Predicting COVID19 Risk: A Retrospective Cohort Study using the National Covid Cohort Collaborative (N3C)

Andrew Barros MD, Jason Adams MS, Robert Link PhD on behalf of the National COVID Cohort Collaborative (N3C) Consortium

RATIONALE We aimed to create and critically examine models to predict the risk of hospitalization at the time of COVID19 diagnosis.

PARTICIPANTS We used the N3C dataset to identify patients with COVID19 from the start of the pandemic until 5/11/2023. We excluded patients younger than 16 years, patients contributed by a data partner with >10% missing geographic data, and patients who were hospitalized on the calendar day of their test.

OUTCOMES Our primary outcome was hospitalization within the 16 days following a COVID19 diagnosis

MEASURES We included demographics, comorbid conditions, medication exposures, and zip-code level social determinants of health.

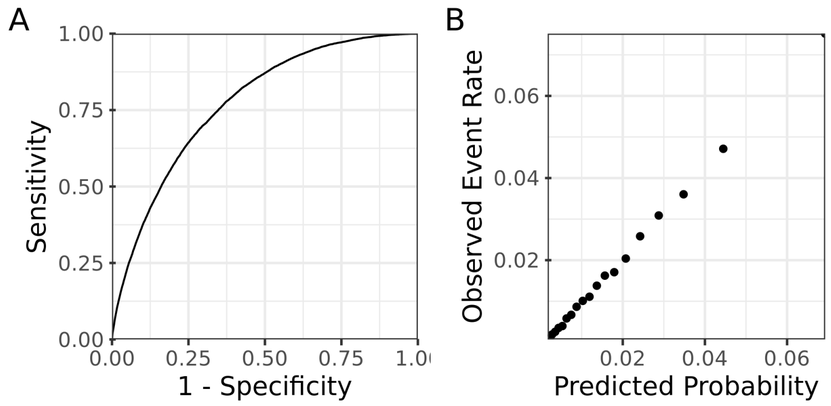
MODEL DEVELOPMENT We compared two methods: a gradient boosted tree (GBT) ensemble model (LightGBM, Redmond, Washington) and machine learning optimized sparse scorecard model (FasterRISK, Durham, North Carolina). We used 80% of the data for training and 20% of the data for model validation.

RESULTS

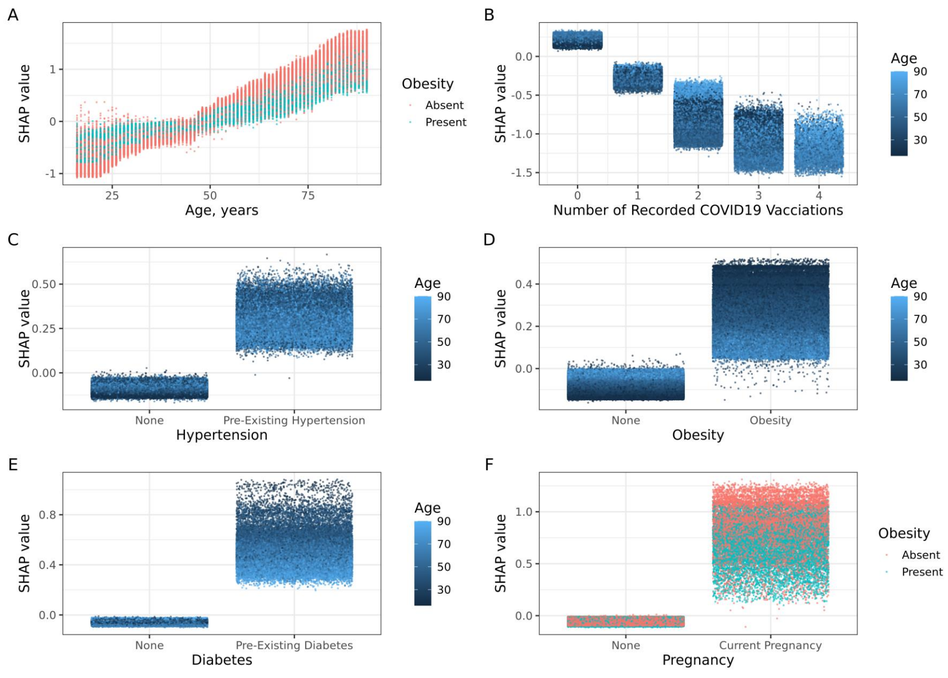
Our final training cohort consisted of 3.6 million patients with a 1.7% rate of hospitalization. Five data partners were excluded for >10% missing patient geographic data. The GBT model had a validation set AUROC of 0.773 and a brier skill score of 0.021 The sparse decision rule performance was slightly lower with a validation AUROC 0.735 and a brier skill score of 0.013. Performance of the GBT model varied across subgroups. Across gender the model had a validation set AUROC of 0.770 in men and 0.777 in women. Across patient self reported race, the model had an AUROC on 0.787 in white subjects and 0.748 in Black subjects.

In the final GBT model the ten most important features were age, count of recorded vaccinations, hypertension, obesity, diabetes, current pregnancy, the % of pop with less than a high school education, the % of the pop that is white, the % of the pop that lives below the federal poverty line, and prior corticosteroid use

AUROC and Calibration for the GBT Model



Feature Importance for the GBT Model



Shap values on the log-odds scale.

← Download the poster
Download the abstract →

Sparse Decision Rule Model

Parameter	Score
Age ≤ 34	-2
Age ≤ 50	-3
Age ≤ 76	-2
Pre-existing kidney disease	2
Pre-existing diabetes	2
Current pregnancy	4
Systemic corticosteroids prior to diagnosis	1
Pre-existing hypertension	1
Receipt of at least 1 vaccination against SARS-CoV-2	-4
% of population in ZCTA with less than a high school education ≤ 23%	-2

Total Score	Model Risk	Event Rate	Sample Size
-13	0.001	0.001	7929
-12	0.002	0.003	1564
-11	0.002	0.001	30,023
-10	0.003	0.004	7455
-9	0.004	0.003	59,561
-8	0.005	0.005	22,060
-7	0.006	0.004	179,333
-6	0.008	0.008	60,607
-5	0.010	0.010	136,291
-4	0.013	0.014	77,067
-3	0.016	0.019	54,694
-2	0.020	0.021	121,319

Total Score	Model Risk	Event Rate	Sample Size
-1	0.026	0.029	68,450
0	0.033	0.037	47,616
1	0.041	0.042	32,326
2	0.052	0.049	20,914
3	0.065	0.065	9971
4	0.082	0.071	8706
5	0.102	0.079	2117
6	0.126	0.073	1708
7	0.155	0.036	28
8	0.190	***	***
9	0.230	***	***
10	0.275	***	***

Cell counts < 20 suppressed due to N3C policy

CONCLUSIONS

- Hospitalization can be predicted with reasonable discrimination and modest overall accuracy.
- A sparse decision rule has similar performance as a gradient boosted tree
- Most of our identified risk factors comport with accepted risk factors
- Identified social determinants are likely proxies for latent causes such as social vulnerability, structural racism, and historical injustice that continue to affect the health of people today.

The N3C Publication committee confirmed that this poster <MSID:1634.108> is in accordance with N3C data use and attribution policies; however, this content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the N3C program. The analyses herein were conducted using the NCATS N3C Data Enclave supported by NCATS U24 TR002306 and made possible because of the patients whose data was contributed by partner organizations (covid.cd2h.org/dtas). We gratefully acknowledge the scientists who have contributed to the on-going development of this community resource (covid.cd2h.org/acknowledgements)

This work was supported by a N3C PHASTR Award to the University of Virginia. The work of Dr. Barros was conducted with the support of the Integrated Translational Health Research Institute of Virginia (iTHRIV) Scholars Program. The iTHRIV Scholars Program is supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health (NIH) under Award Numbers UL1TR003015 and KL2TR003016 as well as by the University of Virginia (UVA). This content is solely the responsibility of the authors and does not necessarily represent the official views of NIH or UVA.