Who Was at Risk for COVID-19 Late in the US Pandemic? Insights From a Population Health Machine Learning Model

Elijah A. Adeoye (elijah.adeoye@providence.org)
Providence St Joseph Health https://orcid.org/0000-0003-4203-8499
Yelena Rozenfeld
Providence St Joseph Health
Jennifer Beam
Providence St Joseph Health
Karen Boudreau
Providence St Joseph Health
Emily Cox
Providence St Joseph Health https://orcid.org/0000-0003-3929-3317
James M. Scanlan
Swedish Center for Research and Innovation

Research

Keywords: COVID-19, infection, risk, social determinants of health

Posted Date: September 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-907939/v1

Abstract

Background: Notable discrepancies in vulnerability to COVID-19 infection have been identified between specific population groups and regions in the United States. The purpose of this study was estimate likelihood of COVID-19 infection using a machine-learning algorithm that can be updated continuously based on health care data.

Methods: Patient records were extracted for all COVID-19 nasal swab PCR tests performed within the Providence St. Joseph Health system from February to October of 2020. Several different machine learning models were tested to evaluate effects of sociodemographic, environmental, and medical history factors on risk of initial COVID-19 infection.

Results: A total of 316,599 participants were included in this study and approximately 7.7% (n = 24,358) tested positive for COVID-19. A gradient boosting model, LightGBM (LGBM), predicted risk of initial infection with an area under the receiver operating characteristic curve of 0.819. Factors that predicted infection were cough, fever, being a member of the Hispanic or Latino community, being Spanish speaking, having a history of diabetes or dementia, and living in a neighborhood with housing insecurity.

Conclusion: A model trained on sociodemographic, environmental, and medical history data performed well in predicting risk of a positive COVID-19 test. This model could be used to tailor education, public health policy, and resources for communities that are at the greatest risk of infection.

Background

Early in the coronavirus disease 2019 (COVID-19) pandemic, a popular interest in predicting risk of infection gave rise to mobile applications and tools for predicting exposure risk. These tools used factors such as medical history, mask compliance, location, demographics, and social activity to predict likelihood of infection or mortality (1). As the pandemic progressed, systematic reviews elucidated additional individual- and population-level characteristics associated with disease progression and mortality. At-risk groups identified by our group and others included people who were older, had laboratory markers of kidney or liver dysfunction, were current smokers, had pre-existing cardiovascular disease, or were Asian, Black, Hispanic or Latino, and non-English-speaking (2–4). These early efforts to categorize at-risk populations were instructive and shaped the initial clinical and population-level responses to the pandemic. However, they generally relied on traditional statistical techniques and limited amounts of data available at the time.

In 2021, mass vaccinations altered risk of COVID-19 infection for much of the US population, but did not eliminate the need for risk prediction. Emergence of vaccine-eluding variants, barriers to accessing vaccines, and widespread vaccine refusal have made it important to continuously re-evaluate risk on an ongoing basis, particularly because disparities in vaccine acceptance may overlap with disparities in infection and/or severe outcomes. For example, older individuals (who were the first to be offered vaccines) are more likely to accept COVID-19 vaccinations than younger individuals, and acceptance

rates are highest among Asian and Alaska Native/American Indian populations, and lowest among Black people (40%) (5, 6).

To respond to the need for updated risk assessments, we have updated our previous risk predictions (2) using a more sophisticated machine learning technique in a larger sample of patient data. Our findings confirm the need for ongoing risk assessment and focusing public resources on the highest-risk communities.

Methods

Ethical approval

The Providence Institutional Review Board (IRB) approved this study and waived the requirement for written informed consent.

Data sources

Data for the development and validation data sets were collected from the electronic medical record (EMR) of Providence St. Joseph Health. Records were included for all people from Alaska, Washington, Oregon, Montana, and California who had at least one COVID-19 PCR test result on a nasal swab sample between February 21, 2020 and October 20, 2020. People with at least one positive test were coded as a positive for infection; people with exclusively negative tests were coded as negative for infection. Location outcomes were evaluated by linking EMR geocoded data to data from the U.S. Census Bureau's 2018 American Community Survey at the census block group or tract level as previously described (2).

Data were split into training and test sets with a 75/25 ratio, respectively, and a random seed for reproducibility. Additional modeling was performed with the final selected model with a train, test, and validation split (80/10/10 ratio, respectively). Two sets of training data were also generated: with clinical symptoms (fever, cough, myalgia, sore throat, chills, and shortness of breath) and without.

Statistics

All major statistical analyses were performed using Python versions 3.6.12 on a 64-bit computer and 3.6.10 leveraging a GPU instance in the Azure Machine Learning ecosystem.

Data cleaning

Continuous variables were standardized or log normalized to address skew and the influence of large values and outliers on the predictive power of trained models. Categorical variables were encoded and dummy variables were created for those variables with more than two classes. Variables were treated mostly as missing not at random (MNAR) except body mass index (BMI) and gender. Missing data for MNAR variables were coded as a separate category, e.g. 'Unknown'. For BMI, median imputation was used to fill in the large amount of missing data (n = 25,646 from initial participant pool, approximately 8%).

Gender was analyzed as legal sex, and missing values were dropped (n = 119; 0.04% of initial participant pool).

Hyperparameter tuning and cross validation

We used a randomized search approach, with cross validation, to tune and identify critical hyperparameters for each model (**Supplementary Material**). A set of hyperparameters that produced the best area under the curve (AUC) on the training set were selected as part of the final ensemble. This was performed with a repeated, stratified k-fold cross validation with 10 splits and 3 repeats. A random seed was set for reproducibility of the cross-validation step. We chose a randomized approach due to the computationally intensive nature of the alternative, more comprehensive grid search approach. We report the best hyperparameters selected for the best model with symptoms (**Supplementary Material**).

Model training and selection

An ensemble approach was used as the predictive model for each possible experiment. Four models – Logistic Regression, Random Forest, and two gradient boosting libraries, XGBoost (XGB) and LightGBM (LGBM) – were used as classifiers for training. We selected the best hyperparameters for each classifier, after hyperparameter tuning, and included these as part of the ensemble for the prediction task. We used a soft-voting ensemble due to the need to compute probabilities of a positive test or event.

Data augmentation

Most COVID-19 test results were negative. Thus, different data augmentation techniques were applied to address class imbalance by over-sampling and/or down-sampling the minority and majority class, respectively. We used a Synthetic Minority Oversampling Technique (SMOTE) and case-control approach to augment the training data as part of multiple modeling experiments. SMOTE is used to create synthetic data that is close, or nearest neighbor, to the minority class in the feature space (7). We also experimented with a case-control (CC) approach typically used in epidemiological studies to create a 1:1 match by down-sampling the majority class (COVID-19 negative) to the size of the minority class. Negative classes were selected using a simple random sample method without replacement. This strategy, unlike SMOTE techniques, uses real, non-synthetic data for model training. These approaches helped to create a 1:1 match of the negative (majority) class and the positive class (Class 0: 19,390, Class 1: 19,390, respectively). No augmentation was performed on the validation/test data set.

Twelve experiments were conducted such that at each experiment, models were fitted on the training set depending on whether data augmentation and dimensionality reduction techniques were applied to that set (Fig. 1). For dimensionality reduction, we applied principal component analysis (PCA) to compute the minimal set of principal components that explained 95% of the variance in the data. Recursive feature elimination (RFE) approach was also used, as part of different experiments, to select the minimal set of predictors that were most predictive for a COVID-19 positive test. Dimensionality reduction techniques

were also applied on the test/validation sets; however, no augmentation was applied to the validation/test data set. PCA was not applied to comparative logistic regression models.

Feature importance

We used the Python implementation of SHAP (SHapley Additive exPlanations) (8) to examine the key predictor variables that contribute to a patient's probability of a positive COVID-19 test result. The library computes Shapley values, which aim to demonstrate the marginal contribution of a feature to the predicted outcome of a vector or an instance (9). This approach examines how much each feature in the model pushes the predicted value of that instance from a baseline, or average, prediction (expected value). Using the SHAP methodology provides a method for improving the interpretability of a machine learning model. SHAP values were computed using the final selected model.

Results

Study participants

A total of 316,599 participants were included in this study and approximately 7.7% tested positive for COVID-19 (n = 24,358). The average age was 47 ± 22 years old, 56.7% (179,381) were female, 63% (199,492) were identified as white or Caucasian, and 55.2% (174,683) had at least one chronic condition (Table 1).

Model performance

In general, models trained with CC augmented data performed better on test/validation sets than SMOTE augmented data. Area under the receiver operating characteristic curve (AUC) scores for models that included symptoms and were trained on augmented data ranged approximately from 0.756–0.816, while the logistic regression model trained on non-augmented data yielded an AUC of 0.767. The gradient boosting library, LightGBM (LGBM), produced an AUC of 0.816. Because this model is computationally lightweight compared to ensembling all models, separate analyses were performed with this model on CC augmented training data split into training/testing/validation sets (80/10/10 ratio, respectively). LGBM AUC on the training set with repeated, stratified k-fold cross validation with 10 splits and 3 repeats gave a mean AUC of 0.811 ± 0.007. AUC was approximately 0.819 on the test set and 0.814 on the validation set.

When symptoms (fever, cough, myalgia, sore throat, chills, and shortness of breath) were not included as predictive variables, AUC on the training set with the same cross validation approach was acceptable, but comparatively poorer (0.735 ± 0.007). AUC on the test and validation sets was 0.734 and 0.727, respectively (Table 2).

Feature importance

Model with symptoms

When symptoms were included as predictors of infection risk, cough and fever were the two most important predictors (Fig. 2A). Being a member of the Hispanic or Latino community, living in the Washington-Montana or Southern California regions, being non-English-speaking and especially Spanish-speaking, polypharmacy, and having shortness of breath were all comparable influences on the risk of a positive COVID-19 test (SHAP scores 0.10–0.30). All of these features except polypharmacy were also directly associated with risk of infection from COVID-19, while polypharmacy, co-morbidity, higher income, and tobacco or alcohol use were inversely associated with risk of infection (Fig. 2B).

Model without symptoms

Because symptom information may not always be available for risk assessments of the population at large, a second model was developed to assess the importance of static population factors. When symptoms were removed from the predictive model, being of Hispanic/Latino ethnicity became the most important predictor of COVID-19 infection (Fig. 3A) in this patient population. Other risk factors with at least two-fold lower SHAP scores included speaking Spanish, being from Montana or a region with housing instability, identifying with an “other” race category, using tobacco, being male, being Christian, and having an “other” BMI. Tobacco use, co-morbidity, polypharmacy, an “other” BMI category, income level, and illicit drug use were inversely associated with risk of infection, while other features were positively associated with this risk (Fig. 3B).

Discussion

Although COVID-19 vaccines are now widely available, predicting the risk of COVID-19 infection remains critical. Unvaccinated populations and new variants of COVID-19 present an ongoing threat to disease control worldwide, and risk prediction is still needed to (1) to assist clinicians and care managers in patient education, (2) guide policy, and (3) allocate resources to the highest risk areas and populations. Our findings indicate that, as expected, fever and cough were the strongest predictors of infection. This validates public guidance to quarantine based on symptoms alone. However, when we removed symptoms from the model to assess static (i.e., not symptom-based) features alone, the following groups in the western US emerged with the highest risk for infection: Hispanic and Latino people, individuals in the “other” race category, non-English-speaking people (particularly Spanish-speaking people), people living in areas with housing insecurity, and people from the Washington-Montana region. Compared to previous similar projects, advantages of the current analysis are the size and geographical spread of the dataset, and the machine learning technique which allows the results to be updated in nearly real-time. We intend to update these results as the pandemic continues.

Immediate recommendations based on the results of this project are as follows. Culturally literate and language-appropriate resources are needed to combat surging infection rates in Hispanic, Latino, and non-English-speaking populations in the western US. Partnering with communities to assure broad availability of information and access to services is critical to reducing disproportionate burden, and such partnership may increase trust in the information that is provided. Clinicians should be aware that

individuals from these populations may be at higher risk and should conduct assessments and provide education accordingly. For example, clinicians may ask their patients whether they have access to masks and cleaning/disinfection supplies, or whether they need assistance accessing vaccine appointment registration systems. Individuals who are not at high risk themselves but have frequent contact with high-risk groups may require more frequent or intense training on infection control precautions. Finally, public efforts to combat the spread of COVID-19 must address issues such as access, physical proximity of vaccine clinics to high-risk populations, and pro-active program development for non-English speaking groups.

These results differ from our previous results from the early period of the pandemic.(2) The present results did not confirm that older, immunocompromised, or Black people were at significantly greater risk of COVID-19 infection in this study population. This difference may reflect the change in technique from traditional logistic regression to a machine learning algorithm. This more sophisticated technique may have elucidated underlying factors that were not immediately apparent with logistic regression, because it focused on predictive performance rather than traditional inference about individual variables and strict cut-off thresholds based on statistical significance. It is also possible that these groups are genuinely at higher risk but became under-represented and under-counted in the larger dataset, and thus their risk levels may have been underestimated.

An additional explanation for the shifting results is the expansion of the window of time over which results were counted. The previous work examined data from February to April of 2020,(2) while the present work extended the data to October of 2020, encompassing the second and early third “waves” of cases occurring between mid-June and October. During this later period, state and local public health departments instituted substantially more stringent transmission-reduction strategies including tight restrictions on public gatherings, remote school and work, universal masking requirements in public spaces, and “stay-at-home” policies. Thus, we may have captured real changes in population risk as the pandemic progressed. This may underlie the finding that young people between 18 and 29 were at higher risk, while older people were no longer at higher risk. As the pandemic progressed, older individuals may have been more compliant with stringent quarantine and isolation precautions due to well-publicized fears of mortality, while younger individuals were perhaps less cautious, and thus continued to become infected.

We developed predictive models both including and excluding symptoms for different purposes. Modeling risk of infection without symptoms was done to evaluate static risk for populations in the western US. The intention of this step was to aid in planning for disease control and prevention within the Providence St. Joseph Health system. In response to this model, Providence St. Joseph Health tailored the selection of sites for COVID testing and vaccination as well as engagement with community organizations. We recommend that other large health systems implement models of this kind to understand underlying risk factors in their patient populations and target infection control responses accordingly.

There are several limitations to this study. First, models were trained based on data that would be available to an outpatient clinician (patient medical history, sociodemographic, self-reportable symptoms, and environmental data). While this was intentional in order to make the model generalizable to various clinical settings, laboratory values such as white blood cell counts (lymphocyte, eosinophil, basophil, and neutrophil values) (10) may have improved performance of the model that included symptoms. Second, the data collection period (February – October 2020) spanned a period of rapidly evolving public health guidelines. This may have influenced some of the findings. For example, the finding that older age was not predictive of a higher risk of COVID-19 infection may reflect greater caution and compliance with stay-at-home orders among older populations. Third, the study did not include the largest part of the third wave, from October 2020 to March 2021; consequently, we intend to update these findings using the same machine learning method as the pandemic continues to progress. Fourth, we suggest that that the population-level characteristics spotlighted by this model (e.g., race, ethnicity, language) are not inherent predictors of risk, but rather are proxy indicators for living conditions (housing density and ability to socially isolate) and social structures, such as systemic racism in healthcare and public policy.

Conclusions

Our results confirm that the following social and demographic factors increased the risk of COVID-19 infection between February to October of 2020: being Hispanic and Latino, being non-English-speaking (and especially Spanish speaking), residing in an area that had housing insecurity, or being from the region of Washington and Montana. These findings confirm that social determinants of health were major drivers of infection risk in the late part of the pre-vaccine US COVID-19 pandemic. Language-appropriate and community-based education is needed to mitigate the effects of social factors on infection risk. Additionally, providers should focus education efforts on patients who fall into high-risk categories or are frequently in contact with individuals from high-risk categories.

Abbreviations

AUC: area under the curve

BMI: body mass index

CC: case-control

COVID-19: coronavirus disease 2019

EMR: electronic medical record

LGBM: gradient boosting model

MNAR: missing not at random

PCA: principal component analysis

RFE: recursive feature elimination

SHAP: SHapley Additive exPlanations

SMOTE: Synthetic Minority Oversampling Technique

XGB: gradient boosting model

Declarations

Ethics approval and consent to participate

The Providence Institutional Review Board (IRB) approved this study and waived the requirement for written informed consent.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due the nature of the data (health record data from electronical medical records).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was not supported by funding.

Authors' contributions

E.A.A. designed the study, performed machine learning experiments, generated the figures, and contributed to the writing and critical reviewing of the manuscript.

Y.R. contributed to the design of the study as well as writing and critical reviewing of the manuscript.

J.B., K.B., E.J.C., and J.M.S. contributed to the writing and critical reviewing of the manuscript.

Acknowledgements

The authors wish to thank Uma Kodali Bhavani and Morgan Goodwin for their incredible efforts providing the data that was critical to this work.

References

1. Eisenstein M. What's your risk of catching COVID? These tools help you to find out.: Nature 589, 158-159 (2021); 2020 [Available from: https://www.nature.com/articles/d41586-020-03637-y.
2. Rozenfeld Y, Beam J, Maier H, Haggerson W, Boudreau K, Carlson J, et al. A model of disparities: risk factors associated with COVID-19 infection. Int J Equity Health. 2020;19(1):126.
3. Zheng Z, Peng F, Xu B, Zhao J, Liu H, Peng J, et al. Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. J Infect. 2020;81(2):e16-e25.
4. Wolff D, Nee S, Hickey NS, Marschollek M. Risk factors for Covid-19 severity and fatality: a structured literature review. Infection. 2021;49(1):15-28.
5. Malik AA, McFadden SM, Elharake J, Omer SB. Determinants of COVID-19 vaccine acceptance in the US. EClinicalMedicine. 2020;26:100495.
6. Khubchandani J, Sharma S, Price JH, Wiblishauser MJ, Sharma M, Webb FJ. COVID-19 Vaccination Hesitancy in the United States: A Rapid National Assessment. J Community Health. 2021;46(2):270-7.
7. Brownlee J. SMOTE for Imbalanced Classification with Python. Machine Learning Mastery 2020 [Available from: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification.
8. Lundberg S. Welcome to the SHAP Documentation — SHAP latest documentation. SHAP. 2018 [Available from: https://shap.readthedocs.io/en/latest/.
9. Molnar C. SHAP (SHapley Additive exPlanations). 2021. In: Interpretable machine learning: A guide for making black box models explainable [Internet].
10. Mickael T, Ahmed M, Rahul MD, Ines S, Guillaume C, Enora G, et al. Pre-test probability for SARS-Cov-2-related Infection Score: the PARIS score. medRxiv. 2020.

Tables

Table 1
Study Participant Demographics and Characteristics

	Tested people		Tested Positive		Tested Negative	
	(N = 316,599)		(N = 24,358)		(N = 292,241)	
	N	% of total[a]	N	In-group %[b]	N	In-group %[b]
Sociodemographic						
Age						
< 18	25640	8.10%	1766	6.89%	23874	93.11%
18–29	51328	16.21%	4992	9.73%	46336	90.27%
30–39	49570	15.66%	3875	7.82%	45695	92.18%
40–49	41634	13.15%	3565	8.56%	38069	91.44%
50–59	45760	14.45%	3707	8.10%	42053	91.90%
60–69	45976	14.52%	2804	6.10%	43172	93.90%
70–79	34057	10.76%	1941	5.70%	32116	94.30%
80+	22634	7.15%	1708	7.55%	20926	92.45%
Gender						
Female	179,381	56.66%	12826	7.15%	166,555	92.85%
Male	137,218	43.34%	11532	8.40%	125,686	91.60%
Education						
Education < 12 years	219,444	69.31%	13409	6.11%	206035	93.89%
Employment						
Student	17475	5.52%	1574	9.01%	15901	90.99%
Employed	131,019	41.38%	10725	8.19%	120,294	91.81%
Not Employed	58380	18.44%	4946	8.47%	53434	91.53%
Retired	63324	20.00%	3864	6.10%	59460	93.90%
Unknown	46401	14.66%	3249	7.00%	43152	93.00%
Race						
White	199,492	63.01%	9742	4.88%	189,750	95.12%

	Tested people		Tested Positive		Tested Negative	
	(N = 316,599)		(N = 24,358)		(N = 292,241)	
	N	% of total[a]	N	In-group %[b]	N	In-group %[b]
American Indian | Alaska Native	4069	1.29%	293	7.20%	3776	92.80%
Asian	13334	4.21%	1044	7.83%	12290	92.17%
Black | African American	12018	3.80%	1095	9.11%	10923	90.89%
Native Hawaiian | Pacific Islander	2700	0.85%	424	15.70%	2276	84.30%
Hispanic | Latino	39997	12.63%	7962	19.91%	32035	80.09%
Unknown	44989	14.21%	3798	8.44%	41191	91.56%
Ethnicity						
Other Ethnic Groups	276,602	87.37%	16396	5.93%	260,206	94.07%
Hispanic or Latino	39997	12.63%	7962	19.91%	32035	80.09%
Religious Affiliation						
Agnostic	90655	28.63%	5585	6.16%	85070	93.84%
Christian	121,557	38.39%	10293	8.47%	111,264	91.53%
Other Religion	10534	3.33%	679	6.45%	9855	93.55%
Unknown	93853	29.64%	7801	8.31%	86052	91.69%
Relationship						
Single	123,850	39.12%	10096	8.15%	113,754	91.85%
Divorced or Legally Separated	37797	11.94%	2412	6.38%	35385	93.62%
Married or Significant Other	128,944	40.73%	9817	7.61%	119,127	92.39%
Unknown	26008	8.21%	2033	7.82%	23975	92.18%
Language						
English	288,252	91.05%	18964	6.58%	269,288	93.42%
Sino-Tibetan	2192	0.69%	244	11.13%	1948	88.87%
Spanish	12435	3.93%	3679	29.59%	8756	70.41%
Other Languages	13720	4.33%	1471	10.72%	12249	89.28%
Clinical						

	Tested people		Tested Positive		Tested Negative	
	(N = 316,599)		(N = 24,358)		(N = 292,241)	
	N	% of total[a]	N	In-group %[b]	N	In-group %[b]
Body Mass Index						
Normal	66179	20.90%	4231	6.39%	61948	93.61%
Underweight	5180	1.64%	296	5.71%	4884	94.29%
Moderately Obese	45918	14.50%	4061	8.84%	41857	91.16%
Overweight	70933	22.40%	5918	8.34%	65015	91.66%
Severely Obese	23334	7.37%	2078	8.91%	21256	91.09%
Very Severely Obese	19981	6.31%	1643	8.22%	18338	91.78%
Unknown	85074	26.87%	6,131	7.21%	78943	92.79%
Number of Chronic Conditions						
0	141,916	44.83%	12551	8.84%	129,365	91.16%
1–2	103,464	32.68%	7629	7.37%	95,835	92.63%
3–4	46632	14.73%	2905	6.23%	43727	93.77%
5+	24587	7.77%	1273	5.18%	23314	94.82%
Clinical Diagnosis						
Diagnosis of Diabetes	34930	11.03%	3340	9.56%	31992	91.59%
Diagnosis of Kidney Disease	789	0.25%	94	11.91%	709	89.86%
Diagnosis of HIV/AIDS	767	0.24%	54	7.04%	718	93.61%
Diagnosis of Dementia	7316	2.31%	910	12.44%	6510	88.98%
Polypharmacy						
0 Prescriptions	104,273	32.94%	9066	8.69%	95207	91.31%
1–9 Prescriptions	160,387	50.66%	12403	7.73%	147,984	92.27%
10–19 Prescriptions	38656	12.21%	2238	5.79%	36418	94.21%
20–29 Prescriptions	9809	3.10%	481	4.90%	9328	95.10%
30 + Prescriptions	3474	1.10%	170	4.89%	3304	95.11%
Mental Health and Substance Use						

	Tested people		Tested Positive		Tested Negative	
	(N = 316,599)		(N = 24,358)		(N = 292,241)	
	N	% of total[a]	N	In-group %[b]	N	In-group %[b]
History of Illicit Drug Use	35588	11.24%	1561	4.39%	34027	95.61%
History of Tobacco Use	40352	12.75%	1836	4.55%	38516	95.45%
Diagnosis of Serious Persistent Mental Illness	30246	9.55%	1286	4.25%	28960	95.75%
Diagnosis of Substance Use Disorder	24757	7.82%	1071	4.33%	23686	95.67%
Primary Care Affiliation						
Internal Primary Care Provider	112,191	35.44%	7017	6.25%	105,174	93.75%
External Primary Care Provider	116,348	36.75%	8708	7.48%	107,640	92.52%
Unknown Primary Care Provider	88060	27.81%	8633	9.80%	79427	90.20%
Symptoms						
Fever	101388	32.02%	15157	14.95%	86231	85.05%
Cough	113047	35.71%	16319	14.44%	96728	85.56%
Breath	107216	33.86%	13642	12.72%	93574	87.28%
Chills	6443	2.04%	950	14.74%	5493	85.26%
Myalgia	8587	2.71%	1686	19.63%	6901	80.37%
Environmental						
Region						
Oregon	83293	26.31%	5018	6.02%	78,275	93.98%
Alaska	17269	5.45%	857	4.96%	16412	95.04%
Puget Sound	34437	10.88%	2144	6.23%	32293	93.77%
Southern California	65815	20.79%	7389	11.23%	58426	88.77%
Washington | Montana	115589	36.51%	8931	7.73%	106,658	92.27%
Unknown	196	0.06%	19	9.69%	177	90.31%
Age-Stratified Communal Living						
Non-Communal Living	230410	72.78%	16624	7.21%	213,786	92.79%

	Tested people		Tested Positive		Tested Negative	
	(N = 316,599)		(N = 24,358)		(N = 292,241)	
	N	% of total[a]	N	In-group %[b]	N	In-group %[b]
Adult Community	12534	3.96%	1055	8.42%	11479	91.58%
Adult and Youth	46996	14.84%	4460	9.49%	42536	90.51%
Multigenerational	15481	4.89%	1535	9.92%	13946	90.08%
Senior Living	2876	0.91%	300	10.43%	2576	89.57%
Other	8302	2.62%	384	4.63%	7918	95.37%
Financial Insecurity	98537	31.12%	10285	10.44%	88252	89.56%
Housing Insecurity	72081	22.77%	8849	12.28%	63232	87.72%
Transportation Insecurity	88401	27.92%	7240	8.19%	81161	91.81%

Legend: Characteristics of the patient population included in this analysis. [a] % of total is the percentage of the total N (316,599. [b] In-group % is the percentage of the total tested people for each row.

Table 2
Area under the curve (AUC) of modeling experiments run to predict COVID-19 risk of infection

Trial	Augmentation / Feature Reduction	Model	AUC	Sensitivity	Specificity
1	RFE	LR	0.767	0.093	0.994
2	CC	LGBM*	0.814	0.718	0.754
3	CC	LGBM*+	0.727	0.623	0.713
4	CC	Ensemble	0.816	0.717	0.760
5	CC	LR	0.800	0.721	0.730
6	CC-PCA	Ensemble	0.805	0.714	0.745
7	CC-RFE	Ensemble	0.816	0.715	0.759
8	CC-RFE	LR	0.800	0.721	0.731
9	SMOTE	Ensemble	0.797	0.552	0.864
10	SMOTE	LR	0.759	0.624	0.759
11	SMOTE-PCA	Ensemble	0.802	0.622	0.823
12	SMOTE-RFE	Ensemble	0.792	0.555	0.858
13	SMOTE-RFE	LR	0.756	0.621	0.760

Legend: Models included symptoms as predictors. Except for the Light Gradient Boosting Machine model (LGBM), reported area under the receiver operating characteristic curve (AUC ROC) scores are for the 25% held-out test set of the 75/25 train/test split. For the LGBM model, a 80/10/10 training/test/validation split was used, and AUC is given for performance on the final validation set.

*****Final selected model.

+Final selected model without symptoms.

RFE = Recursive Feature Elimination; LR = Logistic Regression; CC = Case-Control; LGBM = Light Gradient Boosting Machine; PCA = Principal Component Analysis; SMOTE = Synthetic Minority Oversampling Technique.

Figures

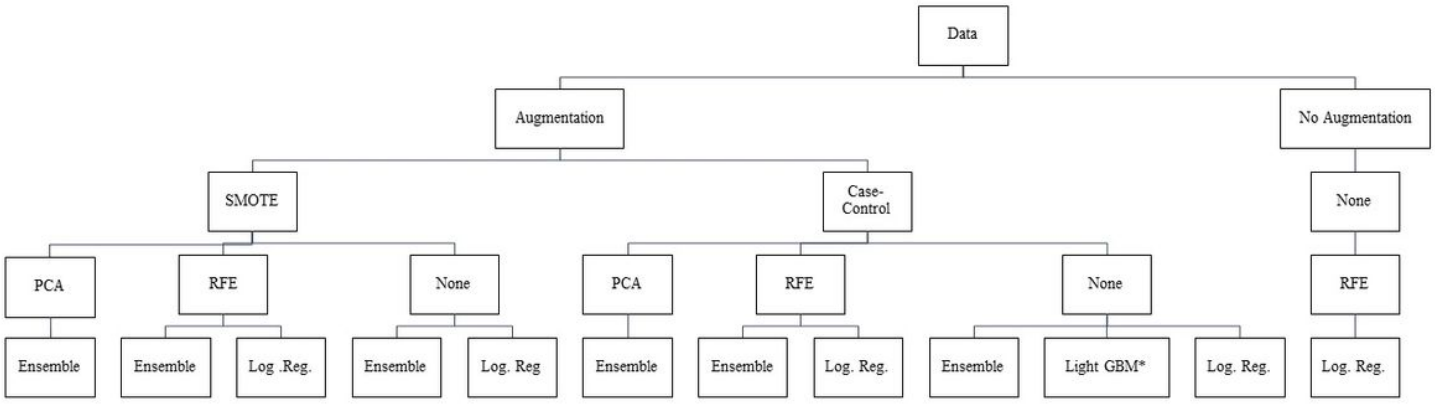
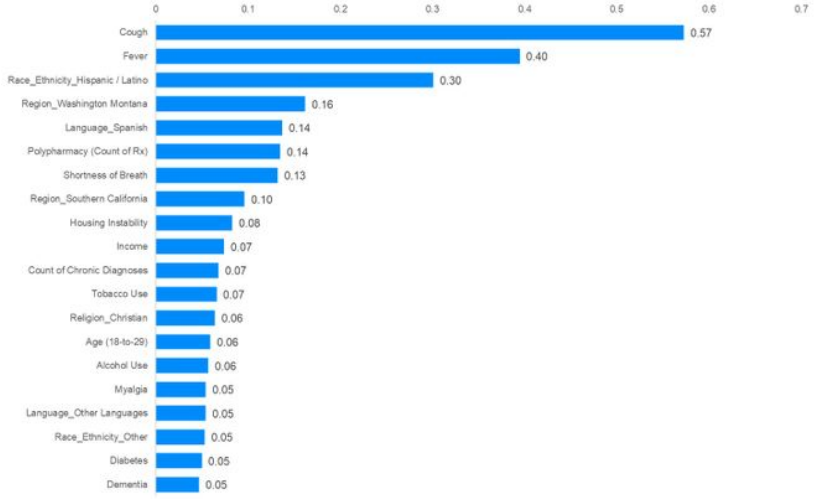
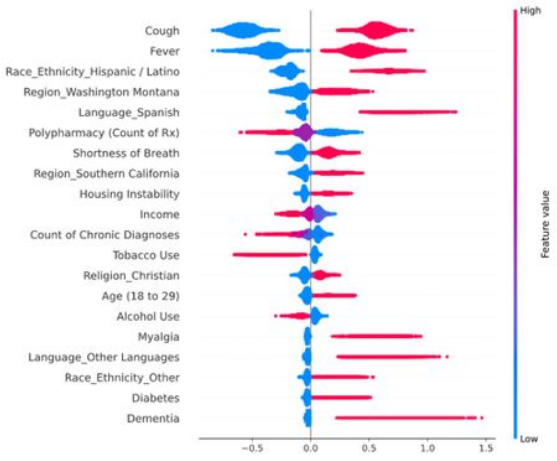


Figure 1

Title: Schematic of predictive modeling experiments performed to predict risk of initial COVID-19 infection Legend: RFE = Recursive Feature Elimination; LR = Logistic Regression; CC = Case-Control; LGBM = Light Gradient Boosting Machine; PCA = Principal Component Analysis; SMOTE = Synthetic Minority Oversampling Technique. *Light GBM (LGBM) was the final selected model.



2a



2b

Figure 2

Title: Relative contribution of predictor variables in a machine learning model trained to predict COVID-19 infection based on symptoms and demographic information Legend: (A) SHapley Additive exPlanations (SHAP) scores for predictor variables input into a Light Gradient Boosting Machine model trained to predict the risk of COVID-19 initial infection. Data displayed are the average magnitude of predictor impact on model output. (B) Chart demonstrates, in descending order, the importance of the top 20 predictors at predicting COVID-19 infection. SHapley Additive exPlanations (SHAP) values were computed using the final Light Gradient Boosting Machine (LGBM) model. The plot is made of a point, or “dots”,

corresponding to each prediction in our training set. Each prediction corresponds to a patient. The horizontal axis shows the impact of each predictor on a low or high prediction value for severe outcomes due to COVID-19 i.e., for each patient. For a given feature, we note the feature's impact on the prediction of the outcome as the value ranges from its lowest (blue) to highest (red) value. Higher SHAP values on the X-axis correspond to increased likelihood of having a positive outcome (i.e., COVID-19 infection). Thus, features with the color scale oriented from blue on the left to red on the right are associated with an increasing probability of infection as the feature increases, such as Cough (0=No Cough, 1=Cough). However, features oriented from red on the left to blue on the right are associated with decreasing risk as the feature increases, such as Polypharmacy.

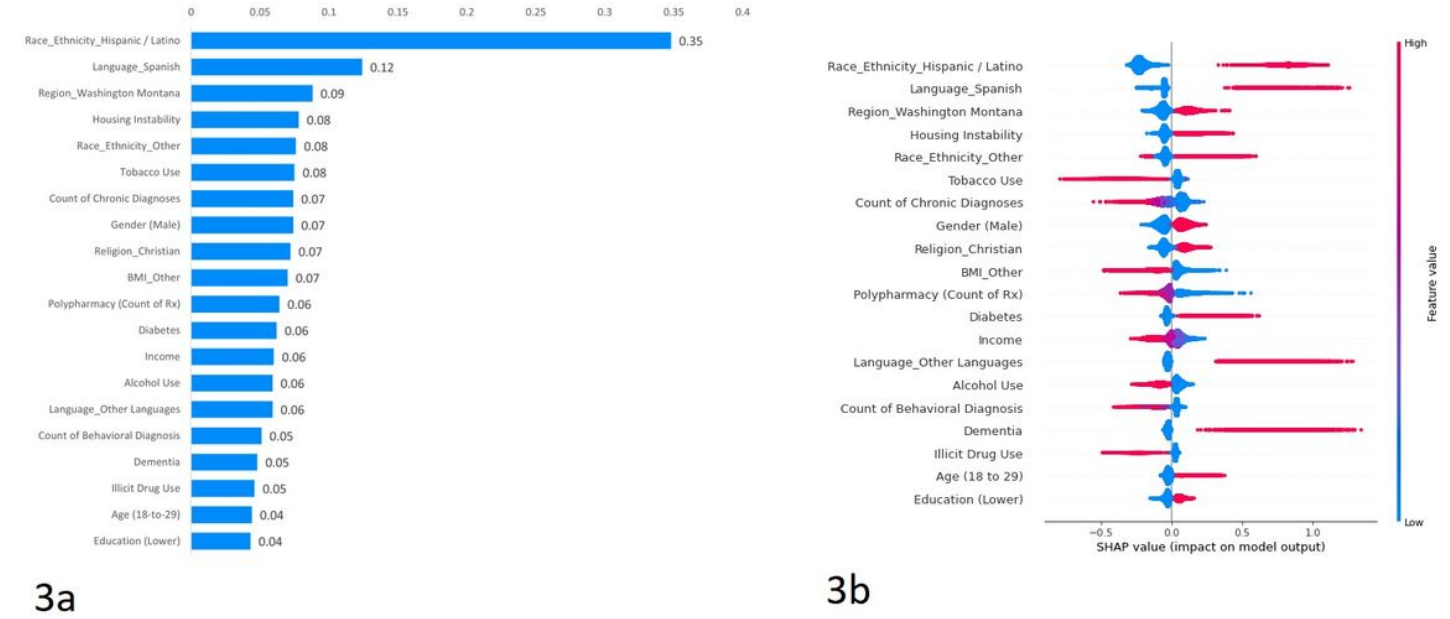


Figure 3

Title: Relative contribution of predictor variables in a machine learning model trained to predict COVID-19 infection based on demographic information Legend: (A) SHapley Additive exPlanations (SHAP) scores for predictor variables input into a Light Gradient Boosting Machine model trained to predict the risk of COVID-19 initial infection. Data displayed are the average magnitude of predictor impact on model output. (B) Chart demonstrates, in descending order, the importance of the top 20 predictors at predicting COVID-19 infection. SHapley Additive exPlanations (SHAP) values were computed using the final Light Gradient Boosting Machine (LGBM) model. The plot is made of a point, or “dots”, corresponding to each prediction in our training set. Each prediction corresponds to a patient. The horizontal axis shows the impact of each predictor on a low or high prediction value for severe outcomes due to COVID-19 i.e., for each patient. For a given feature, we note the feature's impact on the prediction of the outcome as the value ranges from its lowest (blue) to highest (red) value. Higher SHAP values on the X-axis correspond to increased likelihood of having a positive outcome (i.e., COVID-19 infection). Thus, features with the color scale oriented from blue on the left to red on the right are associated with an increasing probability of infection as the feature increases, such as Cough (0=No Cough, 1=Cough). However, features oriented

from red on the left to blue on the right are associated with decreasing risk as the feature increases, such as Polypharmacy.

Supplementary Files

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

- ROIsupplement.docx