Medical & Biological Engineering & Computing (2022) 60:2039–2049
https://doi.org/10.1007/s11517-022-02549-5

ORIGINAL ARTICLE

Who was at risk for COVID-19 late in the US pandemic? Insights from a population health machine learning model

Elijah A. Adeoye[1] · Yelena Rozenfeld[1] · Jennifer Beam[1] · Karen Boudreau[1] · Emily J. Cox[2] · James M. Scanlan[3]

Received: 14 October 2021 / Accepted: 6 March 2022 / Published online: 11 May 2022

Abstract

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

Keywords COVID-19 · Infection · Risk · Social determinants of health

1 Introduction

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 parallel with simpler prediction tools, artificial intelligence (AI) has been used since the early days of the pandemic to classify and predict risk. For example, a recent review of 130 publications found 71 papers related to computational epidemiology of COVID-19, 40 papers related to early detection and diagnosis of COVID-19, and 19 papers related to COVID-19 disease progression [5]. Common techniques used by these studies were deep learning and transfer learning [5]. Elsewhere, an analysis of 264 papers found that the convolutional neural network method was the most frequently applied AI technique in COVID-19 studies, followed by random forest classifier, ResNet, Support Vector Machine, and deep learning [6]. These studies described the rapid expansion in machine learning and AI tools during the COVID-19 pandemic.

✉ Yelena Rozenfeld
Yelena.Rozenfeld@providence.org

Elijah A. Adeoye
elijahaadeoye@gmail.com

[1] Providence St. Joseph Health, 1801 Lind Avenue S.W. Valley Office Park, Morin Bldg, 1st Floor, Renton, WA 98057-9016, USA

[2] Providence Medical Research Center, 105 W 8th Ave, Suite 250E, Spokane, WA 99204, USA

[3] Swedish Center for Research and Innovation, 800 Fifth Ave, 11th floor, Seattle, WA, USA

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 [7, 8].

In 2020, we used logistic regression to examine risk factors associated with COVID-19 infection in 34,503 cases from the Providence health system [2]. As the pandemic evolved, we recognized the need for updated risk assessments and the utility of AI in risk assessment across our growing numbers of cases. Thus, the present paper updates 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.

2 Methods

2.1 Ethical approval

The Providence Institutional Review Board (IRB) approved this study and waived the requirement for written informed consent (IRB identifier STUDY2020000220). The study was conducted in compliance with IRB rules and the Declaration of Helsinki.

2.2 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 US Census Bureau's 2018 American Community Survey at the census block group or tract level as previously described [2].

Two rounds of data splitting were employed. In initial tests, data were split into training and test sets with a 75/25 ratio, respectively, and a random seed for reproducibility (Fig. 1). After we determined that a light gradient boosted model (LGBM) produced the most accurate results, we performed additional modeling with a train, test, and validation split (80/10/10 ratio, respectively). This was done (1) to increase the size of the training set and (2) to avoid overfitting by exploring its performance in both a test and a validation set. Two sets of training data were also generated: with clinical symptoms (fever, cough, myalgia, sore throat, chills, and shortness of breath) and without (Fig. 1).

2.3 Data analysis

2.3.1 Computational environment

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.

2.3.2 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. Count of mental health diagnoses, comorbidities, community size, polypharmacy, and population density each had a skew of 2.58, 1.56, 28.85, 1.04, and − 0.44, respectively. Scaling did not impact the skew for any of these variables. However, log transforming community size reduced its skewness to 4.39. 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).

2.3.3 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 Table 1). 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

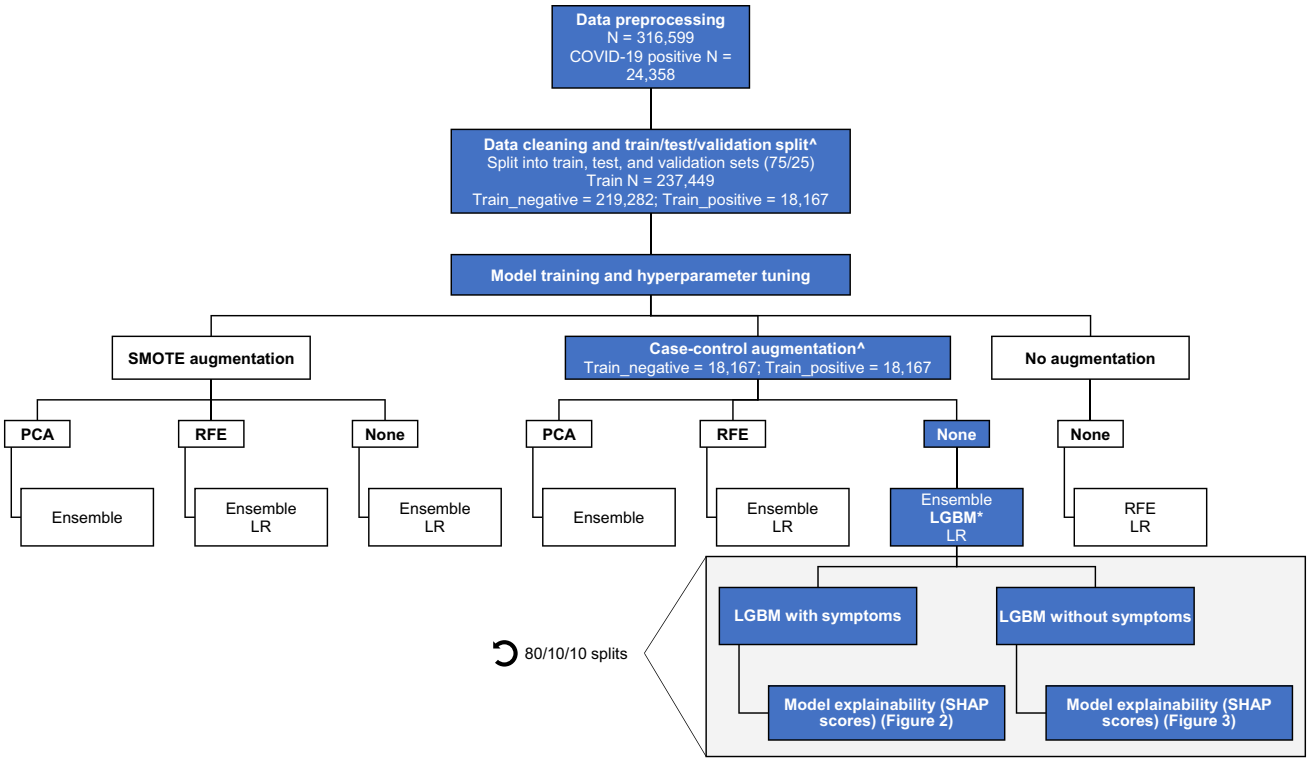

Fig. 1 Schematic of predictive modeling experiments performed to predict risk of initial COVID-19 infection. Legend: ^LGBM outperformed other models on the 25% test set. Thus, we re-trained an LGBM model on a 80/10/10 split (1) to increase the size of the training set and (2) to avoid overfitting by exploring its performance in both a test and a validation set. The training samples, for the 80% split, was 253,279 (train_negative = 233,889; train_positive = 19,390). After case–control augmentation (downsampling the training samples count to the positive samples count), we arrived at a train_negative = 19,390 and train_positive = 19,390. *RFE* = recursive feature elimination, *LR* = logistic regression, *LGBM* = light gradient boosting machine, *PCA* = principal component analysis, *SMOTE* = synthetic minority oversampling technique. *LGBM was the final selected model. The refresh icon indicates that the LGBM model was put through a second round of modeling with a train, test, and validation split of 80/10/10, respectively, for the final steps

selected for the best model with symptoms (Supplementary Material Table 1).

2.3.4 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. This was done to address model bias towards the negative class (i.e., the population of persons who tested negative for COVID-19), which is important to prevent the model from learning to predict the dominant negative class. 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 [9]. 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. 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.

2.3.5 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.

2.3.6 Model explainability

Two LGBMs were generated, one with symptoms and one without symptoms (Fig. 1). We used the Python implementation of SHAP (SHapley Additive exPlanations) [10] 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 [11]. 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.

3 Results

3.1 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).

3.2 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 to 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).

3.3 Feature importance

3.3.1 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).

3.3.2 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).

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	25,640	8.10	1766	6.89	23,874	93.11
18–29	51,328	16.21	4992	9.73	46,336	90.27
30–39	49,570	15.66	3875	7.82	45,695	92.18
40–49	41,634	13.15	3565	8.56	38,069	91.44
50–59	45,760	14.45	3707	8.10	42,053	91.90
60–69	45,976	14.52	2804	6.10	43,172	93.90
70–79	34,057	10.76	1941	5.70	32,116	94.30
80 +	22,634	7.15	1708	7.55	20,926	92.45
Gender						
Female	179,381	56.66	12,826	7.15	166,555	92.85
Male	137,218	43.34	11,532	8.40	125,686	91.60
Education						
Education < 12 years	219,444	69.31	13,409	6.11	206,035	93.89
Employment						
Student	17,475	5.52	1574	9.01	15,901	90.99
Employed	131,019	41.38	10,725	8.19	120,294	91.81
Not employed	58,380	18.44	4946	8.47	53,434	91.53
Retired	63,324	20.00	3864	6.10	59,460	93.90%
Unknown	46,401	14.66	3249	7.00	43,152	93.00%
Race						
White	199,492	63.01	9742	4.88	189,750	95.12
American Indian|Alaska Native	4069	1.29	293	7.20	3776	92.80
Asian	13,334	4.21	1044	7.83	12,290	92.17
Black|African American	12,018	3.80	1095	9.11	10,923	90.89
Native Hawaiian | Pacific Islander	2700	0.85	424	15.70	2276	84.30
Hispanic | Latino	39,997	12.63	7962	19.91	32,035	80.09
Unknown	44,989	14.21	3798	8.44	41,191	91.56
Ethnicity						
Other ethnic groups	276,602	87.37	16,396	5.93	260,206	94.07
Hispanic or Latino	39,997	12.63	7962	19.91	32,035	80.09
Religious affiliation						
Agnostic	90,655	28.63	5585	6.16	85,070	93.84
Christian	121,557	38.39	10,293	8.47	111,264	91.53
Other religion	10,534	3.33	679	6.45	9855	93.55
Unknown	93,853	29.64	7801	8.31	86,052	91.69
Relationship						
Single	123,850	39.12	10,096	8.15	113,754	91.85
Divorced or legally separated	37,797	11.94	2412	6.38	35,385	93.62
Married or significant other	128,944	40.73	9817	7.61	119,127	92.39
Unknown	26,008	8.21	2033	7.82	23,975	92.18
Language						
English	288,252	91.05	18,964	6.58	269,288	93.42
Sino-Tibetan	2192	0.69	244	11.13	1948	88.87
Spanish	12,435	3.93	3679	29.59	8756	70.41
Other languages	13,720	4.33	1471	10.72	12,249	89.28

Table 1 (continued)

	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]
Clinical						
Body mass index						
Normal	66,179	20.90	4231	6.39	61,948	93.61
Underweight	5180	1.64	296	5.71	4884	94.29
Moderately obese	45,918	14.50	4061	8.84	41,857	91.16
Overweight	70,933	22.40	5918	8.34	65,015	91.66
Severely obese	23,334	7.37	2078	8.91	21,256	91.09
Very severely obese	19,981	6.31	1643	8.22	18,338	91.78
Unknown	85,074	26.87	6,131	7.21	78,943	92.79
Number of chronic conditions						
0	141,916	44.83	12,551	8.84	129,365	91.16
1–2	103,464	32.68	7629	7.37	95,835	92.63
3–4	46,632	14.73	2905	6.23	43,727	93.77
5 +	24,587	7.77	1273	5.18	23,314	94.82
Clinical diagnosis						
Diagnosis of diabetes	34,930	11.03	3340	9.56	31,992	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	95,207	91.31
1–9 prescriptions	160,387	50.66	12,403	7.73	147,984	92.27
10–19 prescriptions	38,656	12.21	2238	5.79	36,418	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						
History of illicit drug use	35,588	11.24	1561	4.39	34,027	95.61
History of tobacco use	40,352	12.75	1836	4.55	38,516	95.45
Diagnosis of serious persistent mental illness	30,246	9.55	1286	4.25	28,960	95.75
Diagnosis of substance use disorder	24,757	7.82	1071	4.33	23,686	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	88,060	27.81	8633	9.80	79,427	90.20
Symptoms						
Fever	101,388	32.02	15,157	14.95	86,231	85.05
Cough	113,047	35.71	16,319	14.44	96,728	85.56
Breath	107,216	33.86	13,642	12.72	93,574	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	83,293	26.31	5018	6.02	78,275	93.98
Alaska	17,269	5.45	857	4.96	16,412	95.04
Puget Sound	34,437	10.88	2144	6.23	32,293	93.77
Southern California	65,815	20.79	7389	11.23	58,426	88.77
Washington|Montana	115,589	36.51	8931	7.73	106,658	92.27

Table 1 (continued)

	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]
Unknown	196	0.06	19	9.69	177	90.31
Age-stratified communal living						
Non-communal living	230,410	72.78	16,624	7.21	213,786	92.79
Adult community	12,534	3.96	1055	8.42	11,479	91.58
Adult and youth	46,996	14.84	4460	9.49	42,536	90.51
Multigenerational	15,481	4.89	1535	9.92	13,946	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	98,537	31.12	10,285	10.44	88,252	89.56
Housing insecurity	72,081	22.77	8849	12.28	63,232	87.72
Transportation insecurity	88,401	27.92	7240	8.19	81,161	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: All models included symptoms as predictors except for trial 3. Except for the Light Gradient Boosting Machine model (LGBM), reported area under the receiver operating characteristic curve (AUC ROC) scores is 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

RFE = recursive feature elimination, *LR* = logistic regression, *CC* = case–control, *LGBM* = light gradient boosting machine, *PCA* = principal component analysis, *SMOTE* = synthetic minority oversampling technique

*Final selected model. This was the model that was used for the SHAP scores with symptoms presented in Fig. 2

**Final selected model without symptoms. This was the model that was used for the SHAP scores without symptoms presented in Fig. 3

4 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 USA 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 USA. Partnering with communities to assure broad availability of information and access

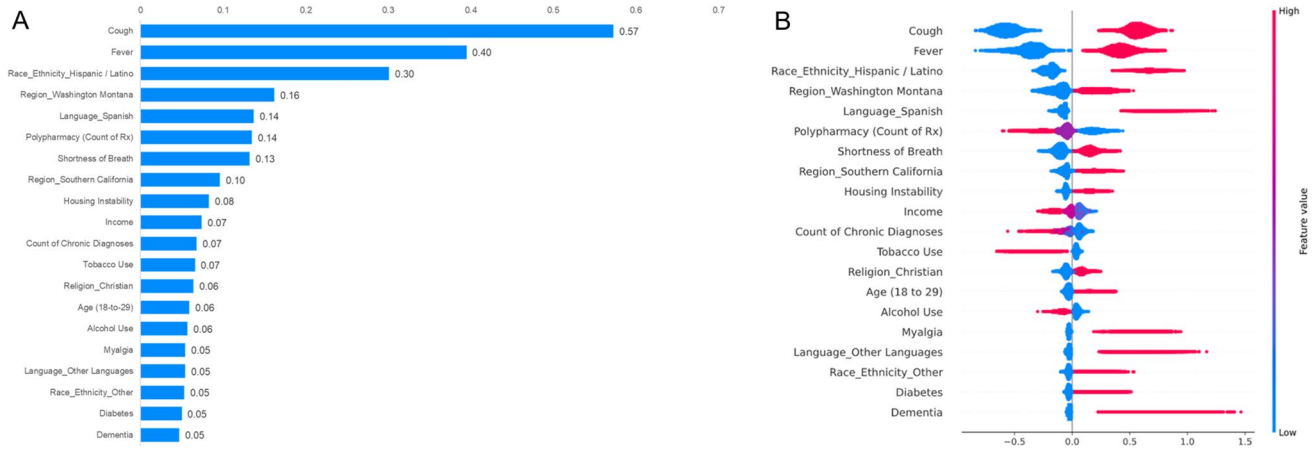

Fig. 2 Relative contribution of predictors in a machine learning model predicting COVID-19 infection based on symptoms and demographic information. Legend: **A** SHapley Additive exPlanations (SHAP) scores showing the average impact of each predictor on the model. SHAP values were computed using the final LGBM model. Higher SHAP values correspond to increased COVID-19 infection risk. **B** The relative importance of the top 20 COVID-19 predictors in descending order is shown here. The plot is made of dots corresponding to each prediction for a single patient. The horizontal axis shows the relative impact of a low or high prediction value for each variable, the impact ranging from blue (least associated with infection) to red (most associated with infection). Blue on the left to red on the right shows increasing infection risk as the feature increases (i.e., Cough: 0 = No Cough, 1 = Cough). Red on the left to blue on the right shows decreasing infection risk as the feature increases (i.e., polypharmacy)

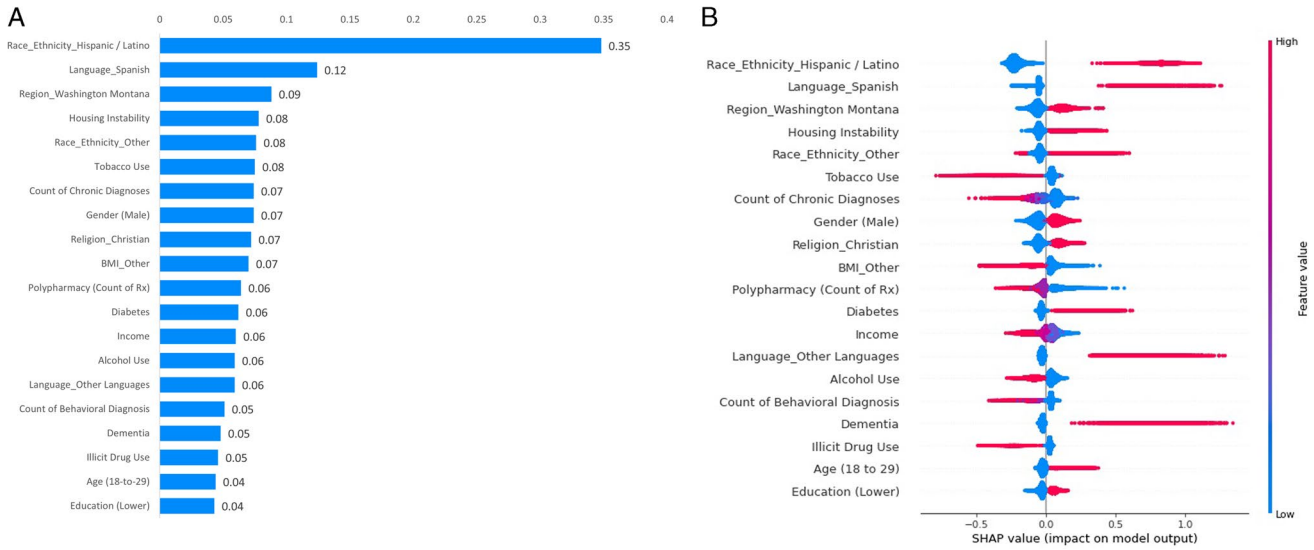

Fig. 3 Relative contribution of predictor variables in a machine learning model trained to predict COVID-19 infection based on demographic information alone. Legend: **A** SHAP scores showing the average impact of each predictor on the model using the final LGBM model. Higher SHAP values correspond to increased COVID-19 infection risk. **B** The top 20 COVID-19 demographic predictors, without symptoms, are shown here in descending order. All other computational and graphic elements (use of dots, color coding, variable score association strength shown by horizontal axis) are identical with those used for Fig. 2a and b

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.

We have previously published modeling work on this topic [2]. The previous model employed a logistic regression (LR) model and achieved an acceptable AUC of 0.78 on the validation set. It is important to note that features selected as strong predictors can be different across different machine and statistical learning approaches. This can be due to factors such as, but not limited to, penalization or regularization methods to reduce overfitting of the model. Other factors include how the model, such as decision tree-based models like LGBM, estimate information gained from all possible splits (using predictor values), different hyperparameters (e.g., tree sizes, number of subsamples, learning rate), etc.

Nevertheless, we computed a comparative logistic regression model and report the output of the model (see Supplementary Table 2). Variables with a $P < 0.25$ were considered for the final model consistent with the previous model [2]. This model was trained on 75% of data and validated on the remaining 25%. AUC on validation data was 0.80 slightly outperforming the previous logistic regression model (AUC = 0.78). Results from the LR and LGBM models are consistent with the previous model with respect to symptoms (cough, fever, shortness of breath, and myalgia), Hispanic or Latino racial/ethnic group, non-English language (specifically Spanish), having housing insecurity, age 18 to 29, and Washington-Montana and Southern California regions being more predictive, or "associated," with a positive COVID-19 test result. Likewise, having a history of tobacco use, higher number of prescription drugs and chronic conditions were more associated with a negative COVID-19 test — also consistent with the previous model (see Supplementary Table 2).

The new LGBM model was notably different from the previous LR model regarding age. There was a relatively small impact of being between ages 18 to 29 on the prediction of a positive test. The comparative new LR model is consistent with the previous model in that adults, 40 and older, have greater adjusted odds of contracting COVID-19 when compared to younger patients (reference group: ages 17 or younger in this LR model vs. 18 to 29 in the previous model). We also observed differences in the impact of existing comorbidities (e.g., diagnoses of diabetes, HIV/AIDS, dementia, and kidney disease) across models. The LGBM and the comparative new LR models do indicate some impact of an existing diagnosis of diabetes and dementia on the increased probability of a COVID-19 infection consistent with the previous model. Also consistent with the previous model is that the LR model shows some impact of having a history of kidney disease (OR 1.70; 95% CI 1.07–2.72, $p = 0.026$) on COVID-19 risk. Neither model, unlike the previous model, indicates that being immunocompromised (HIV/AIDS diagnosis) increases an individual's risk of an initial infection. Notwithstanding, we suspect that comorbidities will be significant predictors of severe illness or mortality *after* a COVID-19 infection.

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. The previous LR model was conducted on data available early in the pandemic between February 28, 2020, and April 27, 2020, with data ten times less than current data. 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 underrepresented 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 suspect that differences in results from current data reflect varied shifts in phased stay-at-home policies across the regions. Providence serves over time. Comparing results from both models is, nevertheless, encouraging as the new model demonstrates a stable and excellent ability to discriminate using new data as the previous model.

In the present study, we developed two predictive models that either included or excluded symptoms for different purposes. Modeling risk of infection without

symptoms was done to evaluate static risk for populations in the western USA. 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 a number of strengths to this study. We used advanced analytic procedures and tested a variety of models seeking the optimal solution. We have a very large data set (319,599 participants) collected across a single hospital system. Our very large data set gave us the statistical power to examine many possible influences on risk of infection simultaneously. The use of a single hospital system ensures that data collection, variable coding, and data extraction was done in a consistent manner, in contrast to meta-analyses and reviews which are forced to merge data sets which can have real methodological differences. Our list of examined variables is long and comprehensive, including age, gender, education, employment, race, ethnicity, religious affiliation, relationship status, language, BMI, chronic illness conditions, drug use, COVID-19 symptoms, geographic region, and living environment. Ours may be the only paper to date which has examined all of these variables, in a single hospital system, with > 300,000 participants.

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) [12] 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.

5 Conclusions

Our results confirm that the following social and demographic factors increased the risk of COVID-19 infection between February and 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

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11517-022-02549-5.

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.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Eisenstein M (2020) What's your risk of catching COVID? These tools help you to find out. June 15, 2021]; Available from: https://www.nature.com/articles/d41586-020-03637-y
2. Rozenfeld Y et al (2020) A model of disparities: risk factors associated with COVID-19 infection. Int J Equity Health 19(1):126
3. Zheng Z et al (2020) Risk factors of critical & mortal COVID-19 cases: a systematic literature review and meta-analysis. J Infect 81(2):e16–e25
4. Wolff D et al (2021) Risk factors for Covid-19 severity and fatality: a structured literature review. Infection 49(1):15–28

5. Syeda HB et al (2021) Role of machine learning techniques to tackle the COVID-19 crisis: systematic review. JMIR Med Inform 9(1):e23811
6. Dogan O, Tiwari S, Jabbar MA, Guggari S (2021) A systematic review on AI/ML approaches against COVID-19 outbreak. Complex Intell Systems 7(5):2655–2678. https://doi.org/10.1007/s40747-021-00424-8
7. Malik AA et al (2020) Determinants of COVID-19 vaccine acceptance in the US. EClinicalMedicine 26:100495
8. Khubchandani J et al (2021) COVID-19 vaccination hesitancy in the United States: a rapid national assessment. J Community Health 46(2):270–277
9. Brownlee J (2020) SMOTE for imbalanced classification with Python. machine learning mastery. Available from: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification
10. Lundberg S (2018) Welcome to the SHAP documentation — SHAP latest documentation. SHAP. Available from: https://shap.readthedocs.io/en/latest/
11. Molnar C (2020) 5.10 SHAP (SHapley Additive exPlanations) | Interpretable machine learning. SHAP (SHapley Additive exPlanations). [cited 2021; Available from: https://christophm.github.io/interpretable-ml-book/shap.html
12. Mickael T, Ahmed M, Rahul MD, Ines S, Guillaume C, Enora G, Robert B, Deborah E, Sebastien B, Guillaume M, Simon B, Antoine F, Fadila M, Benoit D, Robert-Yves C, Luc DJ, Marie-Pierre R (2020) Pre-test probability for SARS-Cov-2-related infection score: the PARIS score. medRxiv 2020.04.28.20081687. https://doi.org/10.1101/2020.04.28.20081687

Elijah A. Adeoye , MPH, Senior Data Scientist for Population Health Care Management with Providence St. Joseph Health (PSJH). He has worked in machine learning and modeling for 5 years.

Yelena Rozenfeld , MPH, Director of Analytics and Data Science for PSJH, has 20 years of experience in statistical analysis, predictive modeling, and leading operational and research projects.

Jennifer Beam , MSSW, MSIE, Vice President of Analytics and Practice Optimization for Population Health Administration in PSJH, is a seasoned analytic leader with over 20 years of industry experience.

Karen Boudreau , MD, Senior Vice President for Enterprise Care Management with Population Health, PSJH, has worked in primary care, population health, and care management for 32 years.

Emily J. Cox , PhD, contributed to the writing of the manuscript. She has over 5 years of experience in statistics and scientific writing.

James M. Scanlan , PhD, contributed to the writing of the manuscript. He has over 30 years of experience in statistics and scientific writing.