Article │ Received 10 October 2024; Accepted 17 December 2024; Published 20 January 2025
https://doi.org/10.55092/bi20250001

Development and validation of a machine learning model elucidating risk factors in severe COVID-19

Claire Y. Zhao[1,*], Xiang (Jay) Ji[2], Shunjie Guan[3], Sima S. Toussi[4], Jennifer Hammond[5] and Subha Madhavan[3,*]

[1] AI/ML, Quantitative and Digital Sciences (AQDS), Global Biometrics and Data Management (GBDM), Pfizer Inc, Cambridge, MA, USA
[2] Data Science and Advanced Analytics, Pfizer Inc, Collegeville, PA, USA
[3] Global Biometrics and Data Management (GBDM), Pfizer Inc, Cambridge, MA, USA
[4] Anti-Infective Research Unit, Pfizer Inc, Pearl River, NY, USA
[5] Global Product Development, Pfizer Inc, Collegeville, PA, USA

* Correspondence authors; E-mails: Claire.Zhao@pfizer.com (C.Y.Z.); Subha.Madhavan@pfizer.com (S.M.).

Highlights:

- Machine learning (ML) models leveraging clinical trial and real-world data were developed to identify multivariate signatures predictive of COVID severe disease (SD) at baseline and to evaluate the generalizability of these findings.
- Nirmatrelvir/ritonavir was the greatest predictor of severe disease (SD).
- Other key baseline risk factors were elevated viral load, hsCRP, ferritin, haptoglobin, and increased age.
- Using ML to identify factors predictive of disease outcomes may aid clinical decision making and trial considerations.

Abstract: **Objectives:** COVID-19 remains a significant healthcare burden. Leveraging the combined power of clinical trial data and big data from the real world, this study elucidated baseline factors predictive of subsequent outcomes relating to severe COVID-19 disease (SD) and the effect of nirmatrelvir/ritonavir (Tx), a protease inhibitor, on disease progression. **Methods:** We retrospectively analyzed data from the Evaluation of Protease Inhibition for COVID-19 in High-Risk Patients (EPIC-HR) clinical trial (NCT04960202) to discern observational associations between baseline factors and subsequent SD outcome. Baseline factors, including demographics, clinical laboratory results, symptoms, medical history, vital signs, and electrocardiogram features, were studied using machine learning (ML) for their importance in predicting hospitalization or death through Day 28, with Tx effects analyzed statistically. Generalizability of results was evaluated using real-world data (RWD) Optum Electronic Health Records. **Results:** Modeling indicated Tx was the greatest predictor of whether a patient progressed to SD. The most important baseline factors associated with increased risk of SD

Copyright©2025 by the authors. Published by ELSP. This work is licensed under Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

were elevated baseline (1) viral load (VL; > ~4 log_{10} copies/mL), (2) hsCRP (> ~1 mg/dL), (3) ferritin (> ~280 ug/L), (4) haptoglobin (> ~210 mg/dL), and (5) increased age (> ~48 years). Tx reduced VL and abnormally high hsCRP and haptoglobin to greater extents than placebo at the measured time points. RWD validation supported findings on increased risk with elevated hsCRP and ferritin and increased age (no data were available on VL and haptoglobin). **Conclusion:** ML analysis identified critical baseline factors immediately before or at the beginning of COVID-19 infection predictive of progression to SD in adults that are common to a heterogeneous population. This study provides insights on multivariate signatures of COVID-19 disease progression and Tx effects, which may aid future studies and inform treatment decision making.

Keywords: big data; real-world evidence; machine learning; precision medicine; COVID-19; EPIC-HR

1. Introduction

The COVID-19 outbreak, officially declared a global pandemic in March 2020, remains a serious public health threat [1,2]. As of early August, 2024, there were over 776 million confirmed COVID-19 cases and 7.1 million COVID-19 deaths worldwide [3], with 4.6 per 100,000 weekly laboratory-confirmed COVID-19 hospitalizations and 498 weekly COVID-19 deaths in the United States alone in the same period [4]. Diagnosis of COVID-19, which typically relies on development of characteristic symptomatology and subsequent confirmatory diagnostic testing, is challenging because SARS-CoV-2 infections demonstrate variable symptomatology, prolonged incubation periods, and high rates of asymptomatic infection [5–7]. Thus, reported COVID-19 cases likely underestimate the true number of global infections and reinfections identified by prevalence surveys, partially due to reduced testing and reporting delays [8]. Severe COVID-19 disease (SD) is associated with increased risk of long-term sequelae, known as Post-COVID Conditions or Long COVID [2]. Known risk factors for SD, hospitalization, or death include age, cardiovascular disease, chronic lung disease, chronic kidney disease, and immunosuppression [9–11]. However, other factors associated with progression of SD, such as baseline laboratory findings, are not well characterized and warrant further study.

We constructed machine learning (ML) models that leveraged the strength of both clinical trial data and real-world data to derive multivariate signatures of SD and elucidate the treatment effect of nirmatrelvir/ritonavir (Tx), which contributed to precision medicine decision making.

2. Methods

2.1. Evaluation of Protease Inhibition for COVID-19 in High-Risk Patients (EPIC-HR) data

The EPIC-HR study was a phase 2–3, double-blind, randomized, controlled trial active between July and December 2021 that enrolled nonhospitalized symptomatic adults with COVID-19 and increased risk of progressing to SD due to age ≥ 60 years, body-mass index > 25, immunosuppression, presence of cardiovascular diseases, and/or diabetes [9–12]. The study enrolled patients unvaccinated against COVID-19 and required initial onset of signs/symptoms attributable to COVID-19, such as cough, shortness of breath or difficulty breathing, and fatigue within 5 days before or on randomization day. In addition, patients' COVID-19 status and treatment efficacy were assessed by baseline and changes in viral load (VL), defined as the level of SARS-CoV-2 RNA (log_{10} copies/mL) in nasopharyngeal swab samples

quantified by PCR, with a lower limit of quantification of 2.0 $\log_{10}$ copies/mL. For statistical analyses, values < 2.0 were imputed to 1.7 $\log_{10}$ copies/mL, and not detected was imputed to 0 $\log_{10}$ copies/mL. A VL of 4.0 $\log_{10}$ copies/mL was considered elevated for this study. The primary endpoint was the proportion of participants with COVID-19–related hospitalization or death from any cause through Day 28 [12], which was used in this study to characterize SD. Participants were randomized 1:1 to receive either Tx or placebo (PBO) every 12 hours for 5 days. The full details of the EPIC-HR study are provided elsewhere (NCT04960202; ClinicalTrials.gov) [12].

The study cohort used herein was extracted based on the modified intent-to-treat 2 population, comprising all participants randomly assigned to the Tx arm who received ≥ 1 dose of study intervention. Participants who did not get hospitalized or die through Day 28 were designated as Class 0; conversely, participants who were hospitalized or died before or on Day 28 were designated as Class 1. Of the 2091 participants ultimately selected for the study cohort, 76 (3.63%) progressed to SD (Appendix Table S1a) such that the PBO arm comprised 1053 participants with 66 (3.16%) SD cases and the Tx arm comprised 1038 participants with 10 (0.48%) SD cases (Appendix Table S1b).

2.2. Development of the predictive SD ML model

In a retrospective analysis, baseline factors (also referred to as features) were ranked by their contribution to SD risk by constructing and interpreting a predictive ML model on the EPIC-HR study cohort, which will be subsequently referred to as the SD ML model. All 99 factors available were used to engineer features for model inputs, including demographics, clinical laboratory values, symptoms, medical history, vital signs, and electrocardiogram features. Treatment arm assignment was input as an indicator variable, with one encoding assignment to the Tx arm and zero the PBO arm (Appendix Table S2). For the ML model to be predictive, only data measured before the occurrence of endpoints were allowed as inputs; thus, the SD model only included baseline measurements because death/hospitalization could have occurred any time post-baseline.

Five ML algorithms were constructed and compared for predictive performance: logistic regression (LogitReg), random forest (RF), balanced random forest (BRF), light gradient boosting machine (LGB), and XGBoost (XGB). Nested cross validation (nCV) framework was used to train and evaluate the models [13]. Feature selection and imputation were completely performed on the training sets from the inner folds of nCV, while the held-out outer folds were completely unseen in model construction. This nCV process resulted in five "outer models" for each ML algorithm. The best-performing algorithm from nCV was subsequently retrained on the full EPIC-HR dataset and designated the "benchmark model." The performance of this model was assumed to be comparable with, if not superior to, the average scores of outer models of the chosen algorithm without inadequately overfitting the training sets [13]. To optimize ML models against high-class imbalance (*i.e.*, prevalence of Class 1 is ~3.5%) with relatively low sample size, performance metrics for model selection emphasized the F1 score, which is the harmonic mean of precision and recall. Furthermore, we focused on constructing ensemble tree-based algorithms with inverse class weighting, such as RF, BRF, LGB, and XGB. These generally excel in imbalanced datasets because their hierarchical structure allows them to learn from both classes. Since these methods tend to overfit to training samples, we also selected for minimum overfit in F1 score for increased potential of model generalizability. See Appendix Section 1.2–1.5 for a more detailed description of the ML model optimization.

2.3. ML model interpretation

The contribution of each feature to SD risk, or feature importance, was determined by Shapley additive explanation (SHAP) values. SHAP values are a measure of individualized feature importance. This measure helps clarify the combined effect of multiple features on the model's output or risk score (*i.e.*, SD risk) because the values add up to the risk score for each participant. The importance of the feature for the entire study cohort, or global feature importance, was evaluated by global SHAP values, computed as the average absolute values of the individualized SHAP values. The contributions of input factors to SD risk were ranked by their respective global feature importance measures.

Statistical analyses were conducted to determine Tx impact on factors prioritized by the SD ML model (Appendix Section 2), designed for post hoc interpretation of multivariate ML results. The factors ranked high in importance to SD risk by ML that were potentially modifiable by treatment were selected. For each selected factor, mean values at each available measured time point and their respective 95% confidence intervals (CIs) were calculated for participants, with the respective factor values falling outside the clinically accepted normal range for the PBO and Tx arms. The mean change from baseline (CFB) on each day was compared between the two arms by two-sided two-sample *t* test ($P \leq 0.05$). If the selected factor had different abnormal thresholds for different subpopulations as defined by the laboratory specification, the analysis was repeated for each subpopulation (Appendix Section 2).

2.4. Real-world data for validation

The Optum Electronic Health Record (EHR) dataset is an in-house collection of US health records that was selected as an independent dataset to assess the generalizability of findings obtained from the SD ML model analysis developed using EPIC-HR data. These real-world data (RWD) were matched wherever possible to the EPIC-HR study. Patients age $\geq$ 18 years diagnosed with COVID-19, as identified by the *International Classification of Diseases, Tenth Revision* code of U07.1, from June 1 to December 15, 2021, were included. Since RWD were sparse, especially for symptom and laboratory measurements, COVID-19–related symptoms were not required for study cohort inclusion. In addition, time windows relative to the indexed COVID-19 diagnosis allowable for feature value inclusion were expanded to catch more data, with symptoms expanded to 10 days and laboratory values to 28 days before the indexed COVID-19 diagnosis date. In case of multiple measurements within the time window, the closest to the index date was chosen. Measurement units were converted to match those from EPIC-HR.

2.5. Evaluation of generalizability of ML modeling results

To evaluate the generalizability of findings based on the SD ML model, additional ML models were constructed with RWD using nCV framework and interpreted by SHAP as described previously. Within the time frame of our RWD extraction, the data that were captured corresponded to Delta as the dominant viral strain, matching that of EPIC-HR; however, this time frame was before the US Food and Drug Administration granted Emergency Authorization of Tx [14]. As a result, Tx effect could not be validated, so validation instead focused on the untreated population. The following three strategies were designed to seek potential evidence supporting the finding of prioritized factors: (1) use EPIC-HR PBO data to train the model via nCV and evaluate the resulting model using RWD; (2) compare RWD study

population means for Class 1 and Class 0 for prioritized factors; and (3) use RWD in nCV to check if prioritized factors were also prioritized in RWD.

3. Results

3.1. Factors critical to SD risk

The BRF algorithm with average test ROC-AUC and F1 scores across five outer nCV folds of 0.859 and 0.630, respectively, and their respective mean absolute deviation (MAD) of 0.032 and 0.034 was selected and retrained on the entire EPIC-HR data to obtain the benchmark model (see Appendix 3). Based on this model, baseline risk factors were ranked by the magnitude of their contribution to SD risk for untreated (Figure 1a) and treated participants (Figure 1b) in descending order of importance. Each dot on the plot represents the contribution of the corresponding feature to SD risk for one study participant: red and blue denote high and low feature values, respectively, and the amount of feature contribution to SD is quantified in the x-axis. The black vertical line at zero denotes no contribution to risk; the farther away features are from this vertical line, the greater the contribution of the feature to model output risk score, with positive x-direction indicating increase in risk and negative direction decrease in risk.

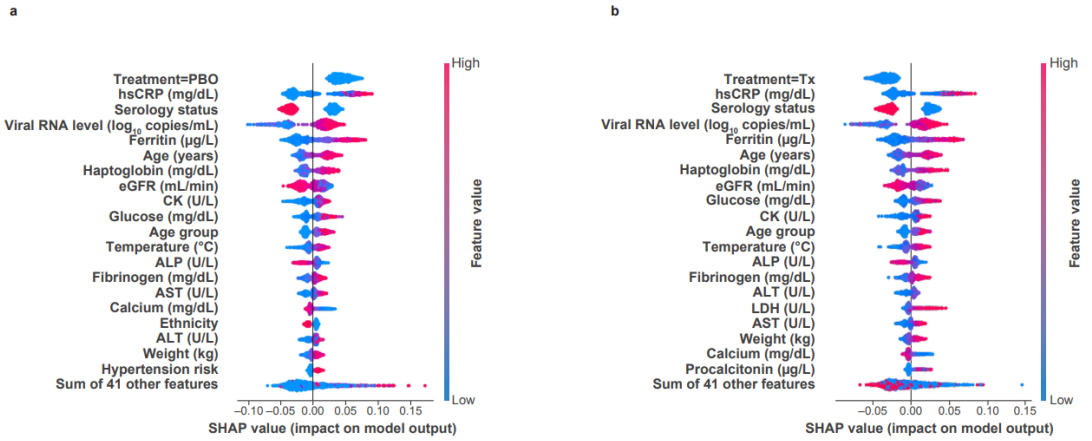


Figure 1. Contribution of baseline factors to SD risk. The top factors contributing to SD risk are illustrated by the beeswarm plot in (a) the PBO arm and (b) the Tx arm. Features are ranked in decreasing order of their respective global SHAP values on each arm based on the benchmark model. The top 20 most important features are shown from a total of 61 input features to the SD ML model after feature selection. Each dot on the plot represents how the corresponding feature contributed to SD risks for one study participant: red and blue colors denote high and low feature values, respectively, and the feature contributions to SD are shown against the x-axis. The black vertical line at zero denotes no contribution to risk; the farther away features are from this vertical line, the greater the contribution of the feature to model-output risk score, with a positive x-direction indicating increase in risk and negative direction decrease in risk. The width of the jitters in the y-axis direction is proportional to the number of participants with the same feature value and feature contribution. Serology status is positive if either one of qualitative IgG-IgM anti-N or anti-S is positive. Age group is defined as 18–44, 45–59, 60–64, 65–74, and ≥ 75 years, with increase in age category encoded by a higher integer value. Ethnicity is encoded as Hispanic or Latino (1) and Not (0). Note the unit for eGFR is mL/min/1.73 m^2 more specifically. Abbreviations: ALP: alkaline phosphatase; ALT: alanine transaminase; AST: aspartate aminotransferase; CK: creatinine kinase; eGFR: estimated glomerular filtration rate calculated by the CKD-EPI (Chronic Kidney Disease-Epidemiology Collaboration) equation using serum creatinine in the Evaluation of Protease Inhibition for COVID-19 in High-Risk Patients trial data; hsCRP: high-sensitivity C-reactive protein; IgG: immunoglobulin G; IgM: immunoglobulin M; LDH: lactate dehydrogenase; PBO: placebo; SD ML: severe COVID-19 disease machine learning; SHAP: Shapley additive explanation; Tx: nirmatrelvir/ritonavir; VL: SARS-CoV-2 viral RNA level, also referred to as viral load.

As shown in Figure 1a, not receiving Tx imposed the greatest contribution to increasing SD risk for the untreated population and increased risks for all untreated participants. Receiving Tx contributed the most to decreasing risk of SD, with all participants contributing negative SHAP values (*i.e.*, decreased SD risk) (Figure 1b). The remaining top ten baseline factors prioritized were the same between the PBO and Tx groups (Figures 1a and b). High baseline high-sensitivity C-reactive protein (hsCRP) increased SD risk, while lower values alleviated risk. Positive serology status (either one of qualitative IgG-IgM anti-N or anti-S is positive) decreased risk, while a negative status increased risk for all untreated participants. Elevated baseline VL, ferritin, haptoglobin, and increased age were also top predictors for increased SD risk. Furthermore, low baseline estimated glomerular filtration rate (eGFR), high creatinine kinase (CK), and high glucose increased SD risk. Higher age group was near the top ten features and mirrored the importance of increased age in SD. Although below the top ten in ranked importance, elevated baseline temperature, fibrinogen, aspartate aminotransferase, alanine transaminase, and weight also contributed to increased risk for SD, as did low baseline alkaline phosphatase and calcium levels. At the bottom of the ranked lists, slightly higher SD risk was observed for participants not of Hispanic or Latino ethnicity and with hypertension risk in the PBO group (Figure 1a) and increased baseline lactate dehydrogenase and procalcitonin for the Tx group (Figure 1b). Together, the above were the top 20 risk factors, while contribution of the remaining ones were small individually, and their combined contributions are shown in the last row of the figure.

3.2. Persistence of prioritized factors

Because the definition of high risk in EPIC-HR included many underlying conditions, we tested if the above identified factors would change due to variation in the population characteristics. We checked feature importance in the five different outer models trained and tested with different, randomly selected, class-prevalence stratified data splits per the nCV framework. Receiving Tx (or not) and baseline ferritin, hsCRP, VL, haptoglobin, and age were among the top ten features in all five outer BRF models, providing evidence that these six variables are common characteristics for progression to SD that can withstand variations in the study population. Other clinical features are sensitive to heterogeneity in the study population and thus model parameterization; these should be cautiously interpreted. In fact, baseline serology status and glucose appeared in four of the five outer models. Baseline CK appeared in three, but no other variables appeared in more than half of the outer models.

Additionally, we obtained the thresholds of the continuous features that were prioritized in all five outer models that lead to increased SD risk. For each feature, this was achieved by scattering SHAP values of the feature from the benchmark model against values of the feature, with the threshold being the feature value that divided positive and negative SHAP values for the majority of participants in the untreated population. We observed that baseline hsCRP > ~1 mg/dL, VL > ~4 $\log_{10}$ copies/mL, ferritin > ~280 µg/L, and haptoglobin > ~210 mg/dL contributed to increased SD risk for most participants. Most participants aged > ~48 years were also at increased SD risk. Similar thresholds were also observed in the Tx arm. Because this is a multivariate analysis, each of the clinical variables should be interpreted in context of the others studied. Having one feature exceeding the threshold increases SD risk for most participants, but not for all. Furthermore, having ≥ 1 feature contributing positively to SD risk does not guarantee SD progression, as risk depends on multiple features. However, having a higher number of the prioritized features in the extremes of the identified abnormal threshold does substantially

increase the probability of progression. As shown in Figure 1, the mixtures of red and blue for all feature values (except treatment and serology status) at any given feature importance (*i.e.*, SHAP value) suggest that for a given level of contribution to SD risk, the feature value may be different between different participants, driven by the values of the other features.

3.3. Characteristics of participants at high risk of SD

The characteristics of participants at high risk of SD are further illustrated in Figure 2. The figure delineates how each feature additively drives changes in SD risk score, f(x), for each participant, thus providing interpretation of each feature in a multivariate context and complementing the feature-centric view in Figure 1. The heatmaps rank SD risk in decreasing order for each untreated participant according to SD progression status (Figure 2). The features are also ranked by decreasing order of importance for the respective groups studied. As shown in Figure 2a, not receiving Tx was the largest contributor to SD risk. Overall, the prioritized factors contributed to increased risk (red) for severe SD participants (Figure 2a), but alleviated risk (blue) for those who did not progress (Figure 2b). In patients who did not experience SD, most of the prioritized factors contributed to decrease in risk, which together sufficiently lowered the risk of SD. There were, however, a small number of exceptions, especially when the risk score was at an intermediate level.

3.4. Impact of treatment

The impact of Tx on the following top seven modifiable features was studied: (1) hsCRP, (2) VL, (3) ferritin, (4) haptoglobin, (5) eGFR, (6) glucose, and (7) CK. Statistically significant results comparing mean BL or CFB between Tx and PBO for subsequent measured days ($P \leq 0.05$) are shown in line plots in Figure 3 (further detailed in Appendix 2.1). The means (solid line) and 95% CIs (shades) are shown for VL, hsCRP, and haptoglobin for participants in the PBO (black) and Tx (blue) groups in Figure 3a–c, respectively. Only participants with hsCRP or haptoglobin values above reference range at baseline were included in the analysis (for sample sizes, see Appendix Table S3). Tx reduced VL to greater extents than PBO, especially on Day 3 and Day 5 (Figure 3a). Similarly, Tx further reduced hsCRP, especially at end of treatment on Day 5 (Figure 3b). In the PBO group, haptoglobin decreased on Day 14 (Figure 3c, black), whereas in the Tx group, it initially decreased on Day 5 and decreased further on Day 14 (Figure 3c, blue). Differences in mean CFB values between the Tx and PBO groups for abnormal levels of glucose, ferritin, CK, and eGFR were not statistically significant (Appendix 4.2).

Furthermore, in participants who did not progress to SD, hsCRP and haptoglobin decreased with time (Appendix Figure S4 and Table S4). Treatment (blue) decreased their levels to greater extents compared with PBO (black) with statistical significance on measured days. By contrast, in SD participants, there was no statistical difference of hsCRP and haptoglobin levels between the PBO and Tx groups; hsCRP and haptoglobin remained high. Compared with PBO (black), Tx (blue) on average decreased hsCRP and haptoglobin on Day 5 but not on Day 14, although not in a statistically significant manner on either day. In addition, at baseline, hsCRP and haptoglobin levels were higher in the SD cohort than the non-SD cohort, consistent with results from ML modeling.

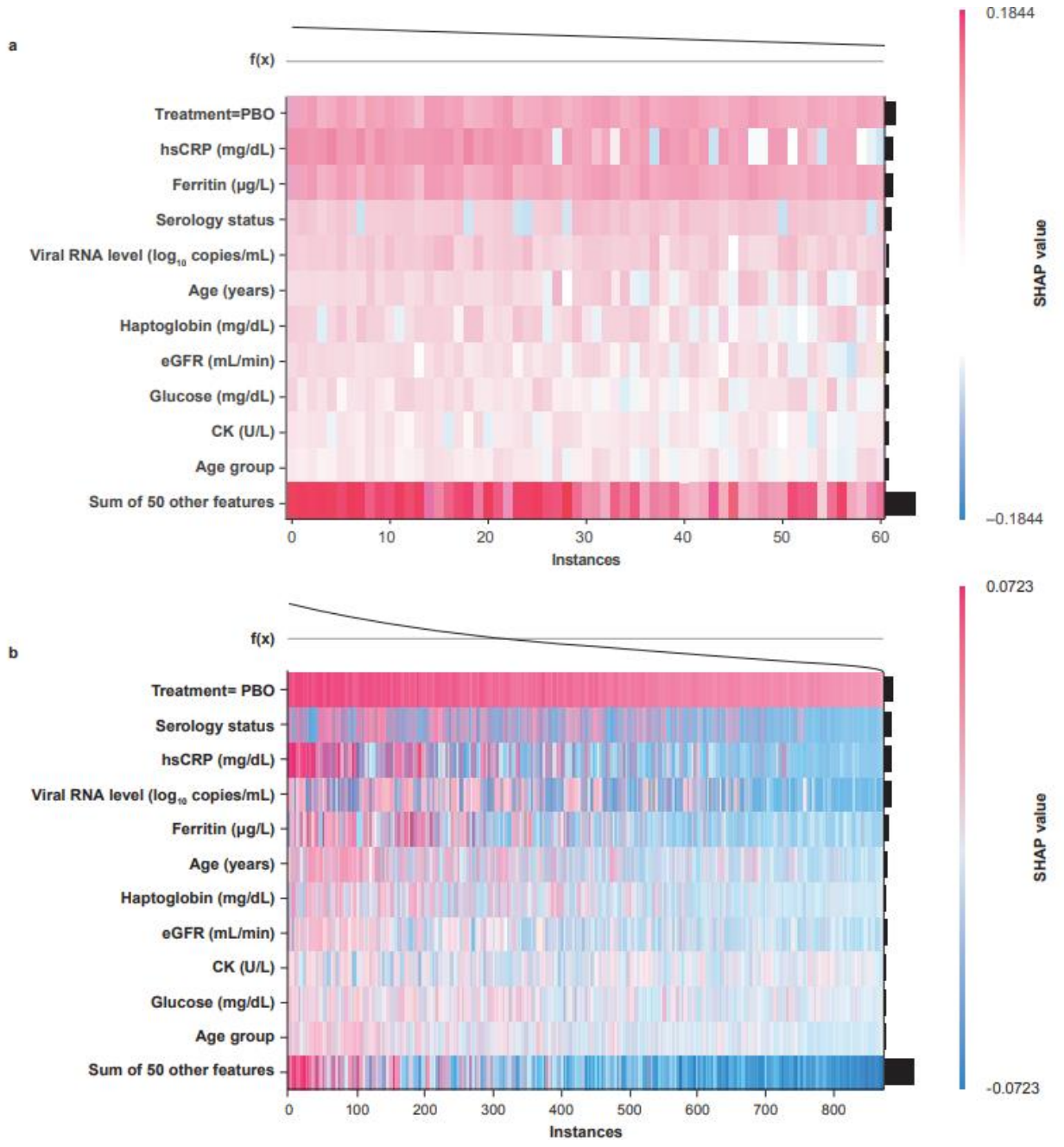


Figure 2. Contribution of prioritized baseline factors to SD risk for each untreated participant. Heatmaps for untreated participants who (a) progressed to SD and (b) who did not progress to SD. Each column represents a participant (*i.e.*, instance), and the model input features are shown on the y-axis. Individualized SHAP values are illustrated by heatmap for the features shown: red shading indicates a positive contribution to risk and blue shading reduced risk. The higher the intensity of the color, the higher the magnitude of contribution to SD risk the feature has for the individual. The global feature importance of each feature is shown as the black bar on the right-hand side of the heatmap and is measured by the mean absolute value of the corresponding individualized SHAP values. Features are ranked in descending order of their respective global feature importance measures. Only instances where the model predicted correctly are shown to guide the correct feature interpretation. Abbreviations: CK: creatinine kinase; eGFR: estimated glomerular filtration rate calculated by the CKD-EPI (Chronic Kidney Disease—Epidemiology Collaboration) using serum creatinine in the Evaluation of Protease Inhibition for COVID-19 in High-Risk Patients trial data; hsCRP: high-sensitivity C-reactive protein; f(x): risk score (*i.e.*, output of the benchmark model); PBO: placebo; SD: severe COVID-19 disease; SHAP: Shapley additive explanation.

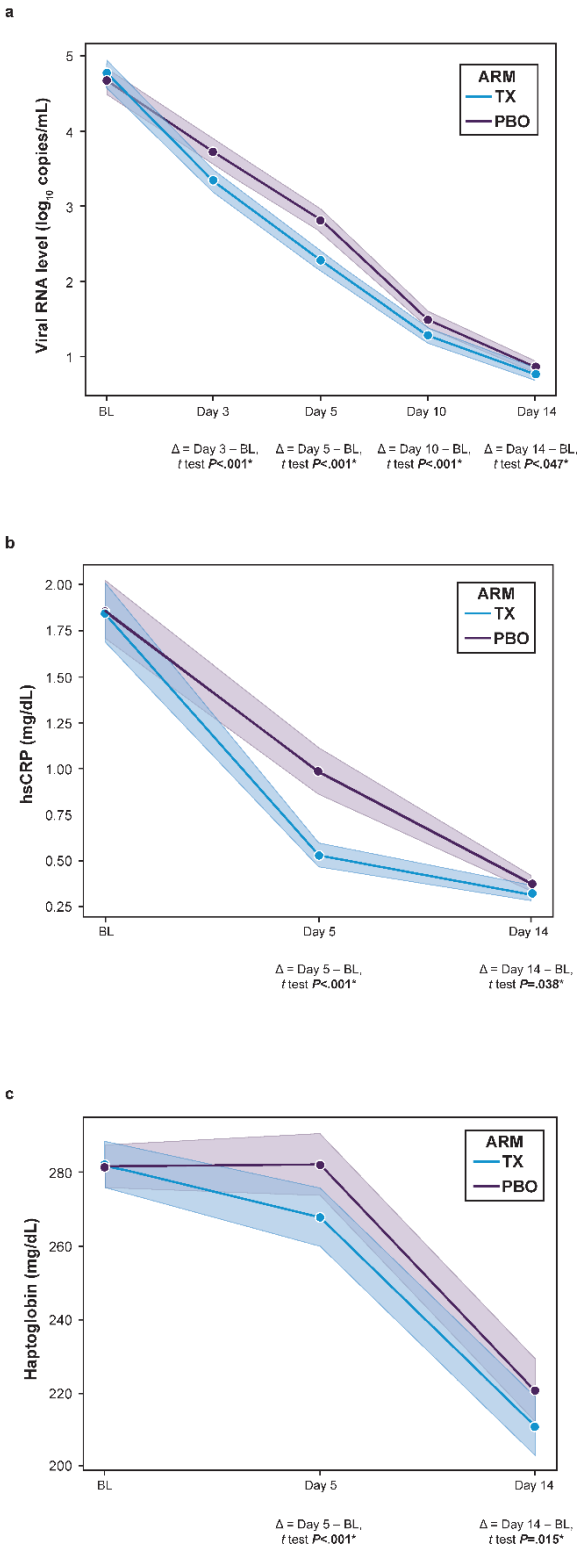
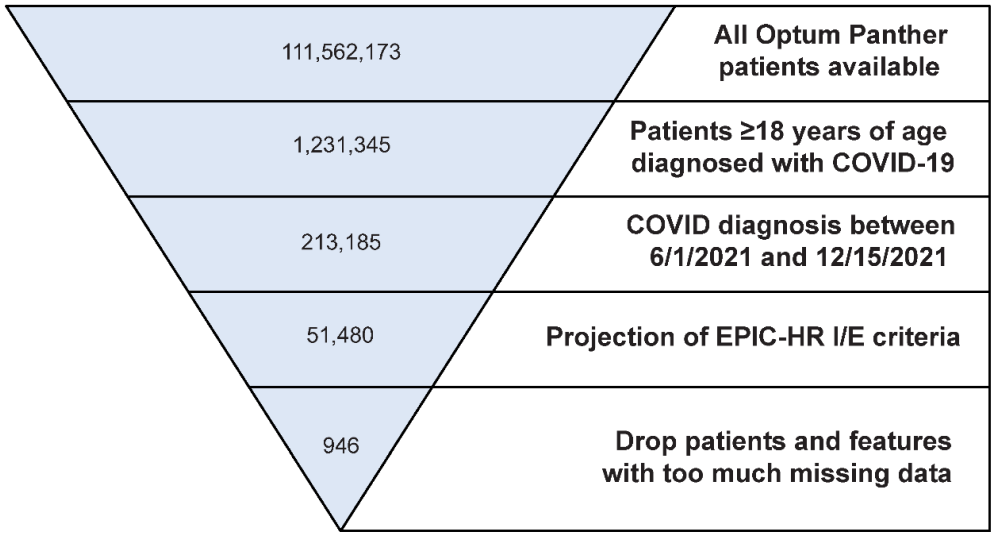


Figure 3. Impact of treatment on prioritized clinical features. Line plots are shown for (a) VL, (b) hsCRP, and (c) haptoglobin. Blue indicates Tx arm and black PBO arm. Only participants with abnormal hsCRP (> 0.5 mg/dL) and haptoglobin (> 200 mg/dL) at baseline are included in the analysis. Means of laboratory values are shown by a solid line with 95% CIs in shaded areas. The dot on the line indicates the time point at which the data were measured. *P* values from two-sample *t* tests on means at baseline or means of change from baseline on subsequent days between the PBO and Tx groups are shown at the bottom of the plot for each subsequent day measured (see Appendix 2.1 for detailed statistical methods). Statistically significant *P* values ($P \leq 0.05$) are bolded and indicated with an asterisk. Abbreviations: BL: baseline; hsCRP: high-sensitivityC-reactive protein; PBO: placebo; Tx: nirmatrelvir/ritonavir; VL: SARS-CoV-2 viral RNA level, also referred to as viral load.

a



b

	ROC-AUC	Precision	Recall	F1 Score
Test Performance on EPIC-HR PBO (Mean [MAD])	0.806 (0.029)	0.616 (0.046)	0.685 (0.021)	0.628 (0.034)
Test Performance on RWD dataset	0.601	0.608	0.601	0.604

c

Prioritized factors	*P* value
hsCRP (mg/dL)	3.07e-18
Ferritin (µg/L)	5.56e-8
Age (years)	3.20e-4

d

Prioritized factors	BRF	RF
hsCRP (mg/dL)	5	5
Ferritin (µg/L)	1	2
Age (years)	1	1

Figure 4. RWD validation results. Panel (a) shows the number of patient records remaining after performing each filtering step in the process of projecting EPIC-HR cohort onto RWD. Panel (b) shows Validation Strategy 1 results (EPIC-HR PBO data to train model via nCV and evaluate the resulting model using RWD). Performance was evaluated by ROC-AUC, precision, recall, and F1 score on the specified unseen test set. First row: test performances averaged across the five outer models in the nCV when training and testing on EPIC-HR PBO data (n = 1053 in total for EPIC-HR PBO arm, of which 80% was reserved for training and validation and 20% were held out for testing). Mean and MAD in brackets are reported for each metric. Second row: test performance of the developed ML model on RWD (RWD test set n = 946). Panel (c) shows the results for Validation Strategy 2 (compare RWD study population means for Class 1 and Class 0): *P* values of the prioritized factors hsCRP, ferritin, and age from two-sample *t* test comparing means of the SD and non-SD subgroups. Missing data were discarded, and the remaining data were log-transformed before hypothesis test. Sample sizes are shown as (Class 0, Class 1): hsCRP (495, 135), ferritin (462, 108), age (778, 168). Panel (d) shows results for Validation Strategy 3 (use RWD data in nCV to check if previously prioritized factors were also prioritized in RWD): the number of folds in the nCV (out of total of five) where the previously prioritized factors appeared among the top ten features ranked by SHAP values for the BRF and RF models constructed using RWD. The higher number of folds a factor appears in, the more likely the factor generalizes across the heterogeneity in the study population. Total sample size is 946 (*i.e.*, the entire RWD study cohort), of which 80% were for training and validation and 20% were held out for testing. Abbreviations: BRF: balanced random forest; EPIC-HR: Evaluation of Protease Inhibition for COVID-19 in High-Risk Patients trial; hsCRP: high-sensitivity C-reactive protein; ML: machine learning; MAD: mean absolute deviation; nCV: nested cross validation; PBO: placebo; RF: random forest; ROC-AUC: area under the curve (AUC) for the receiver operating characteristic (ROC) curve; RWD: real-world data; SD: severe COVID-19 disease; SHAP: Shapley additive explanation.

3.5. Validation using RWD

An independent dataset was used to validate the features (baseline ferritin, hsCRP, VL, haptoglobin, and age) prioritized in all five outer models of the SD ML model developed using EPIC-HR data. Figure 4a shows the process that projects the EPIC-HR study cohort onto Optum-EHR data, with details stated in Appendices 5.1 and 5.2. Ultimately, 63 features remained, with ferritin (39.7% missing), hsCRP (33.4% missing), and age (no missing values) remaining of the five prioritized features. Notably, RWD did not have VL or serology status, and there were limited available symptom data without severity grading (binary variable only). The RWD cohort contained 946 patients with 17.8% having SD (Class 1). COVID-19 vaccination status could not be matched to EPIC-HR, which recruited only unvaccinated individuals, because the RWD source data on vaccination status was unreliable. Based on surveillance readouts, a sizable proportion of the US population would have been vaccinated for COVID-19 by the specified time [4].

For validation Strategy 1, BRF was the best performing model (Appendix 5.3), with average test performance and MAD in brackets on EPIC-HR PBO from the nCV shown in the first row of Figure 4b. Test performance on RWD (Figure 4b, second row) had a lower ROC-AUC, but recall, F1 score, and precision were similar. Recall degraded more than precision due to a larger number of false positives, which was expected since RWD had more Class 1 cases than the EPIC-HR PBO data. This model also prioritized hsCRP and ferritin for all five folds of the nCV and age for four folds. Results indicated that the model's ability to identify patients at risk of SD and its prioritized risk factors were largely generalizable between EPIC-HR and RWD. For Strategy 2, the means of the three prioritized factors were significantly different between SD and non-SD patients in RWD (*P* values are shown in Figure 4c). For Strategy 3, when ML models were trained entirely on RWD using nCV, BRF and RF performed virtually equivalently (Appendix 5.4). Figure 4d shows the number of outer models in the five-fold nCV in which hsCRP, ferritin, and age appeared among the top ten most important features by SHAP value. The importance of hsCRP in SD was clear as it appeared in all five folds. Ferritin appeared in fewer folds, possibly because it was missing for ~40% of the patients in the cohort, and data imputation per nCV framework (Appendix 1.3) degraded the data quality. Age only appeared in one fold each for BRP and RF, possibly because when controlling for missing data, the RWD cohort likely captured a sicker cohort, as characterized by a higher prevalence of hospitalization/death than EPIC-HR. In this case, age may be comparatively less important in determining SD than other factors such as comorbidities. It is to be noted, for both RF and BFR, all three factors were selected by hypothesis-testing-based feature selection across all nCV folds.

4. Conclusion

This study suggests Tx is the greatest predictor of whether a patient progresses to SD. Elevated baseline VL, hsCRP, ferritin, haptoglobin, and increased age are the most important clinical features that are likely common to patients with heterogeneous underlying conditions. Tx reduces VL and abnormal values of hsCRP and haptoglobin to greater extents than PBO. Many of the identified factors in this study are markers of activated immune response, inflammatory response, and multiorgan damage. Prior evidence exists that certain laboratory findings such as elevated CRP and ferritin may predict severe disease, but mostly in a univariate fashion, where relative priority to other factors could not be studied. It was also not always clear if the measurements were obtained at the baseline of COVID-19 infection. The multivariate

analysis here offered by ML methodology provides a holistic view of baseline patient characteristics contributing to increased risk of SD in adult patients with COVID-19.

Previously, CRP, ferritin, and haptoglobin have been separately shown to be abnormal in patients with COVID-19 or SD. CRP, already abnormally elevated in COVID-19, was up to two-fold higher in SD [15–17]. Serum CRP is an acute-phase protein and active regulator of host innate immunity, which was found to be highly predictive of the need for mechanical ventilation and has been proposed to guide escalation of treatment of COVID-19–related uncontrolled inflammation [15,16]. COVID-19 infection is associated with iron overload in patients potentially related to abnormal ferritin levels, hemoglobin denaturation, and/or dysregulated iron channel metabolism [18,19]. Similar to CRP, haptoglobin is considered a positive acute-phase reactant with its concentration elevated during inflammation [20]; however, comparatively, haptoglobin is less studied in COVID-19, with inconsistent findings on its correlation with severity of disease. Our study found elevated baseline haptoglobin values contributed to increased risk of severe disease, supported by previous findings [21]. On the other hand, other studies did not observe differences among varying COVID-19 disease severity [22] or even found it to be significantly lower in deceased patients than in COVID-19 survivors, which authors hypothesized to be caused by haptoglobin depletion from hemolysis [23]. Previous studies also found that risk of hospitalization and death increases with age [24,25]. VL is suggestive of active viral proliferation and is used to identify severe viral infections of the respiratory tract. One study found that SD cases have higher VL than mild to moderate cases [26], while another suggested that high VL was associated with increased risk of intubation and in-hospital mortality [27]. Many studies also found a direct relationship between older age and higher VL [28]. Appendix Section 6.1 includes discussion of other factors identified by ML modeling.

Nirmatrelvir/ritonavir is an antiviral that inhibits SARS-CoV-2 viral replication by targeting its main protease (Mpro, also known as 3C-like protease, 3CL), resulting in clinically reduced VL in patients in treatment of COVID-19 [29]. Although hypothesized in Appendix Section 6.2, the mechanism of action of COVID-19 in affecting serum levels of hsCRP and haptoglobin and the impact of Tx await to be further studied in order to establish causality in the observations made. In our study, only observational correlation between the prioritized baseline factors and SD were established. Causality needs to be further assessed by additional analyses and experiments.

Our study aimed to distinguish the most informative factors associated with increased risk of SD. The methodology was carefully designed in consideration of the high class imbalance in the EPIC-HR data and model generalizability. We cautiously focused the interpretation of factors prioritized in all nCV outer models, so results are persistent despite changes in model parameterization and more representative of the entire study population (Appendix Section 7). Furthermore, patients in the EPIC-HR study were enrolled at 343 sites across the Americas, Europe, Asia, and Africa [12], making the dataset representative of the majority of COVID-19–affected populations. RWD validation also showed generalizability of hsCRP, ferritin, and age relating to SD risk, even in a population with high COVID-19 vaccination rates [4], while VL and haptoglobin were unavailable in the RWD. Even though the RWD did not contain Tx information (Appendix Section 8), the validation still serves to strengthen the findings of the prioritized factors because those factors were at baseline when no treatment had been administered. Limitations of this study included the small sample sizes in both EPIC-HR and RWD datasets and the collection of most of their data from patients infected during the early part of the pandemic, before high rates of vaccination. Some of the population may not have been previously

infected or may have been infected with a different variant, in which case further data collection could improve generalizability. Finally, the RWD dataset lacked some of the measurements available in EPIC-HR (*e.g.*, symptom profile and VL). Because model performance improves with greater breadth and depth of available training data, this effectively capped the level the model could achieve. Additional input data, further validation of the model, and more in-depth interpretation of modeling results will advance the understanding of risk factors contributing to COVID-19 disease progression.

Machine learning modeling and interpretation supported by statistical analysis is a powerful tool for prioritizing factors that enable precision medicine decision making. Careful design of ML training framework and validation using big data is a good approach to strengthen evidence for generalizability of findings based on clinical trial data. Analysis showed receiving Tx was the greatest predictor and most important factor in reducing risk of progression to SD. Elevated baseline VL (> ~4 $\log_{10}$ copies/mL), hsCRP (> ~1 mg/dL), ferritin (> ~280 µg/L), haptoglobin (> ~210 mg/dL), and increased age (> ~48 years) were among the most important baseline factors contributing to increased risk of progressing to SD. Tx reduced VL and abnormal values of hsCRP and haptoglobin to greater extents than PBO at the measured time points. These results provide insights on multivariate signatures of COVID-19 progression, which may aid future studies and healthcare practices.

Supplementary data

The authors confirm that the supplementary data are available at *Biomedical Informatics* online.

Acknowledgment

This study was funded by Pfizer. The coauthors would like to thank all participants in the EPIC study. We also would like to acknowledge the contributions of Konstantinos Tsikkinis, Ioannis Setzis, and Kathy Xu for providing programming support and computing resources to this work. Thanks to Craig Hyde for providing valuable insights in guiding the modeling work. Editorial support was provided by Sheena Hunt, PhD, of ICON (Blue Bell, PA), and was funded by Pfizer.

Conflicts of interests

All authors are Pfizer employees and may hold stock or stock options.

Ethical statement

The study was performed in accordance with the Declaration of Helsinki and was approved by the Independent Ethics Committee and/or Institutional Review Board at each of the participating sites. Ethical conduct and responsibilities of the sponsor are detailed in the clinical trial registration (NCT04960202) and in the Supplementary Appendix of the primary study publication by Hammond *et al.* [12]. As this study is a secondary analysis of trial data, ethical considerations specific to the secondary use of the data are also adhered to as per the original study protocol.

Authors' contribution

Conceptualization, C.Z., J.H., and S.M.; methodology, C.Z., S.G., and J.H.; software, C.Z. and X.J.; validation, C.Z., X.J., S.G., S.T., and S.M.; formal analysis, C.Z., X.J., S.G., and J.H.; investigation, C.Z., J.H., and S.M.; resources, S.M.; data curation, C.Z., S.G., and J.H.; writing—original draft preparation, all authors; writing—review and editing, all authors; visualization, C.Z., S.G., and S.M.; supervision, C.Z., X.J., and S.M.; project administration, C.Z., X.J., and S.M.; funding acquisition, C.Z., J.H., and S.M. All authors have read and agreed to the published version of the manuscript.

References

[1] Cucinotta D, Vanelli M. WHO Declares COVID-19 a Pandemic. *Acta Biomed*. 2020, 91(1):157–160.

[2] Centers for Disease Control and Prevention. Long COVID or Post-COVID Conditions. Availab le at: https://www.cdc.gov/covid/long-term-effects/?CDC_AAref_Val=https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html. (accessed May 7, 2024).

[3] World Health Organization. WHO Coronavirus (COVID-19) Dashboard. Available at: https://covid19.who.int/. (accessed March 12, 2024).

[4] US Centers for Disease Control and Prevention. COVID Data Tracker. Available at: https://covid.cdc.gov/covid-data-tracker/#datatracker-home. (accessed December 5, 2024).

[5] Vetter P, Vu DL, L'Huillier AG, Schibler M, Kaiser L, *et al.* Clinical features of covid-19. *BMJ*. 2020, 369:m1470.

[6] Siordia JA, Jr. Epidemiology and clinical features of COVID-19: A review of current literature. *J. Clin. Virol.* 2020, 127:104357.

[7] Huang C, Wang Y, Li X, Ren L, Zhao J, *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020, 395(10223):497–506.

[8] World Health Organization. Weekly epidemiological update on COVID-19 - 22 March 2023. Available at: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---22-march-2023. (accessed January 10, 2024).

[9] Sanyaolu A, Okorie C, Marinkovic A, Patidar R, Younis K, *et al.* Comorbidity and its impact on patients with COVID-19. *SN Compr. Clin. Med.* 2020, 2(8):1069–1076.

[10] Wang B, Li R, Lu Z, Huang Y. Does comorbidity increase the risk of patients with COVID-19: evidence from meta-analysis. *Aging (Albany N. Y.)* 2020, 12(7):6049–6057.

[11] Biswas M, Rahaman S, Biswas TK, Haque Z, Ibrahim B. Association of sex, age, and comorbidities with mortality in COVID-19 patients: a systematic review and meta-analysis. *Intervirology.* 2020:1–12.

[12] Hammond J, Leister-Tebbe H, Gardner A, Abreu P, Bao W, *et al.* Oral nirmatrelvir for high-risk, nonhospitalized adults with COVID-19. *N. Engl. J. Med.* 2022, 386(15):1397–1408.

[13] Parvandeh S, Yeh HW, Paulus MP, McKinney BA. Consensus features nested cross-validation. *Bioinformatics*. 2020, 36(10):3093–3098.

[14] Centers for Disease Control and Prevention. CDC Museum COVID-19 Timeline. Available at: https://www.cdc.gov/museum/timeline/covid19.html. (accessed January 10, 2024).

[15] Ali N. Elevated level of C-reactive protein may be an early marker to predict risk for severity of COVID-19. *J. Med. Virol.* 2020, 92(11):2409–2411.

[16] Luan YY, Yin CH, Yao YM. Update advances on C-reactive protein in COVID-19 and other viral infections. *Front. Immunol.* 2021, 12:720363.

[17] Tan C, Huang Y, Shi F, Tan K, Ma Q, *et al.* C-reactive protein correlates with computed tomographic findings and predicts severe COVID-19 early. *J. Med. Virol.* 2020, 92(7):856–862.
[18] Cavezzi A, Troiani E, Corrao S. COVID-19: hemoglobin, iron, and hypoxia beyond inflammation. A narrative review. *Clin. Pract*. 2020, 10(2):1271.
[19] Vargas-Vargas M, Cortes-Rojo C. Ferritin levels and COVID-19. *Rev. Panam. Salud Publica*. 2020, 44:e72.
[20] Ceciliani F, Giordano A, Spagnolo V. The systemic reaction during inflammation: the acute-phase proteins. *Protein Pept. Lett*. 2002, 9(3):211–223.
[21] Beimdiek J, Janciauskiene S, Wrenger S, Volland S, Rozy A, *et al.* Plasma markers of COVID-19 severity: a pilot study. *Respir. Res.* 2022, 23(1):343.
[22] Chrostek L, Gan K, Kazberuk M, Kralisz M, Gruszewska E, *et al.* Acute-phase proteins as indicators of disease severity and mortality in COVID-19 patients. *Sci. Rep.* 2024, 14(1):20360.
[23] Yağcı S, Serin E, Acicbe Ö, Zeren Mİ, Odabaşı MS. The relationship between serum erythropoietin, hepcidin, and haptoglobin levels with disease severity and other biochemical values in patients with COVID-19. *Int. J. Lab. Hematol.* 2021, 43(suppl 1):142–151.
[24] Henkens MTHM, Raafs AG, Verdonschot JAJ, Linschoten M, van Smeden M, *et al.* Age is the main determinant of COVID-19 related in-hospital mortality with minimal impact of pre-existing comorbidities, a retrospective cohort study. *BMC Geriatr.* 2022, 22(1):184.
[25] Romero Starke K, Reissig D, Petereit-Haack G, Schmauder S, Nienhaus A, *et al.* The isolated effect of age on the risk of COVID-19 severe outcomes: a systematic review with meta-analysis. *BMJ Glob. Health.* 2021, 6(12):e006434.
[26] Layden JE, Ghinai I, Pray I, Kimball A, Layer M, *et al.* Pulmonary illness related to E-Cigarette use in Illinois and Wisconsin - final report. *N. Engl. J. Med.* 2020, 382(10):903–916.
[27] Magleby R, Westblade LF, Trzebucki A, Simon MS, Rajan M, *et al.* Impact of severe acute respiratory syndrome coronavirus 2 viral load on risk of intubation and mortality among hospitalized patients with coronavirus disease 2019. *Clin. Infect. Dis.* 2021, 73(11):e4197–e4205.
[28] Dadras O, Afsahi AM, Pashaei Z, Mojdeganlou H, Karimi A, *et al.* The relationship between COVID-19 viral load and disease severity: a systematic review. *Immun. Inflamm. Dis*. 2022, 10(3):e580.
[29] Toussi SS, Hammond JL, Gerstenberger BS, Anderson AS. Therapeutics for COVID-19. *Nat. Microbiol*. 2023, 8(5):771–786.