increased expression of ACE-2, a major entry site for the virus. Feature interaction plots with eGFR show that patients with low urea (Fig. 4A), elevated total protein (Fig. 4D), and hyponatremia (Fig. 3F) are at higher risk of being symptomatic. These results unveil the complexity of the acute disease phase upon admission, in which patients might experience multiple severe inflammatory processes and a negative fluid balance, as a result of impaired renal function [38].

A general limitation of the present study is the population size and potential selection bias toward our study population. That said, our data were collected from the official COVID-19 treatment hospital (*i.e.*, Jaber Al-Ahmad Al-Sabah Hospital), which makes this population representative of the whole state of Kuwait. Furthermore, our ML pipeline is incapable of characterizing the uncertainties in the model predictions well. Methods such as Bayesian additive regression tree (BART) are more robust in quantifying such uncertainties, although they are limited by their requirement for larger datasets and demanding computations [39]. An advantage of the present analytical pipeline is the remarkable applicability of Shapley values to interpreting, at a finer scale, what our model means in terms of classifying symptom status (*i.e.*, why a specific infected individual developed COVID-19 symptom, while the other did not?). For example, for a randomly selected patient from our cohort (Fig. 5A), having high CPR, low eGFR, hyponatremia, and diffuse opacification were associate with that COVID-19 patients becoming symptomatic.

By providing deeper insights into the underlying disease process that dictate patients' clinical course, our ML pipeline can potentially be used to risk-stratify patients. Biomarkers and demographic data can be used as a proxy for disease status, potentially eliminating the need for extensive testing, which has exhausted healthcare resources, particularly for COVID-19 worldwide. Also, ML models can be robust tools for COVID-19 case definitions, and therefore, may help avoid inaccurate mapping of epidemic trajectories through public health surveillance activities [40]. It worth noting that while our ML model identified community type as an important feature (Fig. 1), it was insignificantly associated (p-value =

0.270; Additional file 1) with the study outcome using traditional statistical methods. Thus, commonly used p-values to assess the statistical significance of the association between two variables might not be a reliable measure of inference in population-based studies [41].

Conclusions

In conclusion, we have shown that inflammatory markers, respiratory signs, transmission dynamics, and demographics were essential predictors of symptom status in COVID-19 patients. Nevertheless, our ML model further revealed that the non-linear relationships between impaired renal function and other clinical biomarkers, were critical in shaping the risk of being a symptomatic COVID-19 patient. We demonstrated a superior predictive performance of our ML model, over traditional statistical methods, such as logistic regression. Further application of our ML pipeline in defining COVID-19 cases may help guide public health interventions, which will help reduce the grave implications of this devastating virus globally.

List Of Abbreviations

SARS-CoV-2
acute respiratory syndrome 2; COVID-19:coronavirus disease 2019; ML:Machine learning; AUC:Area Under the Curve; SE:Standard Error; LR:Logistic Regression; RF:Random Forest; SVM:Support Vector Machine; GBM:Gradient Boosting; XGB:Extreme Gradient Boosting; ROC:receiver operator characteristic, Acc:accuracy, Se:sensitivity, Sp:specificity;:predicted probability; PD:partial dependence;:Shapley values; cICE:centered individual condition expectation; CRP:C-reactive protein; CxR:chest x-ray; eGFR:Estimated glomerular filtration rate; ALP:alkaline phosphatase.

Declarations

Ethics approval and consent to participate

Ethical approval and consent to participate for this study was obtained from the Ministry of Health of Kuwait Ethical Review Board.

Consent for Publication

Due to the retrospective nature of this study, the requirement for obtaining informed consent from study subjects was waived by the IRB (Ministry of Health Kuwait)

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

Authors declare no competing interests.

Funding

This project was funded by the Kuwait Foundation for the Advancement of Science, grant number: Corprop-35. The grant that was provided was used to pay data collectors for their time.

Author's contribution

The study was designed by S.Y, M.A.A and S.K.S. The literature review and data were collected and organized by S.M, A.K, M.H and M.H.J. All statistical analyses were conducted by M.A.A. All authors contributed in writing the first draft of the manuscript.

Acknowledgments

The authors would like to acknowledge the National Center for Health Information in Kuwait for assistance with data collection.

References

1. Dowd JB, Andriano L, Brazel DM, Rotondi V, Block P, Ding X, Liu Y, Mills MC: **Demographic science aids in understanding the spread and fatality rates of COVID-19**. *Proc Natl Acad Sci U S A* 2020, **117**(18):9696-9698.
2. Molnar C: **Interpretable machine learning**, Retrieved edn; 2018.
3. Wiens J, Shenoy ES: **Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology**. *Clinical Infectious Diseases* 2018.
4. Wei Y, Lu Y, Xia L, Yuan X, Li G, Li X, Liu L, Liu W, Zhou P, Wang CY *et al*: **Analysis of 2019 novel coronavirus infection and clinical characteristics of outpatients: An epidemiological study from a fever clinic in Wuhan, China**. *J Med Virol* 2020.
5. Oran DP, Topol EJ: **Prevalence of Asymptomatic SARS-CoV-2 Infection**. *Annals of Internal Medicine* 2020:M20-3012.
6. Yang X, Dong N, Chan EW, Chen S: **Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries**. *Emerg Microbes Infect* 2020, **9**(1):1287-1299.
7. Chau NVV, Thanh Lam V, Thanh Dung N, Yen LM, Minh NNQ, Hung LM, Ngoc NM, Dung NT, Man DNH, Nguyet LA *et al*: **The natural history and transmission potential of asymptomatic SARS-CoV-2 infection**. *Clin Infect Dis* 2020.
8. Myers JF, Snyder RE, Porse CC, Tecle S, Lowenthal P, Danforth ME, Powers E, Kamali A, Jain S, Fritz CL *et al*: **Identification and Monitoring of International Travelers During the Initial Phase of an Outbreak of COVID-19 - California, February 3-March 17, 2020**. *MMWR Morb Mortal Wkly Rep* 2020, **69**(19):599-602.

9. Wu Z, McGoogan JM: **Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases From the Chinese Center for Disease Control and Prevention**. *JAMA* 2020.
10. Gold Jaw Fau - Wong KK, Wong Kk Fau - Szablewski CM, Szablewski Cm Fau - Patel PR, Patel Pr Fau - Rossow J, Rossow J Fau - da Silva J, da Silva J Fau - Natarajan P, Natarajan P Fau - Morris SB, Morris Sb Fau - Fanfair RN, Fanfair Rn Fau - Rogers-Brown J, Rogers-Brown J Fau - Bruce BB *et al*: **Characteristics and Clinical Outcomes of Adult Patients Hospitalized with COVID-19 - Georgia, March 2020**. *MMWR Morb Mortal Wkly Rep* 2020(1545-861X (Electronic)).
11. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, Barnaby DP, Becker LB, Chelico JD, Cohen SL *et al*: **Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area**. *JAMA* 2020(1538-3598 (Electronic)).
12. Hu Z, Song C, Xu C, Jin G, Chen Y, Xu X, Ma H, Chen W, Lin Y, Zheng Y *et al*: **Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China**. *Science China Life Sciences* 2020, **63**(5):706-711.
13. Kong W, Wang Y, Hu J, Chughtai A, Pu H, Clinical Research Collaborative Group of Sichuan Provincial People's H: **Comparison of clinical and epidemiological characteristics of asymptomatic and symptomatic SARS-CoV-2 infection: A multi-center study in Sichuan Province, China**. *Travel Med Infect Dis* 2020:101754.
14. Iniesta R, Stahl D, McGuffin P: **Machine learning, statistical learning and the future of biological research in psychiatry**. *Psychol Med* 2016, **46**(12):2455-2465.
15. Wiens J, Guttag J, Horvitz E: **A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions**. *J Am Med Inform Assoc* 2014, **21**(4):699-706.
16. Wiens J, Campbell WN, Franklin ES, Guttag JV, Horvitz E: **Learning Data-Driven Patient Risk Stratification Models for Clostridium difficile**. *Open Forum Infect Dis* 2014, **1**(2):ofu045.
17. Almazeedi S, Al-Youha S, Jamal MH, Al-Haddad M, Al-Muhaini A, Al-Ghimlas F, Al-Sabah S: **Characteristics, risk factors and outcomes among the first consecutive 1096 patients diagnosed with COVID-19 in Kuwait**. *EClinicalMedicine*.
18. Fountain-Jones NM, Machado G, Carver S, Packer C, Recamonde-Mendoza M, Craft ME: **How to make more from exposure data? An integrated machine learning pipeline to predict pathogen exposure**. *Journal of Animal Ecology* 2019, **88**(10):1447-1461.
19. Stekhoven DJ, Bühlmann P: **Missforest-Non-parametric missing value imputation for mixed-type data**. *Bioinformatics* 2012, **28**(1):112-118.
20. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, Marquéz JRG, Gruber B, Lafourcade B, Leitão PJ *et al*: **Collinearity: a review of methods to deal with it and a simulation study evaluating their performance**. *Ecography* 2013, **36**(1):27-46.
21. Kursa MB, Rudnicki WR: **Feature selection with the Boruta Package**. *Journal of Statistical Software* 2010, **36**(11):1-13.

22. Degenhardt F, Seifert S, Szymczak S: **Evaluation of variable selection methods for random forests and omics data sets**. *Briefings in Bioinformatics* 2017:1-12.
23. Haibo He H, Garcia EA: **Learning from imbalanced data**. *IEEE Transactions on Knowledge and Data Engineering* 2009, **21**(9):1263-1284.
24. Kuhn M: **Building Predictive Models in R Using the caret Package**. *Journal of Statistical Software* 2008, **28**(5):1-26.
25. Breiman L, Cutler A: **R Package 'randomForest'**. In., 4.6-14 edn; 2018.
26. Molnar C: **iml: An R package for Interpretable Machine Learning**. *Journal of Open Source Software* 2018, **3**(26):786-786.
27. Breiman L: **Random Forests**. *Machine Learning* 2001, **45**(1):5-32.
28. Goldstein A, Kapelner A, Bleich J, Pitkin E: **Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation**. *Journal of Computational and Graphical Statistics* 2015, **24**(1):44-65.
29. Friedman JH, Popescu BE: **Predictive learning via rule ensembles**. *Ann Appl Stat* 2008, **2**(3):916-954.
30. Shapley LS: **Stochastic games**. *Proceedings of the National Academy of Sciences of the United States of America* 1953, **39**(10):1095-1100.
31. Henry BM, de Oliveira MHS, Benoit S, Plebani M, Lippi G: **Hematologic, biochemical and immune biomarker abnormalities associated with severe illness and mortality in coronavirus disease 2019 (COVID-19): a meta-analysis**. *Clin Chem Lab Med* 2020, **58**(7):1021-1028.
32. Yoshimura Y, Sasaki H, Horiuchi H, Miyata N, Tachikawa N: **Clinical characteristics of the coronavirus disease 2019 (COVID-19) outbreak on a cruise ship**. *J Infect Chemother* 2020.
33. Alkhamis MA, Youha SA, Khajah MM, Haider NB, Alhardan S, Nabeel A, Mazeedi SA, Al-Sabah SK: **Spatiotemporal Dynamics of COVID-19 epidemic in the State of Kuwait**. *Int J Infect Dis* 2020.
34. Mamelund SE: **1918 pandemic morbidity: The first wave hits the poor, the second wave hits the rich**. *Influenza Other Respir Viruses* 2018, **12**(3):307-313.
35. Heinzerling A, Stuckey MJ, Scheuer T, Xu K, Perkins KM, Resseger H, Magill S, Verani JR, Jain S, Acosta M *et al*: **Transmission of COVID-19 to Health Care Personnel During Exposures to a Hospitalized Patient - Solano County, California, February 2020**. *MMWR Morb Mortal Wkly Rep* 2020, **69**(15):472-476.
36. Uribarri A, Nunez-Gil IJ, Aparisi A, Becerra-Munoz VM, Feltes G, Trabattoni D, Fernandez-Rozas I, Viana-Llamas MC, Pepe M, Cerrato E *et al*: **Impact of renal function on admission in COVID-19 patients: an analysis of the international HOPE COVID-19 (Health Outcome Predictive Evaluation for COVID 19) Registry**. *J Nephrol* 2020.
37. Cheng Y, Luo R, Wang K, Zhang M, Wang Z, Dong L, Li J, Yao Y, Ge S, Xu G: **Kidney disease is associated with in-hospital death of patients with COVID-19**. *Kidney Int* 2020, **97**(5):829-838.
38. Schuster L, Nuñez DE: **Using clinical pathways to aid in the diagnosis of necrotizing soft tissue infections synthesis of evidence**. *Worldviews Evid Based Nurs* 2012, **9**(2):88-99.

39. Chipman HA, George EI, McCulloch RE: **BART: Bayesian additive regression trees**. *The Annals of Applied Statistics* 2010, **4**(1):266-298.
40. Tsang TK, Wu P, Lin Y, Lau EHY, Leung GM, Cowling BJ: **Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study**. *Lancet Public Health* 2020, **5**(5):e289-e296.
41. Cohen HW: **P Values: Use and Misuse in Medical Literature**. *American Journal of Hypertension* 2011, **24**(1):18-23.

Figures

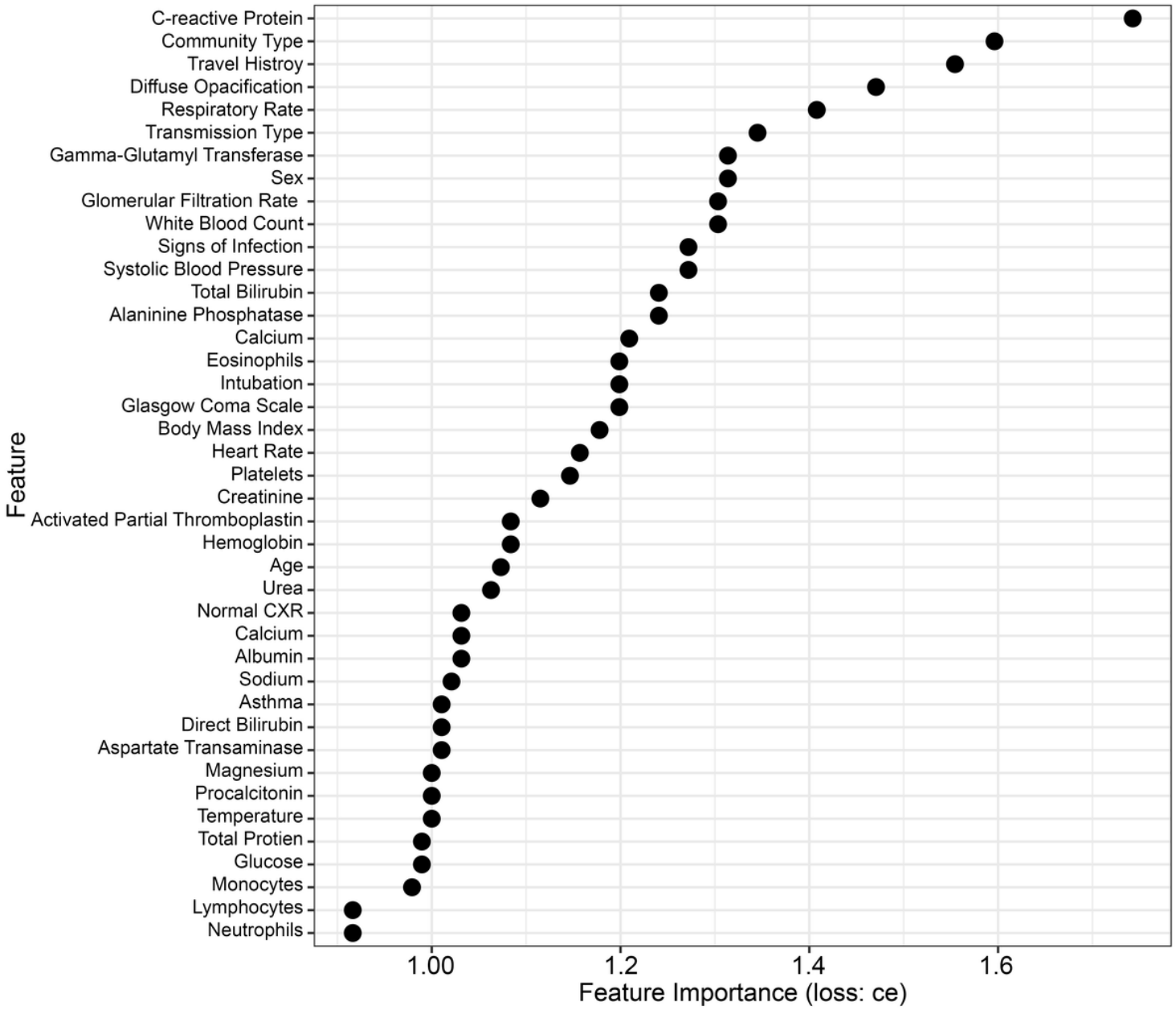


Figure 1

Plot showing feature importance that contribute to the prediction of symptom status of COVID-19 positive patients.

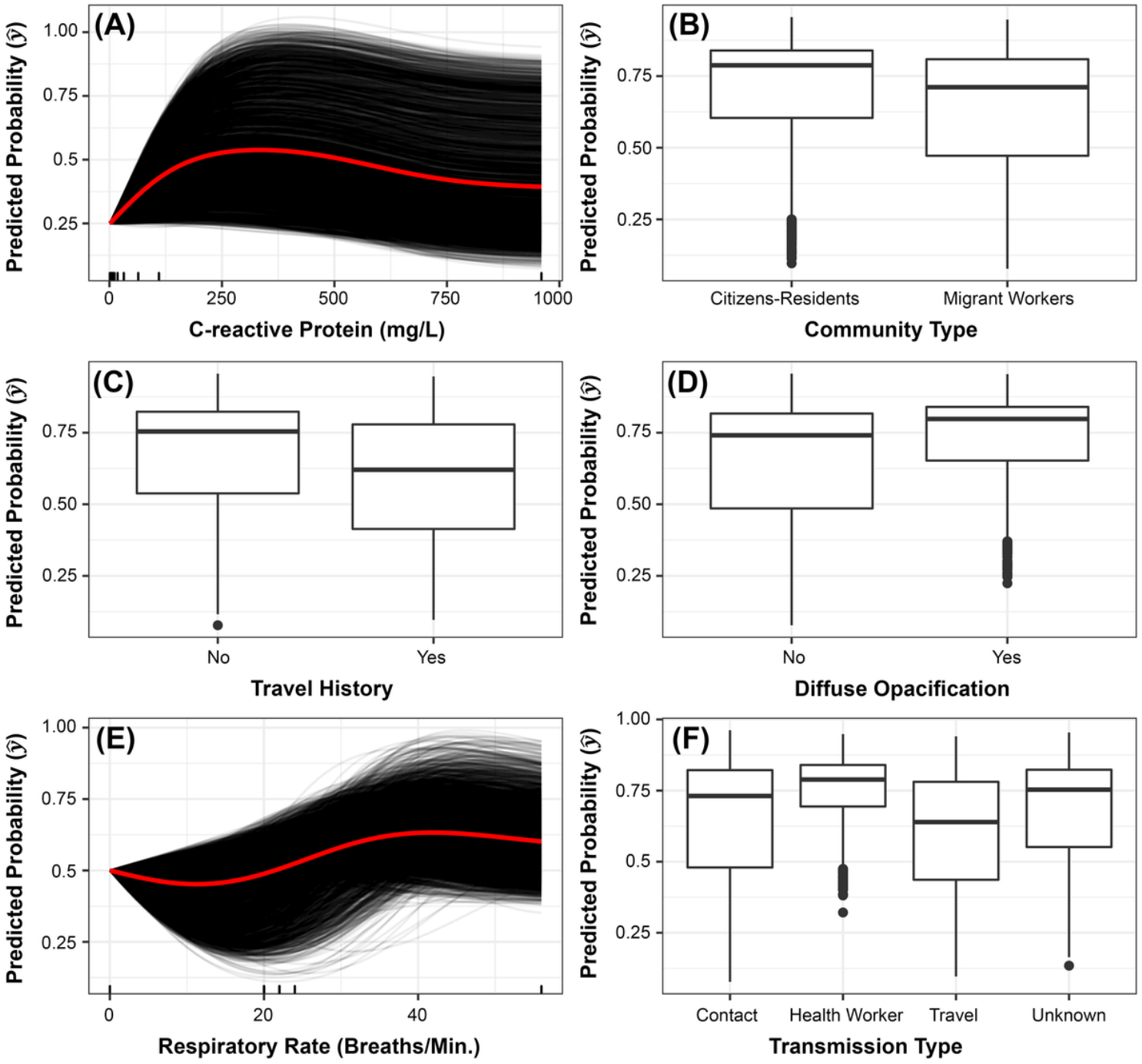


Figure 2

Centered individual conditional expectation (ICE) plots for the top six important features (see figure 1) that contribute to the prediction of symptom status of COVID-19 positive patients. The plots show the relationship between the predicted risk of symptom status and each corresponding feature, and are ordered according to their importance (A-F). (A & E) The black lines indicate the predicted risk of symptom status in each patient, while the red line indicates the partial dependence calculated as the average risk across all patients.

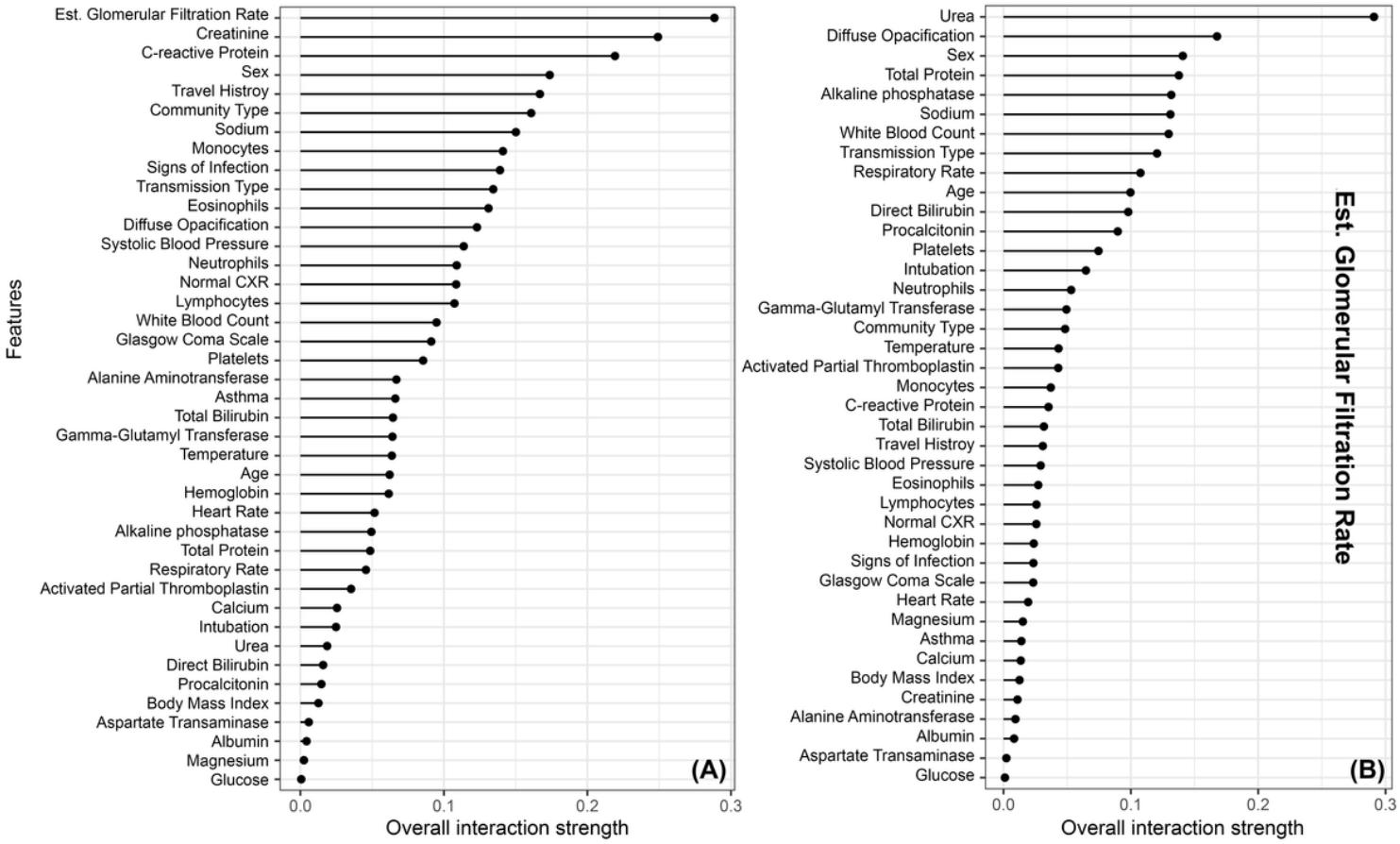


Figure 3

Feature interaction plots calculated by Friedman's H-statistic. (A) indicates overall interaction strength of each feature with the other features. (B) demonstrate the overall interaction strength of estimate glomerular filtration rate (eGFR) with the other features.

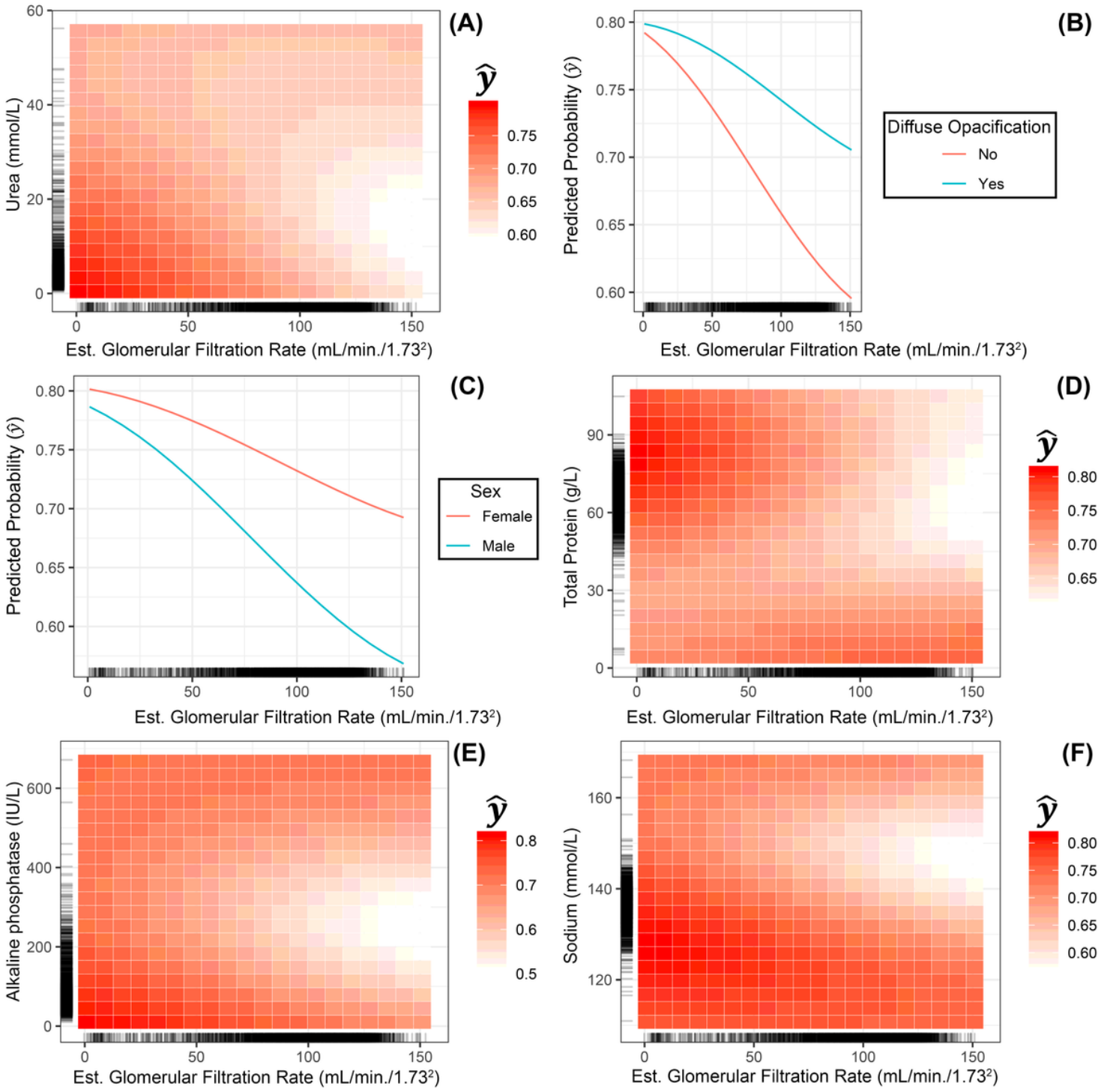


Figure 4

Partial dependence plots for the top six interactions with estimate glomerular filtration rate (eGFR) that shaped the risk of symptom status for COVID-19 positive patients. (A-F) plots are ordered according to the feature interaction strength with eGFR. (A, D, E, & F) The heat matrix corresponds to the risk of developing symptoms, in which lighter shades of red indicate lower risks of developing symptoms, and darker shades of reds indicate higher risks of developing symptoms. The bar on the right indicates ($\hat{y}$) the relative risk of being dead with all other feature combinations marginalized. (B & C) interaction between diffuse opacification status and sex on one side and eGFR on the other. Red and green partial

dependence curves represent the two features, while the y-axis indicates the predicted risk of developing symptoms.

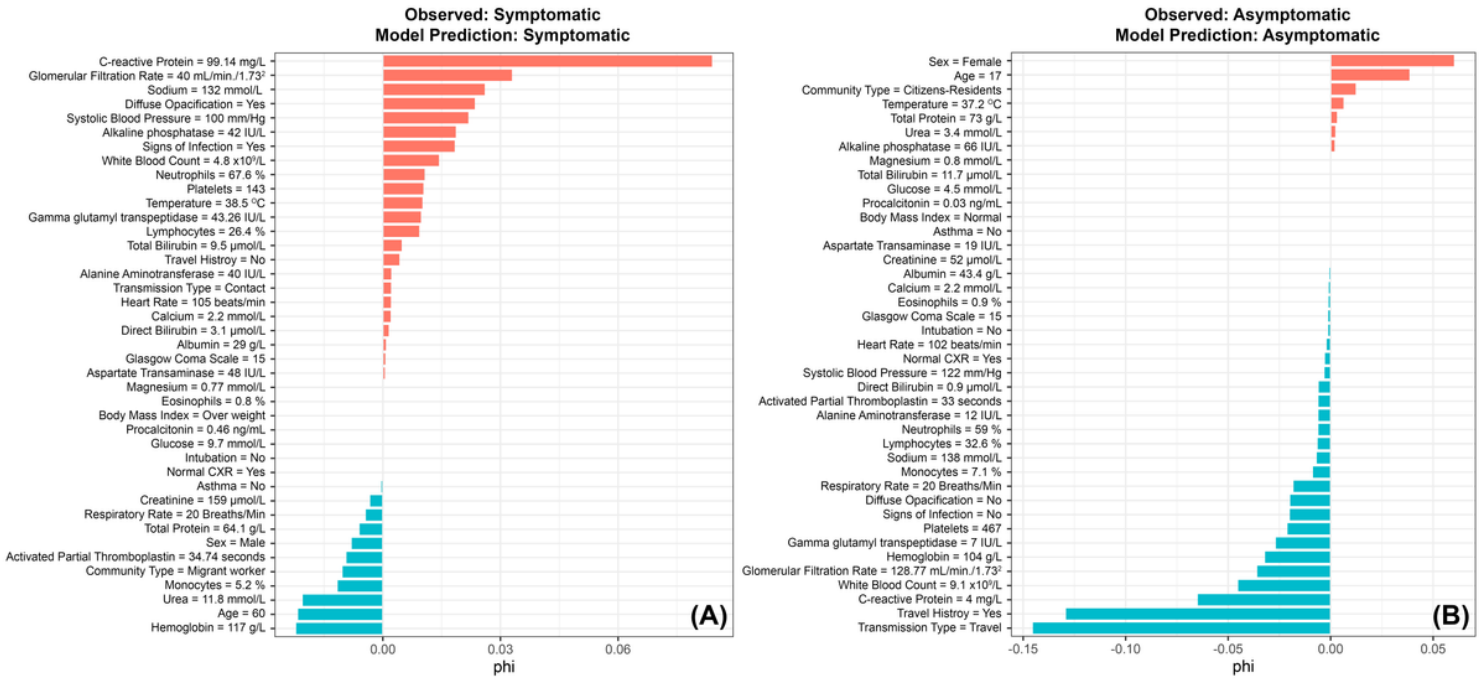


Figure 5

Plots showing feature value contributions for the respective risk of developing symptoms based on Shapley values (phi) for two individual patients. (A) observed symptomatic patient. (B) observed asymptomatic patient. Red bars indicate positive Shapley values for variables that increased the risk of developing symptoms, while green bars indicate negative Shapley values that lowered the risk of developing symptoms.

Supplementary Files

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

- Additionalfile.docx