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

Supplementary materials

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.).

Table of Contents

1. COVID SD ML model development using EPIC-HR 4
1.1. Study cohort and class definition 4
1.2. Input features/factors and feature preprocessing 4
1.3. Nested cross-validation framework 7
1.4. Threshold optimization for the ML model 7
1.5. ML methods for addressing high class imbalance 8
2. Statistical analysis methods and results 8
2.1. Statistical methods 8
3. COVID SD ML model performance 10
3.1. Benchmark model selection 11
4. Supplementary results on treatment effect 12
4.1. Impact of treatment on prioritized factors 12
4.2. Impact of treatment on prioritized factors (not statistically significant) 12
4.3. Further analysis on impact of treatment on hsCRP and haptoglobin between non-SD and SD populations 13
5. Independent validation in RWD 14
5.1. RWD study cohort extraction 14
5.2. Alignment of EPIC-HR data to RWD 16
5.3. Additional results on Validation Strategy 1 16
5.4. Additional results on Validation Strategy 3 17
6. Supplementary discussion on prioritized factors 18
6.1. Further discussion of factors identified by ML modeling 18
6.2. Hypothesized mechanisms of action of Nirmatrelvir/ritonavir 19
7. Supplementary discussion on ML SD model methods 20
8. Supplementary discussion on RWD model validation 21
8.1. Differences of extracted RWD from EPIC-HR PBO data 21
8.2. Selection of independent dataset 22
9. Citations 22

List of Tables

Table S1: Class definition and prevalence based on EPIC-HR data.. 4

Table S2: Input feature/factor list based on EPIC-HR.. 4

Table S3: Sample size for each data point in Figure 3 line plots... 12

Table S4: Corresponding sample size for lines plots in Figure S4. 14

Table S5: Features available in the RWD study cohort and their corresponding missingness. ... 15

List of Figures

Figure S1: ML model performance.. 10

Figure S2: ML benchmark model selection.. 11

Figure S3: Impact of treatment on prioritized factors (not statistically significant)................... 12

Figure S4: Levels of hsCRP and haptoglobin between non-SD and SD subgroups with impact of treatment. 13

Figure S5: Strategy 1 validation ML model performance... 17

Figure S6: Strategy 3 validation ML model benchmark model selection................................. 18

1. COVID SD ML model development using EPIC-HR

1.1. Study cohort and class definition

Table S1. Class definition and prevalence based on EPIC-HR data.

(a) Class definition and prevalence across the entire study cohort extracted from EPIC-HR. (b) Class prevalence separated by PBO and Tx arms.

a)

	Definition	**n**	**Total, %**
Class 0	Participants who were not hospitalized or did not die through Day 28	2015	96.37
Class 1	Participants who were hospitalized or died before or on Day 28	76	3.63
Total	Entire study population	2091	100

b)

	Subgroup	**n**	**Arm, %**	**Total, %**
PBO	Class 0	987	93.73	47.20
	Class 1	66	6.27	3.16
	PBO total	1053	100	50.36
Tx	Class 0	1028	99.04	49.16
	Class 1	10	0.96	0.48
	Tx total	1038	100	49.64
Total	--	2091	--	100

1.2. Input features/factors and feature preprocessing

Features inputted to the ML model pipeline at the feature preprocessing stages are shown in Table S2 below, along with their corresponding feature type.

Table S2. Input feature/factor list based on EPIC-HR.

The table lists all input features studied in the SD ML model and their corresponding feature encoding method.

	Feature name	**Feature type**
1	COVID19 mAb Treatment (expected/received)	Binary
2	Viral RNA Level ($\log_{10}$ copies/mL)	Continuous
3	Chills or Shivering	Ordinal
4	Cough	Ordinal
5	Diarrhea	Ordinal
6	Feeling Hot or Feverish	Ordinal
7	Headache	Ordinal
8	Low Energy or Tiredness	Ordinal
9	Muscle or Body Aches	Ordinal

Table S2. *Cont.*

	Feature name	Feature type
10	Nausea	Ordinal
11	Sense of Smell	Ordinal
12	Sense of Taste	Ordinal
13	Shortness of Breath or Difficulty Breathing	Ordinal
14	Sore Throat	Ordinal
15	Stuffy or Runny Nose	Ordinal
16	Vomit	Ordinal
17	Age (years)	Continuous
18	Height (cm)	Continuous
19	Weight (kg)	Continuous
20	BMI (kg/m^2)	Continuous
21	Duration Since First Diagnosis (days)	Ordinal
22	Duration Since First Symptom (days)	Ordinal
23	Sex	Binary
24	Race	One-hot
25	Ethnicity	Binary
26	Country	One-hot
27	Treatment	Binary
28	Serology Status	Binary
29	Cardiovascular Risk	Binary
30	Chronic Kidney Risk	Binary
31	Chronic Lung Risk	Binary
32	Cigarette Smoker Risk	Binary
33	Hypertension Risk	Binary
34	Immunosuppression Risk	Binary
35	Device Dependence Risk	Binary
36	HIV Risk	Binary
37	Sickle Cell Risk	Binary
38	Neurodevelopmental Risk	Binary
39	Cancer Risk	Binary
40	Diabetes Risk	Binary
41	Number of Risk Factors	Ordinal
42	Age Group	Ordinal
43	Diastolic BP (mmHg)	Continuous
44	Oxygen Saturation (%)	Continuous
45	Pulse Rate (beats/min)	Continuous
46	Respiratory Rate (breaths/min)	Continuous
47	Sitting Diastolic BP (mmHg)	Continuous
48	Sitting Systolic BP (mmHg)	Continuous
49	Supine Diastolic BP (mmHg)	Continuous
50	Supine Systolic BP (mmHg)	Continuous
51	Systolic BP (mmHg)	Continuous
52	Temperature (Celsius)	Continuous
53	Heart Rate (beats/min)	Continuous
54	PR Interval (msec)	Continuous

Table S2. *Cont.*

	Feature name	Feature type
55	QRS Interval (msec)	Continuous
56	QT Interval (msec)	Continuous
57	QTcB Interval (msec)	Continuous
58	QTcF Interval (msec)	Continuous
59	RR Interval (msec)	Continuous
60	APTT (sec)	Continuous
61	ALT (U/L)	Continuous
62	Albumin (g/dL)	Continuous
63	ALP (U/L)	Continuous
64	AST (U/L)	Continuous
65	Basophils (10^9/L)	Continuous
66	Bicarbonate (mEq/L)	Continuous
67	Bilirubin (mg/dL)	Continuous
68	hsCRP (mg/dL)	Continuous
69	Calcium (mg/dL)	Continuous
70	Calcium Corrected (mg/dL)	Continuous
71	Chloride (mEq/L)	Continuous
72	CK (U/L)	Continuous
73	Creatinine (mg/dL)	Continuous
74	D-Dimer (ng/mL)	Continuous
75	Eosinophils (10^9/L)	Continuous
76	Erythrocytes (10^{12}/L)	Continuous
77	Ferritin (µg/L)	Continuous
78	Fibrinogen (mg/dL)	Continuous
79	eGFR (mL/min)	Continuous
80	Glucose (mg/dL)	Continuous
81	Haptoglobin (mg/dL)	Continuous
82	Hematocrit (%)	Continuous
83	Hemoglobin (g/dL)	Continuous
84	LDH (U/L)	Continuous
85	Leukocytes (10^9/L)	Continuous
86	Lymphocytes (10^9/L)	Continuous
87	Monocytes (10^9/L)	Continuous
88	Neutrophils (10^9/L)	Continuous
89	Platelets (10^9/L)	Continuous
90	Potassium (mEq/L)	Continuous
91	Procalcitonin (µg/L)	Continuous
92	Protein (g/dL)	Continuous
93	PT/INR	Continuous
94	PT (sec)	Continuous
95	Sodium (mEq/L)	Continuous
96	Thyrotropin (mIU/L)	Continuous
97	Thyroxine, Free (ng/dL)	Continuous
98	BUN (mg/dL)	Continuous
99	Treatment	Binary

1.3. Nested cross-validation framework

ML models were trained using the nested cross-validation (nCV) framework [1]. Specifically, five outer folds were split, with 80% as the training set and 20% as the test set, with each split stratified by class prevalence. Each training set from the outer folds was split again into five inner folds in the same manner to generate the inner training sets and the validation sets. Feature selection was performed on the training sets in each of the outer cross-validation (CV) folds. Features with more than 10% missing in the training set were removed. Subsequently, features were selected based on *P* values from hypothesis testing comparing Class 0 and Class 1 on each of the training sets, with the Welch *t* test and Wilcoxon rank sum test deployed for continuous variables and the Chi-squared test and Fisher exact test used for categorical variables. To be more permissive, features were included if one of the corresponding *P* values was ≤ 0.05. Feature/factor selection at this stage is lenient by design; the downstream ML nCV framework is to further prioritize the features selected by this first pass.

Missing values were imputed based on the training set within the nCV. As shown, missing values for continuous (*i.e.*, numerical) features were imputed by corresponding medians of the training set. Binary and ordinal features were imputed by their corresponding modes of the training set. Subsequently, binary features were one-hot encoded, with 1 designating presence of the condition and 0 absence of the condition. Ordinal features were not one-hot encoded to preserve the severity scoring associated with each numbered category.

Each of the above ML algorithms was trained within the inner CV folds, with the inner training set used for parameter optimization and the validation set for hyperparameter tuning via grid search, resulting in “inner models.” The inner model with the best F1 score on the validation set was then evaluated on the corresponding test set and chosen as the “outer model.” This process resulted in five “outer models” from the five outer folds for each ML algorithm tested. After each ML algorithm was evaluated in the nCV framework, one ML algorithm was selected based on the least average overfits and best average F1 from all five “outer models.” The chosen algorithm from nCV was subsequently retrained on the full dataset, with the input features set and hyperparameters carried over from the corresponding least overfitted “outer model.” This single retrained model is referred to as the “benchmark model,” which was assumed to be of comparable if not superior performance to the average of outer models without inadequately overfitting the training sets.

1.4. Threshold optimization for the ML model

The following four classification thresholds were tried out for evaluation of outer models based on inner training and validation sets:

(1) Default 0.5

(2) Threshold at intersection between precision & recall

(3) Threshold that maximizes geometric mean of TPR & (1 − FPR) (*i.e.*, sensitivity & specificity)

(4) Threshold that maximizes F1 macro

Option 3 above achieved better performance on the validation tests and thus was chosen to be the method incorporated in model evaluation on the test set, that is, used in calculation of threshold-dependent metrics, such as the F1 score, accuracy, precision, and recall.

1.5. ML methods for addressing high class imbalance

To optimize ML models against the high class imbalances (*i.e.*, prevalence of Class 1 is ~3.5%) in situations of relatively low sample size, performance metrics for model selection emphasized the F1 score, which is the harmonic mean of precision and recall. We also focused more on optimizing ensemble tree-based algorithms, such as RF, BRF, LGB, and XGB, which in general excel in imbalanced datasets because their hierarchical structure allows them to learn from both classes. Because these methods tend to overfit to training samples, we also prioritized overfit on F1 score as a key evaluation metric to increase potential for generalization of results. Hyperparameters that constrain tree structures, such as tree depth and minimum samples per split, were also tuned in grid search. Furthermore, for hyperparameter tuning, inverse class weighting was deployed to give more weight to the training samples from the minority Class 1 during the optimization process, which penalized misclassification of Class 1 by an amount proportional to the level of underrepresentation.

Up-sampling of the minority class was initially attempted but did not result in marked performance improvement. This is believed to be due to the combined effect of small sample size, low temporal resolution of the data, and relatively heterogeneous participant characteristics. In addition, generation of synthetic data was not considered desirable in this case and thus was not pursued further. Down-sampling of the majority class was also tested without marked performance improvements observed. With eliminating valuable data perceived as undesirable, this method was also disregarded in early stages. Deep learning methods were not explored because they were not well suited for the small sample size, sparsity of data, and in-depth level of interpretation required in this study.

2. Statistical analysis methods and results

2.1. Statistical methods

The following steps were followed to produce the line plots shown in Figures 3 and S3.

1. Select features based on ML model outputs.
   Only analyze prioritized features from preceding analysis.
   Only select features that can potentially be modified by treatment.
   i.e., Treatment, age, and age group cannot be modified, so eliminate from analysis.

2. Select participants who have abnormal lab values at baseline.
   -BNRLO: baseline normal range lower limit
   -BNRHI: baseline normal range upper limit

   For the following features, select subjects with
   Value < BNRLO
   eGFR
   For the rest, select subjects with

Value > BNRHI

Exception:

VL does not have a normal range, so values from all participants were taken.

3. Perform normality test on lab values of selected participants from Step 2.

Normality test is only done on data at baseline, on combined data from both Tx and PBO arms.

Exception:

VL is already in $\log_{10}$, does not need to be transformed (*i.e.*, does not need normality test).

Calculate skewness from linear scale:

(1) If skewness falls within –1 and 1, then no transformation needed.

(2) If skewness is outside of –1, 1 range, then log transform the data and check skewness of the log transformed data.

(3) If skewness of the log-transformed data is closer to 0 than the untransformed skewness, then calculate mean and CI using log-transformed data.

(4) If log transformation did not improve skewness (decrease in absolute values of skewness), then calculate mean and CI using untransformed data.

4. Calculate:

a. Mean

b. 95% CI

For each arm (*i.e.*, PBO & Tx separately), for each time point from original or transformed dataset depending on outcome of normality test (*i.e.*, Step 3).

5. Plotting in linear space for interpretation.

Transform back into linear space if mean and 95% CI of the mean were obtained from log-transformed data per normality test.

6. Hypothesis test to evaluate changes from baseline for selected labs between PBO and Tx arms.

(1) Determine whether to use log transformation or not. Transform all data if necessary (see Step 3, normality test).

(2) Calculate change from baseline.

(3) Use two-sided two-sample *t* test to compare Tx vs PBO using change from baseline data at each individual day separately (*e.g.*, Day 5, Day 3, Day 14, as data availability permits).

(4) Overlay the *P* value and significance level on the mean and 95 CI line plot on Day 5 and Day 14 visits.

Subgroup analysis for line plots for ferritin, eGFR, CK

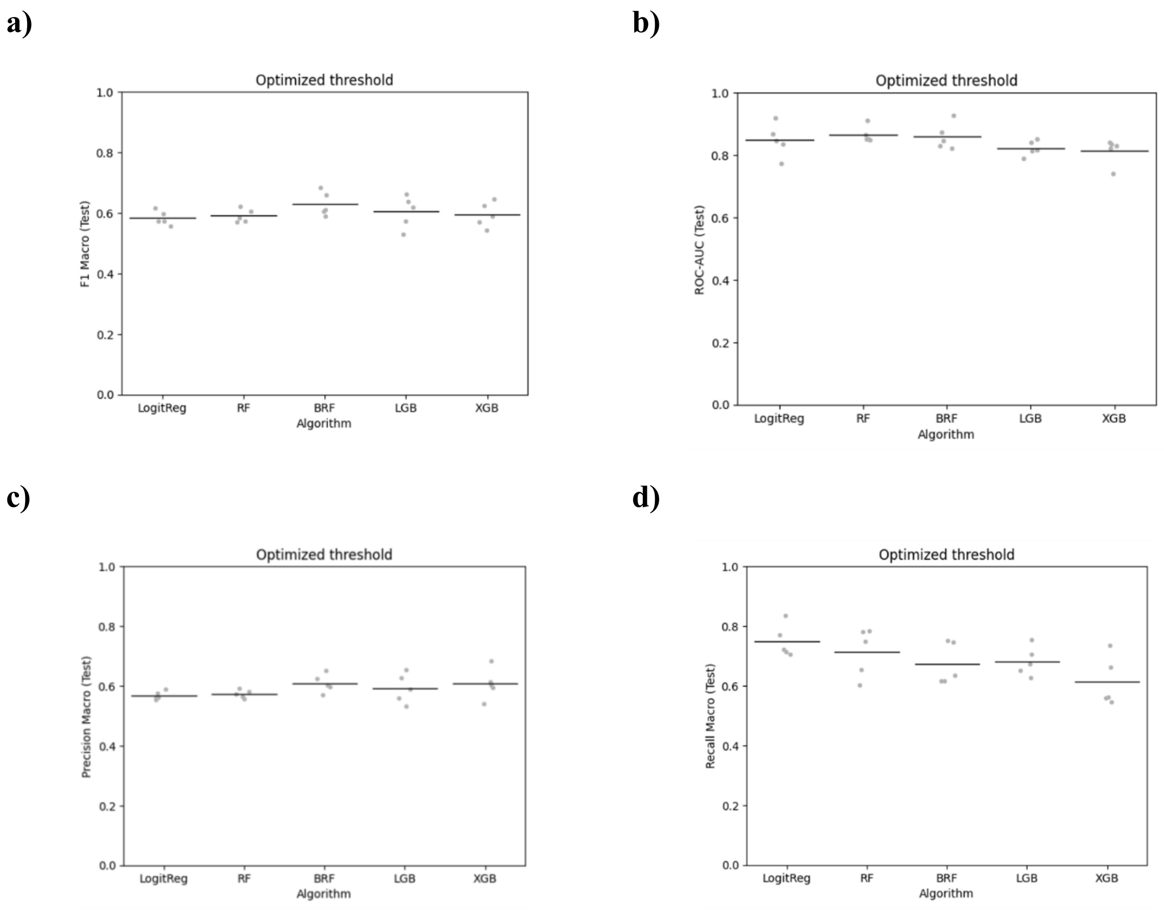
Perform statistical analysis specified above for the following subgroups (subgroups are identified by participants with designated thresholds defined by the laboratory specification of the trial):

BNRHI for ferritin, CK

BNRLI for eGFR

1. Ferritin
    a. Participants with threshold of 400 µg/L
    b. Participants with threshold of 291 µg/L
2. eGFR
    a. Participants with threshold of 85 mL/min/1.73 m^2
    b. Participants with threshold of 75 mL/min/1.73 m^2
3. CK
    a. Participants with threshold of 207 U/L
    b. Participants with threshold of 169 U/L

3. COVID SD ML model performance



e)

	ROC_AUC	**F1 score**	**Precision**	**Recall**
Average	**0.859**	**0.630**	**0.609**	**0.673**
MAD	**0.032**	**0.034**	**0.022**	**0.060**

Figure S1. ML model performance. Test **(a)** F1 score, **(b)** ROC-AUC, **(c)** precision, and **(d)** recall are shown for the five algorithms tested in the nCV framework. Grey dots denote individual model performance, and the horizonal bar denotes the average across the five outer models. **(e)** Average (first row) test performance and mean absolute deviation (MAD) (second row) for each of the performance metrics reported.

The performances of the outer models for each algorithm were evaluated on held-out test sets. The metrics were macro F1 score, area under the receiver operating characteristic curve (ROC-AUC), precision, and recall, as shown in Figures 1a–d, respectively. ROC-AUC provides an aggregate measure of performance across all possible classification thresholds. The F1 score is the harmonic mean of precision and recall. All metrics range from 0 to 1, and the higher these scores are, the better performing the ML model is.

BRF had the highest average F1 scores (~0.65), with the other algorithms closely following (Figure S1a). BRF, along with RF, also achieved the highest average AUC (~0.85; Figure S1b). Precision for BRF was slightly higher than the rest (~0.6), but with a lower recall at ~0.65, which is not surprising due to precision-recall trade off. As a result, BRF was selected as the benchmark algorithm. Note that optimized threshold is the threshold that maximizes geometric mean of TPR & (1 − FPR) (*i.e.*, sensitivity & specificity) on inner training and validation sets (See Appendix Section 1.4).

3.1. Benchmark model selection

BRF was determined to be the best algorithm out of the five algorithms tested based on average performance across the five outer models. The overfit for each of the five BRF outer models, as defined by test F1 minus training F1, was further investigated for selection of the benchmark model (Figure S2). As shown, Fold 1 and 2 are the least overfitted; we then selected Fold 2 arbitrarily. As a result, the BFR outer model from Fold 2 was retrained on the full EPIC-HR dataset with its hyperparameters and input features carried over.

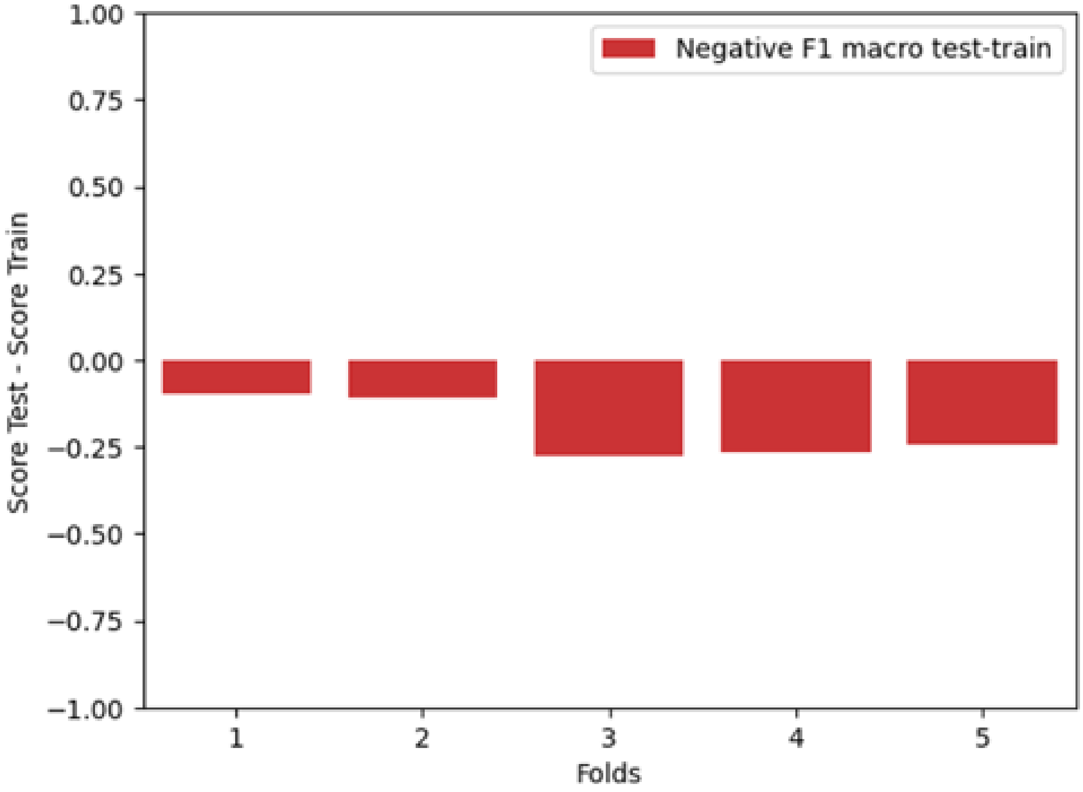


Figure S2. ML benchmark model selection.

4. Supplementary results on treatment effect

4.1. Impact of treatment on prioritized factors

Table S3. Sample size for each data point in Figure 3 line plots.

Samples were obtained after selection for abnormal lab values. (a) hsCRP; (b) haptoglobin.

a)

		n
Arm	**Day**	
PBO	**BL**	**543**
	Day 5	**476**
	Day 14	**398**
Tx	**BL**	**530**
	Day 5	**469**
	Day 14	**395**

b)

		n
Arm	**Day**	
PBO	**BL**	**537**
	Day 5	**468**
	Day 14	**376**
Tx	**BL**	**506**
	Day 5	**451**
	Day 14	**356**

4.2. Impact of treatment on prioritized factors (not statistically significant)

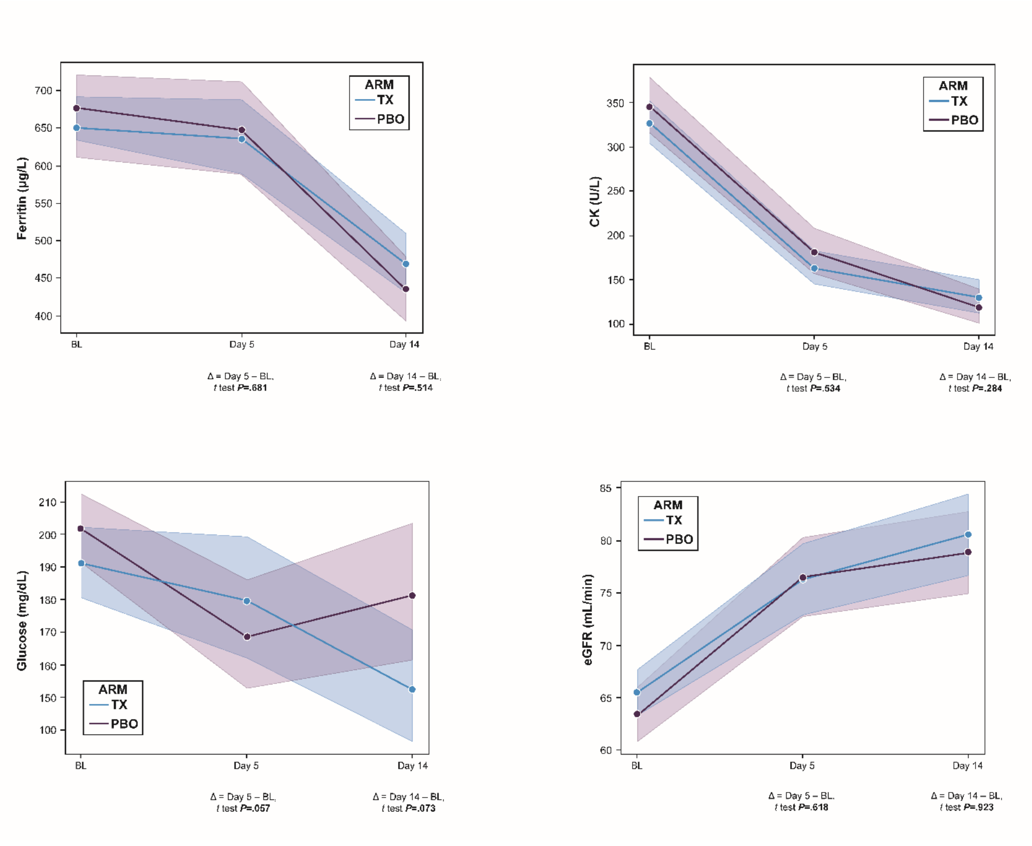


Figure S3. Impact of treatment on prioritized factors (not statistically significant). Lines plots for abnormal (a) ferritin (> 400 ng/mL males, > 291 ng/mL females), (b) CK (> 207 U/L males, > 169 U/M females), (c) glucose (> 138 mg/dL), and (d) eGFR (< 85 mL/min/1.73 m^2 males, < 75 mL/min/1.73 m^2 females) are shown. Blue indicates the Tx arm and black the PBO arm. Means of abnormal lab values are shown by solid lines with 95% CIs in shades. The dot on the line indicates the time point at which the data were measured. *P* values from two-sample *t* test on means of change from baseline between the PBO and Tx groups are shown at the bottom of the plot for each subsequent day measured.

4.3. Further analysis on impact of treatment on hsCRP and haptoglobin between non-SD and SD populations

Line plots (Figure S4) were generated according to the method specified in Appendix 2.1. As shown in Figure S4a and b, in participants who did not progress to severe COVID-19 disease (SD), hsCRP and haptoglobin decreased with time. Furthermore, treatment (blue) decreased their levels to greater extents compared with PBO (black) with statistical significance on measured days. By contrast, in SD patients, there was no statistical difference of hsCRP and haptoglobin levels between PBO and Tx groups (Figure S4c and d); hsCRP and haptoglobin remained high. Compared with PBO (black), treatment (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 (Figure S4c and d). In addition, at baseline, hsCRP and haptoglobin levels were higher in the SD cohort (Figure S4c and d) than in the non-SD cohort (Figure S4a and b), consistent with results from ML modeling.

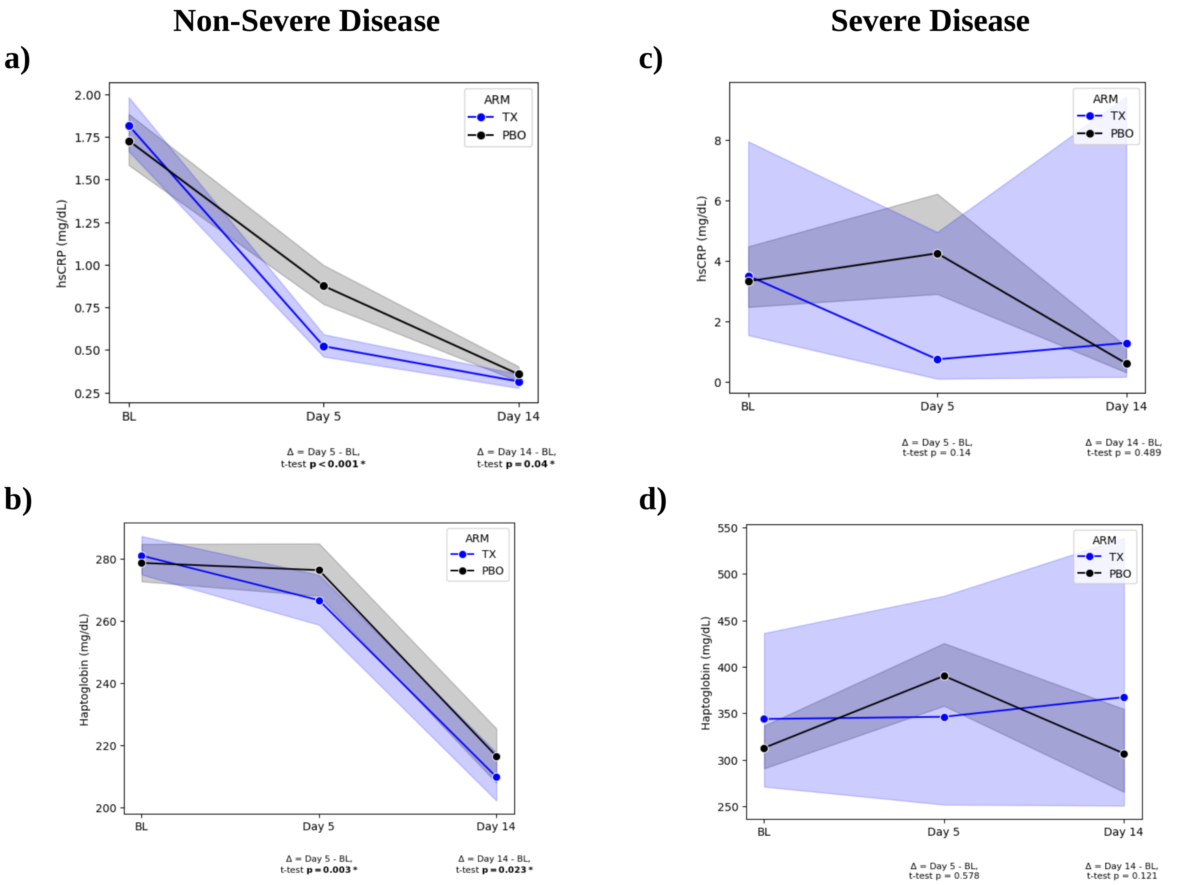
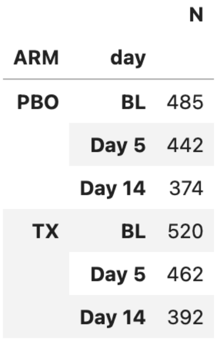


Figure S4. Levels of hsCRP and haptoglobin between non-SD and SD subgroups with impact of treatment. Line plots are shown for participants of non-SD (first column) and SD (second column) for hsCRP (a, c) and haptoglobin (b, d). Blue indicates the Tx arm and black the PBO arm. Only participants with abnormal hsCRP (> 0.5 mg/dL) and haptoglobin (> 200 mg/dL) at baseline were 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* test 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.

Table S4. Corresponding sample sizes for lines plots in Figure S4.

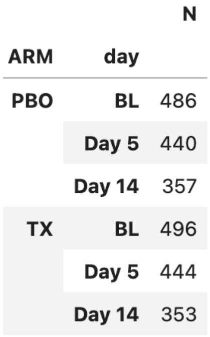
Non-Severe Disease

a) hsCRP



ARM	day	N
PBO	BL	485
	Day 5	442
	Day 14	374
TX	BL	520
	Day 5	462
	Day 14	392

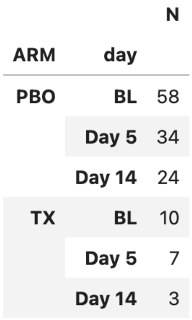
b) Haptoglobin



ARM	day	N
PBO	BL	486
	Day 5	440
	Day 14	357
TX	BL	496
	Day 5	444
	Day 14	353

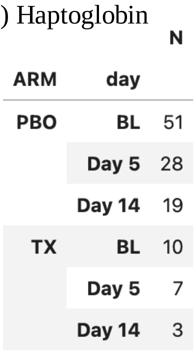
Severe Disease

c) hsCRP



ARM	day	N
PBO	BL	58
	Day 5	34
	Day 14	24
TX	BL	10
	Day 5	7
	Day 14	3

d) Haptoglobin



ARM	day	N
PBO	BL	51
	Day 5	28
	Day 14	19
TX	BL	10
	Day 5	7
	Day 14	3

5. Independent validation in RWD

5.1. RWD study cohort extraction

The following considerations were taken during Optum-EHR cohort extraction. Measurements were taken as the value closest to the COVID-19 index date (0) within the following defined time windows.

- Height, weight, BMI: [–90, 0] days
- Vitals: [–28, 0] days
- Labs: [–28, 0] days

Presence of symptoms and comorbidities were noted if any conditions arose within the following time windows relative to COVID-19 index date (0).

- Symptoms [–10, 0] days
- Comorbidities
  - [–360, 0] days for most comorbidities
  - [–inf, 0] for long-term/chronic conditions (*e.g.*, cancer, neuro).

In the final step, participants with too much missing data were dropped according to the following rules: (1) only kept patients with at least one of hsCRP, haptoglobin, or ferritin measured; (2) discarded features that were missing for more than 50% of the remaining patients; (3) discarded patients who were missing more than 10% of the remaining features; (4) selected the common feature set between EPIC-HR and RWD for both datasets. Before Step 1, all lab measures available were already at least 88% missing, so selection for any lab measures would have reduced sample size by at least 88%. Haptoglobin had to be eliminated in Step 2 among a total of 11 features that were missing in more than 50% of the patients. Specifically, the 11 features eliminated in Step 2 were haptoglobin (mg/dL), fibrinogen (mg/dL), thyroxine-free (ng/dL), procalcitonin (µg/L), APTT (sec), bicarbonate (mEq/L), CK (U/L), LDH (U/L), PT (sec), PT/INR, and D-Dimer (ng/mL). Original eGFR data were sparse and often included misspelled/ambiguous units or missing units altogether; consequently, eGFR was calculated based on the CKD-EPI equation [2]. In the end, the RWD cohort contained 946 patients with 17.8% SD (Class 1). The input features in the final RWD design matrix are listed in the following table, together with their corresponding missingness as a percentage of cohort size.

Table S5. Features available in the RWD study cohort and their corresponding missingness.

	Feature	**Missing values**	**Missing, %**
0	Duration Since First Symptom (days)	453	47.9
1	**Ferritin (µg/L)**	**376**	**39.7**
2	**hsCRP (mg/dL)**	**316**	**33.4**
3	Respiratory Rate (breaths/min)	309	32.7
4	Ethnicity	188	19.9
5	Temperature (Celsius)	152	16.1
6	Pulse Rate (beats/min)	82	8.7
7	eGFR (mL/min)	67	7.1
8	Basophils (10^9/L)	66	7.0
9	Race	65	6.9
10	Eosinophils (10^9/L)	64	6.8
11	Height (cm)	58	6.1
12	Monocytes (10^9/L)	55	5.8
13	Lymphocytes (10^9/L)	54	5.7
14	Neutrophils (10^9/L)	54	5.7
15	Sitting Systolic BP (mmHg)	43	4.5
16	Sitting Diastolic BP (mmHg)	43	4.5
17	Bilirubin (mg/dL)	42	4.4
18	Protein (g/dL)	38	4.0
19	ALP (U/L)	37	3.9
20	Albumin (g/dL)	36	3.8
21	AST (U/L)	36	3.8
22	ALT (U/L)	34	3.6
23	BMI (kg/m^2)	32	3.4
24	Weight (kg)	29	3.1
25	Glucose (mg/dL)	8	0.8
26	Calcium (mg/dL)	6	0.6
27	Hematocrit (%)	5	0.5
28	Platelets (10^9/L)	5	0.5
29	Hemoglobin (g/dL)	4	0.4
30	Sex	2	0.2
31	Diarrhea	0	0.0
32	Feeling Hot or Feverish	0	0.0
33	Creatinine (mg/dL)	0	0.0
34	Chloride (mEq/L)	0	0.0
35	Potassium (mEq/L)	0	0.0
36	Sodium (mEq/L)	0	0.0
37	BUN (mg/dL)	0	0.0
38	Cough	0	0.0
39	Sense of Taste	0	0.0
40	Sense of Smell	0	0.0
41	TARGET	0	0.0
42	Nausea	0	0.0

Table S5. *Cont.*

	Feature	**Missing values**	**Missing, %**
43	Headache	0	0.0
44	Shortness of Breath or Difficulty Breathing	0	0.0
45	Cigarette Smoker Risk	0	0.0
46	Sore Throat	0	0.0
47	Stuffy or Runny Nose	0	0.0
48	Vomit	0	0.0
49	**Age (years)**	**0**	**0.0**
50	Muscle or Body Aches	0	0.0
51	Cardiovascular Risk	0	0.0
52	Chronic Kidney Risk	0	0.0
53	Chronic Lung Risk	0	0.0
54	Hypertension Risk	0	0.0
55	Low Energy or Tiredness	0	0.0
56	Immunosuppression Risk	0	0.0
57	Device Dependence Risk	0	0.0
58	HIV Risk	0	0.0
59	Sickle Cell Risk	0	0.0
60	Neurodevelopmental Risk	0	0.0
61	Cancer Risk	0	0.0
62	Diabetes Risk	0	0.0
63	Chills or Shivering	0	0.0
64	SUBJID	0	0.0

5.2. Alignment of EPIC-HR data to RWD

For validation purposes (Validation Strategy 1: Using EPIC-HR PBO data to train the model via nCV and evaluate the resulting model using RWD), only features in EPIC-HR that were also found in the RWD design matrix were kept. In addition, since RWD only have symptom presence but not severity gradings, EPIC-HR symptom data were converted to binary variables, with Grade 0 encoded as 0 and higher numerical grades encoded as 1.

5.3. Additional results on Validation Strategy 1

Results are shown below for Validation Strategy 1: Using EPIC-HR PBO data to train the model via nCV and evaluate the resulting model using RWD. Figure S5 shows test performance from the nCV on EPIC-HR data. BRF has a slightly higher F1, but RF and BRF performed almost the same. This is expected because EPIC-HR PBO is a more balanced dataset in class prevalence than the entire EPIC-HR with PBO and Tx arms combined (Table S1).

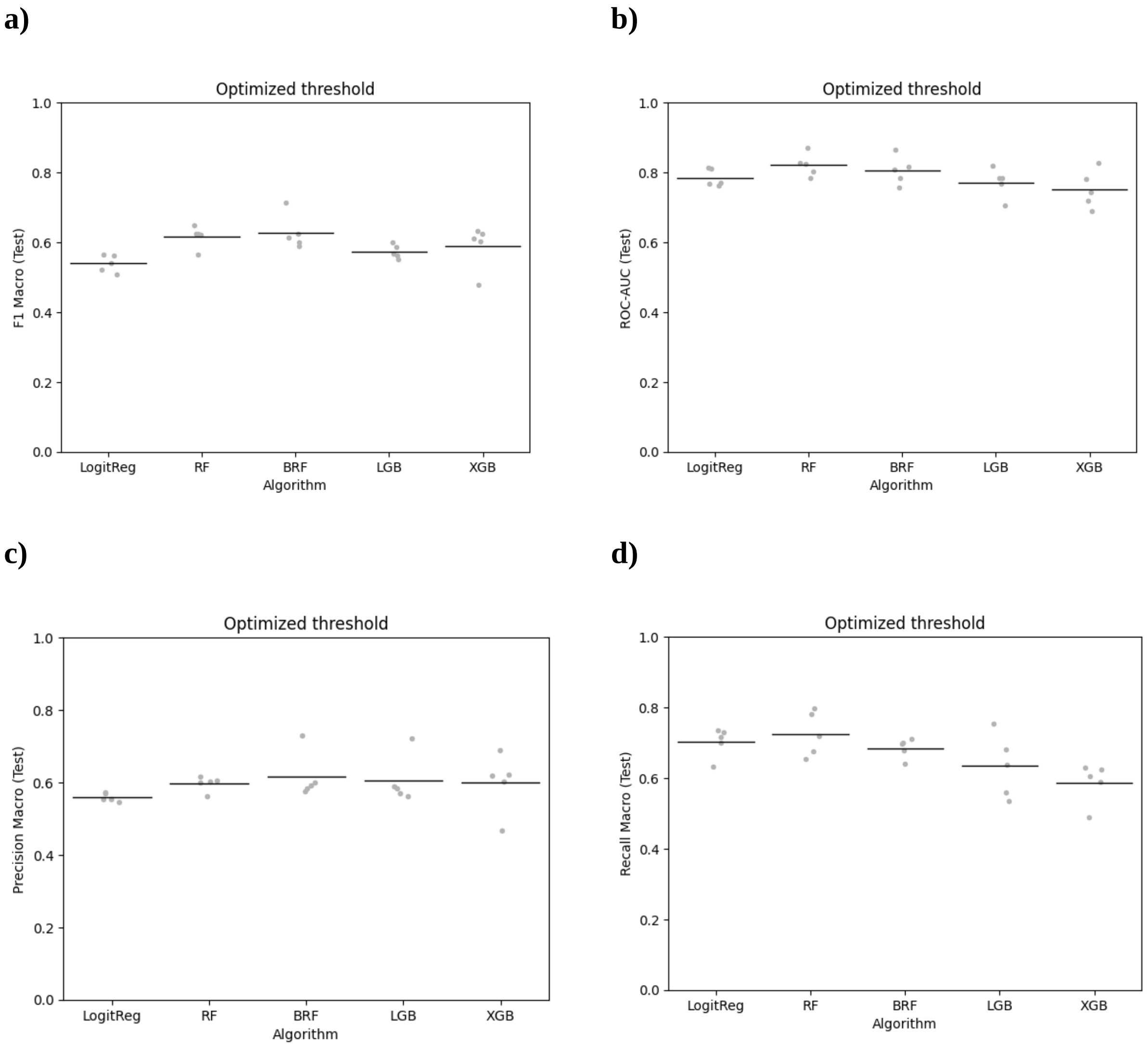


Figure S5. Strategy 1 validation ML model performance. Performances of ML model developed using EPIC-HR PBO data via nCV, specifically test performances on EPIC-HR PBO test data held out in the nCV process. (a) F1 score, (b) ROC-AUC, (c) precision, and (d) recall are shown for the five algorithms tested in the nCV framework. Grey dots denote individual model performance, and the horizonal bar denotes the average across the five outer models.

5.4. Additional results on Validation Strategy 3

Results are shown below for Validation Strategy 3, using RWD in nCV to check if previously prioritized features in EPIC-HR are also prioritized in the RWD outer models.

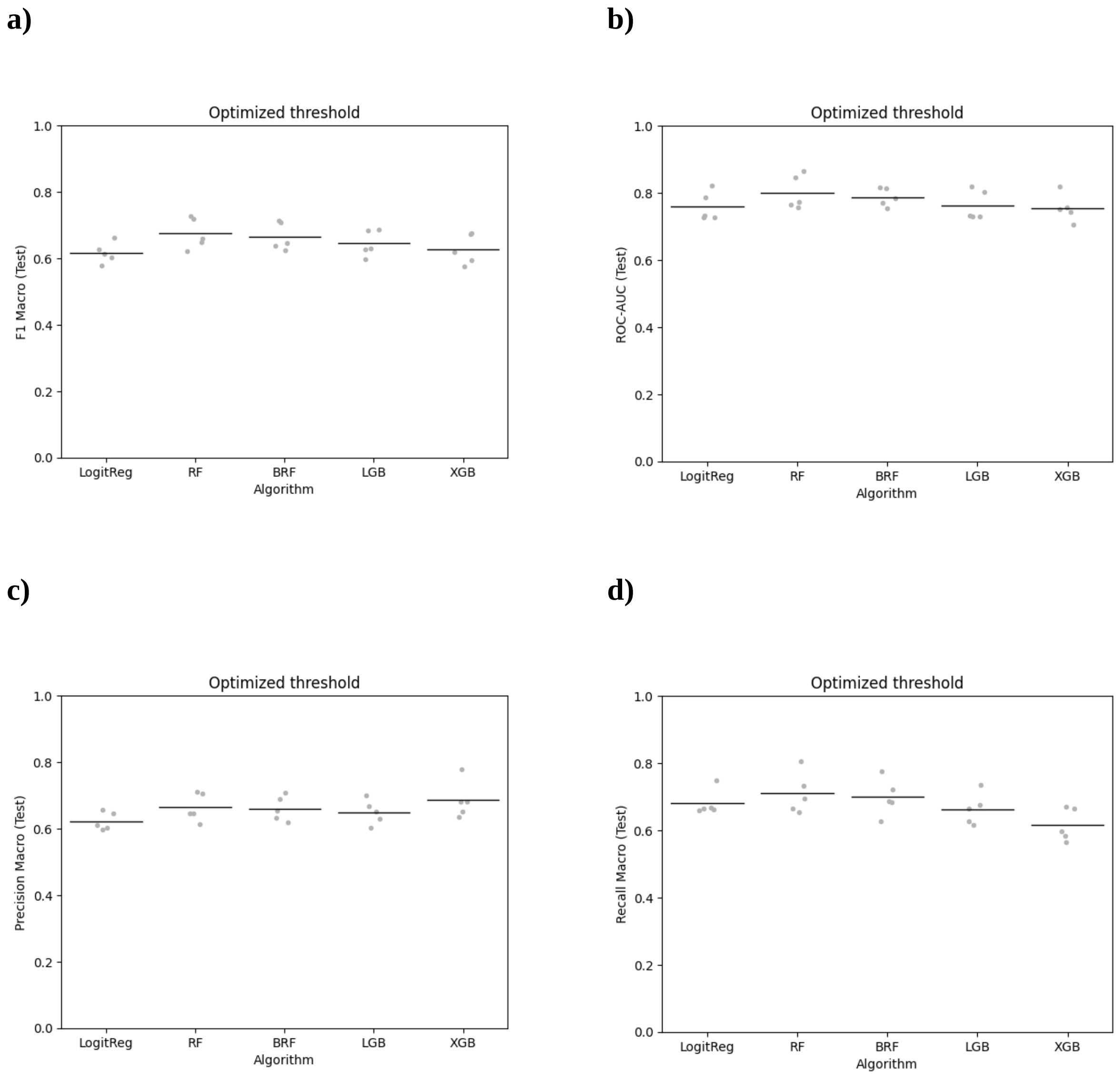


Figure S6. Strategy 3 validation ML model benchmark model selection. Performances of ML model developed using extracted RWD via nCV, specifically test performance on RWD test set held out during the nCV process. Test (a) F1 score, (b) ROC-AUC, (c) precision, and (d) recall are shown for the five algorithms tested in the nCV framework. Grey dots denote individual model performance, and the horizonal bar denotes the average across the five outer models.

6. Supplementary discussion on prioritized factors

6.1. Further discussion of factors identified by ML modeling

Prior literature on features ranked lower in priority and not consistently observed in all of the nCV folds is discussed below. These prior studies also serve to provide evidence for the validity of ML analysis in this work. Multiple studies have found that elevated CK was associated with a more severe COVID-19 prognosis [3,4] and could be the sole initial presentation in patients with COVID-19 [5]. It was believed that the virus may directly invade muscles and the nervous system via the same immune-mediated pathway, causing muscle injury in these patients [4]. Reduced eGFR is a marker of renal dysfunction and acute kidney injury (AKI) in critically ill patients with COVID-19 [6]. AKI induced by COVID-19 was proposed to be caused by the interaction of several mechanisms, including dehydration causing impaired renal blood flow, coagulation activation with microthrombotization, and immune activation of inflammatory neutrophil polymorphonuclear cells [6].

A previous study also found that elevated plasma fibrinogen was associated with severe COVID-19. Fibrinogen levels were elevated in patients with COVID-19 at admission, especially in those with severe disease, and sometimes warranted ICU admission [7]. Fibrinogen was reported to be a marker of inflamation and was associated with elevated IL-6 levels [7]. Low serum calcium levels have been associated with more severe COVID-19 in early stages of illness [8] and have been observed more frequently in those with COVID-19 than a control group [9]. Changes in intracellular calcium homeostasis are believed to promote the activation of inflammatory pathways, leading to the increase in inflammatory factors such as IL-1β, tumor necrosis factor (TNF), and IL-6 [8]. Low calcium levels have been associated with cytokine storm [8]. Vital signs, such as temperature, are commonly regarded as essential signs predictive of clinical deterioration [10]. Elevated LDH is associated with severe *versus* mild COVID, ICU admission *versus* non-admission, and non-survival *versus* survival [11,12]. It is a potential marker of vascular permeability in immune-mediated lung injury and various inflammatory states (*e.g.*, infection, malignancies, MI, sepsis, cardiopulmonary compromise) [11,12]. Procalcitonin was believed to be an indicator of disease severity [13]. Hypertension has been identified as the most prevalent cardiovascular comorbidity in patients infected with COVID-19 and is associated with increased risk of hospitalization and death [14]. Liver enzymes, ALP, ALT, and AST in patients diagnosed with COVID-19 were not found to be signficantly affected by COVID-19 in a meta-analysis of exisiting literature [15].

Serology status and glucose were also among the top ten most important identified factors. They are also characteristics of the entire population because they appeared in all but one of the outer models in the nCV. On average, glucose contributed a smaller extent to SD risk than serology status, barely making the top ten. Positive serology status and IgG and IgM antibody values have been found to be significantly associated with COVID-19–related symptoms, such as cough, fever, and chills [16], and is believed to be a useful marker of COVID-19 severity [17]. Furthermore, studies have found that severe COVID-19 was associated with higher blood glucose [18]. It was proposed that glucose-induced metabolic reviewing potentiates SARS-Cov-2 replication and cytokine production [19]. In metabolically stressed monocytes cultured under high glucose conditions, SARS-CoV-2 infection enhances glycolysis by aberrant production of mitochondrial ROS, which leads to activation of transcription factor hypoxia-inducible factor-1α (HIF-1α) to ensure its rapid replication [19].

6.2. Hypothesized mechanisms of action of nirmatrelvir/ritonavir

The following describes our hypothesis on mechanisms of enhanced reduction of CRP and haptoglobin levels in treated COVID-19 patients, which needs to be further validated. Nirmatrelvir/ritonavir inhibits SARS-CoV-2 viral replication by targeting its main protease (Mpro, also known as 3C-like protease, 3CL), resulting in reduced viral load (VL) in patients receiving treatment for COVID-19 [20]. Serum CRP is closely correlated with COVID-19 disease progression [21] and has been observed to be elevated in COVID-19, including severe COVID-19 [21–24]. Furthermore, it was shown 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 [21,24]. SARS-CoV-2 infected cells, together with free virus, activate proinflammatory mediators (*e.g.*, cytokines IL-6 and TNF-a, among others) of the innate and adaptive immune system [25,26]. Besides engagement of anti-inflammatory mediators, which contribute to resolving the proinflammatory response, the proinflammatory response also causes the accumulation of alveolar cell damage due to the inflammatory death of infected and bystander alveolar cells [27,28]. This will further activate the immune response in a positive feedback loop [27,28], where a cytokine response storm (CRS) can be triggered [25,26]. This uncontrolled release of cytokines stimulates hepatocytes to produce CRP, which is then released into systemic circulation [21,27]. Indeed, the main cause of critical illness and death in patients with COVID-19 is considered to be excessive inflammation, in which serum CRP levels are markedly increased [21]. As VL is decreased by nirmatrelvir/ritonavir in treating COVID-19, this release of CRS is believed to be ablated and inflammation controlled, thereby decreasing CRP levels in the process.

Similar to CRP, haptoglobin is considered a positive acute-phase reactant with its concentration elevated during inflammation [29]. However, comparatively, haptoglobin is less studied in COVID-19, with inconsistent findings on its correlation with severity of disease [29]. Our study found elevated baseline haptoglobin values contributed to increased risk of severe disease. Supporting this finding, a previous study has found that haptoglobin levels were higher in COVID-19 patients compared with controls in a statistically significant manner [22]. Additionally, albeit without statistical significance, haptoglobin was observed to increase in a COVID-19 severity-dependent manner [22]. On the other hand, other studies did not observe differences among varying COVID disease severity [30] 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 [31]. The understanding of the role of haptoglobin in COVID-19 is incomplete, but possible hypothesis can be formulated from studies of (patho-)physiology and other diseases. In contrast to the sharp rise in CRP upon inflammatory stimulus, the increase in haptoglobin is far less in magnitude and more gradual [32]. In reversing the course of COVID-19, nirmatrelvir/ritonavir may have ablated inflamation-induced elevation of haptoglobin. Nonetheless, the above stated hypothesis needs to be validated, the role and mechanism of action of haptoglobin in COVID-19 remains to be clarified, and results should be interpreted in the clinical context of the individual.

7. Supplementary discussion on ML SD model methods

We err on the side of caution by focusing more on markers prioritized in all of the outer models in the nCV validation. The more frequently a factor appears in the nCV folds, the more representative of the entire study population it is believed to be because it is less perturbed by the divisions of training and testing sets during the ML model training process. It may also be interpreted that factors of this nature are able to reduce the consequences of certain unintended bias arising from the chosen model training and selection methodology. However, this is not to say that the factors that were prioritized only by some rather than all of the outer models are not informative. In fact, it is possible that they signify characteristics of a particular patient segment within the overall study population. Indeed, by trial design, patients with different baseline risks factors were included with different comorbidities, concomitant medications, and clinical profiles [33]. On the other hand, these results need to be interpreted with caution. Patient segmentation has not been studied in this paper because further division of the study population, which was already limited in sample size, together with the rare occurrence of the endpoint would likely lead to further deterioration of model performance and reduction in power for feature prioritization. This would render any subsequent result inconclusive.

Generalizability of the model and results was partially evaluated via the nCV framework, which effectively simulated potential heterogeneity in study cohort by training and evaluating the model in different random stratified splits of the data. Furthermore, the test sets in the outer folds were entirely unseen by the model training process, including the feature selection process. Model performances were reported from all CV folds, with a range of possible performances rather than only focusing on a single held-out test set. Even though a “final benchmark model” was selected for interpretation of factors on individual patients, repeatability of top factors was evaluated by frequency of factors appearing in top ten features across all outer models. Because the “final benchmark model” was trained on data from the entire study cohort, not just an 80% split, it is expected to perform better than the average performance of the “outer models” of the corresponding algorithms on unseen datasets. Additional ML optimization techniques and training framework can be tested for possible improvements of model performance. Further validation of modeling results can also be obtained by deeper interpretation of underlying pathophysiology. For instance, by moving across biological scales, one can observe if signaling pathways regulating certain clinical laboratory markers are also changed by the disease through proteomics analysis. Subsequently, more in-depth interpretation of modeling results will further the understanding of COVID-19.

Our study was targeted for factor prioritization for the most informative factors most frequently associated with SD in treated and untreated groups. The models may not be feasible for use in broader clinical decision support during the care process and would require additional validation. This is because some of the data used in modeling are unique to the clinical trial environment. Laboratory data, such as CRP, ferritin, or haptoglobin, are informative factors for SD risk but may not be available for most patients at the time of treatment decisions. A future analysis could build ML models with data based only on criteria available before a patient would be prescribed any treatment medication. For instance, a model could be constructed using only demographic data, the 14-test comprehensive metabolic panel, and concomitant medications. Conversely, the performance of the ML models will likely degrade because of the absence of highly informative clinical laboratory data that are typically available only in a clinical trial environment.

8. Supplementary discussion on RWD model validation

8.1. Differences of extracted RWD from EPIC-HR PBO data

Overall, the RWD captured the Delta dominant time frame and most features measured in the EPIC-HR. The extracted RWD study cohort data have no nirmatrelvir/ritonavir data, so validation was constrained to the PBO arm only. This was acceptable for validating risk factors prioritized by the SD model because all risk factors studied were at baseline when no treatment had yet been administered. However, it was not possible to observe the treatment effect of nirmatrelvir/ritonavir on the risk factors in the RWD. Additionally, the RWD study cohort data nonetheless were notably different from the EPIC-HR PBO data, namely unknown COVID-19 vaccination status (although with the majority expected to be vaccinated according to US census data), some missing features (*e.g.*, VL, haptoglobin, serology status), and impreciseness of timing of measurement relative to indexed COVID-19 infection. It likely also had captured a cohort with more advanced disease as a byproduct of selecting for patients with fewer missing data; this is manifested in that prevalence of SD patients increased to 17.8% after filtering from 7.99% before filtering. In addition, the quality of RWD in general will be inferior to clinical trial data because it is generated in a less controlled environment. The above discrepancies between the RWD and the EPIC-HR study cohort data made this assessment of generalizability a conservative one. This domain shift in input will likely degrade the performance of the AI/ML validation models in RWD. As such, we designed three validation strategies to gather evidence for the claims made using EPIC-HR data in an independent dataset.

8.2. Selection of independent dataset

Despite its limitations, Optum-EHR RWD were our best option at hand. We evaluated the suitability of two other COVID-19–related datasets available for the purposes of validating of our findings. One had very limited lab data and was in a younger population than EPIC-HR because participants in the study underwent testing as a requirement for work. The other dataset did not have hospitalization/death endpoint or lab data recorded, was mostly collected during the Omicron time period, and did not select for high-risk patients. As such, the Optum-EHR data offered the potential of closest match to EPIC-HR and thus were chosen for the validation work.

Citations

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

[2] Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, *et al*. A new equation to estimate glomerular filtration rate. *Ann. Intern. Med*. 2009, 150(9):604–612.

[3] Orsucci D, Trezzi M, Anichini R, Blanc P, Barontini L, *et al*. Increased Creatine Kinase May Predict A Worse COVID-19 Outcome. *J. Clin. Med*. 2021, 10(8):1734.

[4] Akbar MR, Pranata R, Wibowo A, Lim MA, Sihite TA, *et al*. The prognostic value of elevated creatine kinase to predict poor outcome in patients with COVID-19 – A systematic review and meta-analysis. *Diabetes Metab. Syndr*. 2021, 15(2):529–534.

[5] Chan KH, Farouji I, Abu Hanoud A, Slim J. Weakness and elevated creatinine kinase as the initial presentation of coronavirus disease 2019 (COVID-19). *Am. J. Emerg. Med*. 2020, 38(7):1548.e1–1548.e3.

[6] Larsson AO, Hultstrom M, Frithiof R, Nyman U, Lipcsey M, *et al*. Differential Bias for Creatinine- and Cystatin C—Derived Estimated Glomerular Filtration Rate in Critical COVID-19. *Biomedicines*. 2022, 10(11):2708.

[7] Sui J, Noubouossie DF, Gandotra S, Cao L. Elevated Plasma Fibrinogen Is Associated With Excessive Inflammation and Disease Severity in COVID-19 Patients. *Front. Cell. Infect. Microbiol*. 2021, 11:734005.

[8] Zhou X, Chen D, Wang L, Zhao Y, Wei L, *et al*. Low serum calcium: a new, important indicator of COVID-19 patients from mild/moderate to severe/critical. *Biosci. Rep*. 2020, 40(12).

[9] Elham AS, Azam K, Azam J, Mostafa L, Nasrin B, *et al*. Serum vitamin D, calcium, and zinc levels in patients with COVID-19. *Clin. Nutr. ESPEN*. 2021, 43:276–282.

[10] Brekke IJ, Puntervoll LH, Pedersen PB, Kellett J, Brabrand M. The value of vital sign trends in predicting and monitoring clinical deterioration: A systematic review. *PLoS One*. 2019, 14(1):e0210875.

[11] Szarpak L, Ruetzler K, Safiejko K, Hampel M, Pruc M, *et al*. Lactate dehydrogenase level as a COVID-19 severity marker. *Am. J. Emerg. Med*. 2021, 45:638–639.

[12] Henry BM, Aggarwal G, Wong J, Benoit S, Vikse J, *et al*. Lactate dehydrogenase levels predict coronavirus disease 2019 (COVID-19) severity and mortality: A pooled analysis. *Am. J. Emerg. Med*. 2020, 38(9):1722–1726.

[13] Hu R, Han C, Pei S, Yin M, Chen X. Procalcitonin levels in COVID-19 patients. *Int. J. Antimicrob. Agents*. 2020, 56(2):106051.

[14] Peng M, He J, Xue Y, Yang X, Liu S, *et al*. Role of Hypertension on the Severity of COVID-19: A Review. *J. Cardiovasc. Pharmacol*. 2021, 78(5):e648–e655.

[15] Bzeizi K, Abdulla M, Mohammed N, Alqamish J, Jamshidi N, *et al*. Effect of COVID-19 on liver abnormalities: a systematic review and meta-analysis. *Sci. Rep*. 2021, 11(1):10599.

[16] Haghi Ashtiani MT, Sadeghi Rad P, Asnaashari K, Shahhosseini A, Berenji F, *et al*. Role of serology tests in COVID-19 non-hospitalized patients: A cross-sectional study. *PLoS One*. 2022, 17(4):e0266923.

[17] Edouard S, Colson P, Melenotte C, Di Pinto F, Thomas L, *et al*. Evaluating the serological status of COVID-19 patients using an indirect immunofluorescent assay, France. *Eur. J. Clin. Microbiol. Infect. Dis*. 2021, 40(2):361–371.

[18] Chen J, Wu C, Wang X, Yu J, Sun Z. The Impact of COVID-19 on Blood Glucose: A Systematic Review and Meta-Analysis. *Front. Endocrinol*. 2020, 11:574541.

[19] Codo AC, Davanzo GG, Monteiro LB, de Souza GF, Muraro SP, *et al*. Elevated Glucose Levels Favor SARS-CoV-2 Infection and Monocyte Response through a HIF-1alpha/Glycolysis-Dependent Axis. *Cell Metab*. 2020, 32(3):498–499.

[20] Toussi SS, Hammond JL, Gerstenberger BS, Anderson AS. Therapeutics for COVID-19. *Nat. Microbiol*. 2023, 8(5):771–786.

[21] Luan YY, Yin CH, Yao YM. Update Advances on C-Reactive Protein in COVID-19 and Other Viral Infections. *Front. Immunol*. 2021, 12:720363.

[22] 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.

[23] Shi F, Wu T, Zhu X, Ge Y, Zeng X, *et al*. Association of viral load with serum biomakers among COVID-19 cases. *Virol*. 2020, 546:122–126.

[24] 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.

[25] Tang Y, Liu J, Zhang D, Xu Z, Ji J, Wen C. Cytokine Storm in COVID-19: The Current Evidence and Treatment Strategies. *Front. Immunol*. 2020, 11:1708.

[26] Hu B, Huang S, Yin L. The cytokine storm and COVID-19. *J. Med. Virol*. 2021, 93(1):250–256.

[27] Dai W, Rao R, Sher A, Tania N, Musante CJ, Allen R. A Prototype QSP Model of the Immune Response to SARS-CoV-2 for Community Development. *CPT: Pharmacometrics Syst. Pharmacol*. 2021, 10(1):18–29.

[28] Rao R, Musante CJ, Allen R. A quantitative systems pharmacology model of the pathophysiology and treatment of COVID-19 predicts optimal timing of pharmacological interventions. *NPJ Syst. Biol. Appl*. 2023, 9(1):13.

[29] Ceciliani F, Giordano A, Spagnolo V. The systemic reaction during inflammation: the acute-phase proteins. *Protein Pept. Lett*. 2002, 9(3):211–223.

[30] 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.

[31] Yagci S, Serin E, Acicbe O, Zeren MI, Odabasi 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(Suppl 1):142–151.

[32] Gabay C, Kushner I. Acute–phase proteins and other systemic responses to inflammation. *N. Engl. J. Med*. 1999, 340(6):448–454.

[33] 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.