



# COVID-19 Infection Risk Among Previously Uninfected Adults: Development of a Prognostic Model

Richard Sloane<sup>1</sup> , Carl F Pieper<sup>1,2</sup>, Richard Faldowski<sup>1</sup>, Douglas Wixted<sup>3</sup>, Coralei E Neighbors<sup>4</sup> , Christopher W Woods<sup>4,5</sup>, and L Kristin Newby<sup>3,6,7</sup>

## Abstract

**Background:** Few models exist that incorporate measures from an array of individual characteristics to predict the risk of COVID-19 infection in the general population. The aim was to develop a prognostic model for COVID-19 using readily obtainable clinical variables.

**Methods:** Over 74 weeks surveys were periodically administered to a cohort of 1381 participants previously uninfected with COVID-19 (June 2020 to December 2021). Candidate predictors of incident infection during follow-up included demographics, living situation, financial status, physical activity, health conditions, flu vaccination history, COVID-19 vaccine intention, work/employment status, and use of COVID-19 mitigation behaviors. The final logistic regression model was created using a penalized regression method known as the least absolute shrinkage and selection operator. Model performance was assessed by discrimination and calibration. Internal validation was performed via bootstrapping, and results were adjusted for overoptimism.

**Results:** Of the 1381 participants, 154 (11.2%) had an incident COVID-19 infection during the follow-up period. The final model included six variables: health insurance, race, household size, and the frequency of practicing three mitigation behavior (working at home, avoiding high-risk situations, and using facemasks). The c-statistic of the final model was 0.631 (0.617 after bootstrapped optimism-correction). A calibration plot suggested that with this sample the model shows modest concordance with incident infection at the lowest risk.

**Conclusion:** This prognostic model can help identify which community-dwelling older adults are at the highest risk for incident COVID-19 infection and may inform medical provider counseling of their patients about the risk of incident COVID-19 infection.

## Keywords

COVID-19, infection, prediction model

<sup>1</sup> Center for the Study of Aging and Human Development, Duke University Medical Center, Durham, North Carolina, USA

<sup>2</sup> Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, North Carolina, USA

<sup>3</sup> Duke Clinical and Translational Science Institute, Duke University, Durham, North Carolina, USA

<sup>4</sup> Hubert-Yeargan Center for Global Health, Duke University, Durham, North Carolina, USA

<sup>5</sup> Departments of Medicine and Pathology, Duke University Medical Center, Durham, North Carolina, USA

<sup>6</sup> Duke Clinical Research Institute, Duke University, Durham, North Carolina, USA

<sup>7</sup> Division of Cardiology, Department of Medicine, Duke University Medical Center, Durham, North Carolina, USA

Submitted November 15, 2022. Revised January 7, 2023. Accepted January 14, 2023.

## Corresponding Author:

Richard Sloane, Center for the Study of Aging and Human Development, Box 3003, Duke University Medical Center, Durham, NC 27710, USA.

Email: richard.sloane@duke.edu



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

## Introduction

As of December 2022, it is estimated there have been over 650,000,000 COVID-19 infections from the SARS-CoV-2 virus worldwide and almost 100,000,000 in the USA (<https://covid19.who.int/> accessed 29 December 2022).<sup>1</sup> Researchers worldwide have been collecting population-based data from individuals who are uninfected with COVID-19 along with infected people and tracking their status over time to assess COVID-19 prevalence and incidence rates. As COVID-19 variants develop, the dynamics of the infection process change and public health officials and clinicians are faced with estimating the risk for COVID-19 infection.<sup>2</sup>

At the outset of the COVID-19 pandemic beginning in early 2020, public health agencies developed and disseminated guidelines for reducing the risk of infection. Prior to and early in the distribution of COVID-19 vaccines in the United States in December 2020–January 2021<sup>3</sup> Centers for Disease Control (CDC) issued guidelines comprised of ten behavioral measures including masking, social distancing, working at home, staying home, avoiding crowds, washing hands, avoiding high-risk situations, avoiding restaurants, avoiding touching people, and wiping surfaces.<sup>4</sup> On a global basis, similar behavioral and nonpharmaceutical guidelines were studied and recommended by public health organizations<sup>1,5,6</sup> and other investigators.<sup>7–9</sup> The efficacy of such behavioral and nonpharmaceutical interventions in the time frame previous to the initiation of the rollout and widespread availability of vaccines in the United States and Europe were largely found to be the most useful tools to attenuate the rate of Covid-19 infection.<sup>10–16</sup> Community studies of adherence to behavioral and nonpharmaceutical interventions contributed to building an evidence base about the uptake of the nonvaccine guidelines.<sup>9,12,17–27</sup>

Utilizing the power of extant databases (eg, formal registries, electronic health records, or various types of surveys), data of all types can be fed into programs that model the risk of a given outcome, such as hospitalization or death<sup>28,29</sup>. However, few models exist that incorporate measures from an array of individual characteristics, including clinical and socio-demographic features and use of CDC-recommended protective behaviors, among other items, to predict the risk of COVID-19 infection in the general previously uninfected population. Ultimately clinicians can translate the predictions to effectively counsel patients as to their risk with respect to modifiable risk factors. To address this issue, we developed a prognostic model for incident COVID-19 infection among uninfected participants in a longitudinal, community-based cohort study.

## Methods

### Study Sample

All data from this study came from the Cabarrus County COVID-19 Prevalence and Immunity (C3PI) Study. The C3PI Study was a community COVID-19 surveillance study

that enrolled 1410 individuals from the Measurement to Understand the Reclassification of Disease of Cabarrus/Kannapolis (MURDOCK) Study Community Registry and Biorepository longitudinal cohort<sup>30,31</sup> and was conducted in North Carolina by Duke University with funding from the North Carolina Department of Health and Human Services (NCDHHS). The design and methods of this study were published previously.<sup>32</sup> Briefly, each participant completed a baseline survey covering demographics, current health status, household features, lifestyle, and employment and their perceptions of the COVID-19 pandemic, use of COVID-19 mitigation behaviors, and attitudes about COVID-19 vaccination. Follow-up after the baseline survey occurred on a biweekly basis for up to 74 weeks. Of all participants, 29 (2.1%) did not respond to most or all of the use of COVID-19 mitigating behavior items (see Measures) and were dropped from the analysis. This left 1381 (97.9%) that were ultimately included in this analysis with an age range from 24 to 98 years. Over the course of the follow-up period, there were 154 COVID-19 infections. COVID-19 infections were identified from C3PI Study-related testing in a subset of 300 individuals in a COVID-19 testing subcohort or from self-report of infection in the biweekly surveys.

### Measures

The baseline survey comprised nearly 370 unique items querying a variety of dimensions before a vaccine for COVID-19 was available (**Appendix 1: the baseline survey**). For the current analysis, a subset of representative items was identified by the authors, and in some cases, summary variables were constructed:

- Demographics
- Living situation
- Financial status
- Physical activity
- Health conditions
- Flu vaccination history
- COVID-19 vaccine intention
- Work/employment status
- Use of COVID-19 mitigation behaviors from CDC.<sup>4</sup>

After a data reduction, data quality, and adjudication decision-making process, 34 separate items were included in the analysis (see Table 2).

### Statistical Analysis

Descriptive statistics comparing participants with and without incident COVID-19 infections during the follow-up period were presented as means (*SD*) for continuous variables or frequencies for categorical variables. To develop our prediction model, we included the selected variables described above in a multivariable logistic regression<sup>33</sup> with any infection during the follow-up period as the outcome variable. Missing values were imputed using single-imputation maximum

likelihood estimation. We then developed a parsimonious model by using a penalized regression method known as the least absolute shrinkage and selection operator (LASSO)<sup>34</sup> specifying 500 bootstrap samples and included those predictors that were retained in more than 10% of the bootstrapped samples<sup>35</sup> for the final logistic model. There are several methods for selecting a set of independent variables for use to develop the “best” regression model, but some such as the family of methods known as stepwise regression have several problems including inflated  $R^2$  values, invalid F and Chi-square distributions, underestimated standard errors and confidence intervals with resultant too small  $p$ -values, inflated parameter estimates, and exacerbated issues around multicollinearity.<sup>33</sup> LASSO selection arises from a constrained form of ordinary least squares regression where the sum of the absolute values of the regression coefficients is constrained to be smaller than a specified parameter.<sup>34</sup> Using conventional methods, model performance was assessed through discrimination and calibration.

Discrimination of the logistic models refers to the ability of the model to separate individuals who develop an infection from those who do not.<sup>36</sup> Calibration refers to the graphical association between the observed risk of infection and predicted risk. This was visually assessed via a calibration plot with the  $x$ -axis displaying the predicted estimate from the model and the  $y$ -axis displaying the observed proportion of infection.<sup>37</sup> Internal validation was assessed by evaluating the c-statistic of results of bootstrapped samples.<sup>33</sup> Estimates from the final model were used to develop several hypothetical scenarios based on a participant’s specific characteristics in order to illustrate the risk of infection. All statistical analyses were performed using SAS statistical software, version 9.4 (SAS Institute, Inc, Cary, NC).

## Ethics

Both the parent MURDOCK Study (Approval Number: Pro00011196) and Phase 1 and 2 of the C3PI Study (Approval Number: Pro00105703) were approved by the Duke Health Institutional Review Board. Participants provided electronic informed consent within REDCap<sup>®</sup> to participate in the C3PI Study.

## Results

### Baseline Characteristics

The baseline characteristics of the 1381 participants included in this study are summarized in Table 1. Of the total sample, 154 (11.2%) participants reported or tested positive for a COVID-19 infection.

For most of the demographic, living situations, employment, health conditions, and attitudes and behaviors associated with the COVID-19 pandemic, the infected and noninfected groups were similar in their characteristics (Table 1). The group with an infection was on average about 2.1 years

younger ( $p = .054$ ), had a lower proportion of Whites ( $p = .005$ ), a higher proportion with Hispanic ethnicity ( $p = .03$ ), reported more people living in their household ( $p = .05$ ), and a lower proportion who had any health insurance coverage ( $p < .0001$ ). Of the CDC-recommended COVID-19 mitigation behaviors, compared with the not-infected group the infected group reported lower average frequencies of practicing using facemasks ( $p = .03$ ), maintaining six-foot distances ( $p = .04$ ), avoiding high-risk situations ( $p = .008$ ), working at home ( $p = .002$ ), and avoiding touching people ( $p = .04$ ).

### Modeling COVID-19 Infection Using LASSO

The candidate variables were analyzed for predicting COVID-19 infection using LASSO. The final model retained six variables (Table 2): Any current health insurance coverage (yes or no); Race (white, black, or other); the number of people living in the household; and the frequency of practicing three CDC-recommended mitigation behavior (working at home, avoiding high-risk situations, and using facemasks). The c-statistic for this model was 0.631. After bootstrap internal validation, the optimism-corrected estimate of the c-statistic was 0.617. A calibration plot is shown in Figure 1. Based on the parameter values from the final logistic model, we present a set of estimates for risk of infection for a set of hypothetical patients with different baseline characteristics (Table 3) and graphically in Figure 2.

## Discussion

Using the LASSO method in a community-based cohort we developed a prognostic model for incident COVID-19 infection during one-year follow-up surveillance after the baseline assessment. In our final model the effects of not having current insurance, identifying as non-White race, having four or more people in the household, and not frequently using a face mask significantly contributed to the prognostic information for incurring an incident COVID-19 infection, while not working from home fell just out of the range of statistical significance.

As noted above, the set of items captured in the baseline survey (**Appendix 1: the baseline survey**) was wide-ranging and included candidate risk factors prevalent in the current literature of the time as well as the behavioral and nonpharmaceutical interventions, for example.<sup>5,7</sup> Many of these variables included in the final model are consistent with previous literature that employed LASSO to process a wide selection of epidemiologic variables to model incident COVID-19 infection, particularly those related to race, number of people in the household, and level of compliance with some of the CDC mitigation behavior guidelines.<sup>38</sup> While in that study many of the same or similar demographic characteristics were included in the initial set of variables, the most informative one in our final model, health insurance coverage, was not included although it is likely that insurance status is associated as a marker or proxy for social deprivation, poorer health, and

**Table 1.** Baseline Characteristics of Participants by Infection status During Follow-up.

	Infected (n = 154)	Not infected (n = 1227)	p-value
Age, mean (SD)Range	58.3 (12.9)24–98	60.4 (12.5)26–91	0.054
BMI, mean (SD)	29.5 (6.7)	28.6 (6.1)	0.12
Sex, % female	68.8	69.2	0.92
Race (%)			
White	77.9	87.2	0.005
Black	11.7	7.7	
Other	10.4	5.1	
Hispanic ethnicity	8.4	4.4	0.026
Annual income < 50,000 (%)	18.6	21.9	0.36
Education: Less than college grad (%)	31.4	28.7	0.48
Number living in household, mean (SD)	2.7	2.5	0.05
Any people under 18 years old living in household, %	30.9	24.1	0.07
Covered by some type of insurance (%)	90.2	97.5	<.0001
How serious a problem is the COVID-19 pandemic (from 1 to 4), mean (SD):			
For you personally?	2.18 (0.98)	2.13 (0.97)	0.48
For people in your community?	1.86 (0.87)	1.81 (0.90)	0.50
For people in the United States?	1.68 (0.85)	1.57 (0.83)	0.14
For people of the world?	1.69 (1.02)	1.56 (0.87)	0.08
Self-rated health pre-COVID-19, 1 = excellent 5 = poor, mean (SD)	2.21 (0.97)	2.20 (0.87)	0.94
Self-rated health over the past two weeks, 1 = excellent 5 = poor, mean (SD)	2.29 (0.89)	2.22 (0.86)	0.35
Has your work situation changed since the COVID-19 pandemic began in North Carolina? (% Yes)	30.1	26.5	0.34
Employment status pre-COVID-19 (%)			0.22
Full-time	44.8	42.4	
Part-time	12.3	8.8	
Unemployed	0	1.3	
Retired	28.6	35.8	
Homemaker	7.8	6.9	
Other	6.5	4.9	
Received flu vaccine past year (%)	75.3	79.2	0.27
Intend to get vaccine for COVID-19 when one becomes available (%)			
Yes	57.5	61.8	0.56
No	9.2	7.6	
Do not know	33.3	30.6	
Engage in moderate physical exercise on a weekly basis (%)	72.1	72.6	0.89
Currently smoke tobacco (%)	2.7	3.8	0.46
Ever smoke tobacco (%)	26.0	29.9	0.31
Currently consume alcohol at least once a week (%)	45.8	43.5	0.59
CDC recommended COVID-19 mitigating Behaviors, 1 = never 5 = always, mean (SD)			
Use facemask	4.09 (1.15)	4.28 (1.03)	0.03
Washed hands	4.71 (0.54)	4.70 (0.57)	0.91
Maintain six feet	4.28 (0.70)	4.40 (0.62)	0.04
Avoid crowds	4.60 (0.71)	4.69 (0.63)	0.14
Avoid high-risk situations	4.49 (0.90)	4.65 (0.67)	0.008
Avoid restaurant food	3.16 (1.24)	3.22 (1.24)	0.57
Work at home	3.22 (1.74)	3.67 (1.62)	0.002
Avoid touching people	4.52 (0.85)	4.64 (0.69)	0.04
Staying home	4.78 (0.73)	4.85 (0.66)	0.22
Wiped surfaces	4.26 (1.02)	4.19 (0.95)	0.36

inadequate resources.<sup>39</sup> Liang et al.<sup>28</sup> used a cohort infected with COVID-19 and constructed a calculator to estimate the risk of critical illness, and used primarily clinical characteristics upon presentation at hospital admission to calculate the risk

score. The objective of our study was to develop an algorithm from a model as a proof-of-concept that given a sufficiently large evidence base a valid prognostic model can be used to inform a simple risk calculator so that clinicians can effectively

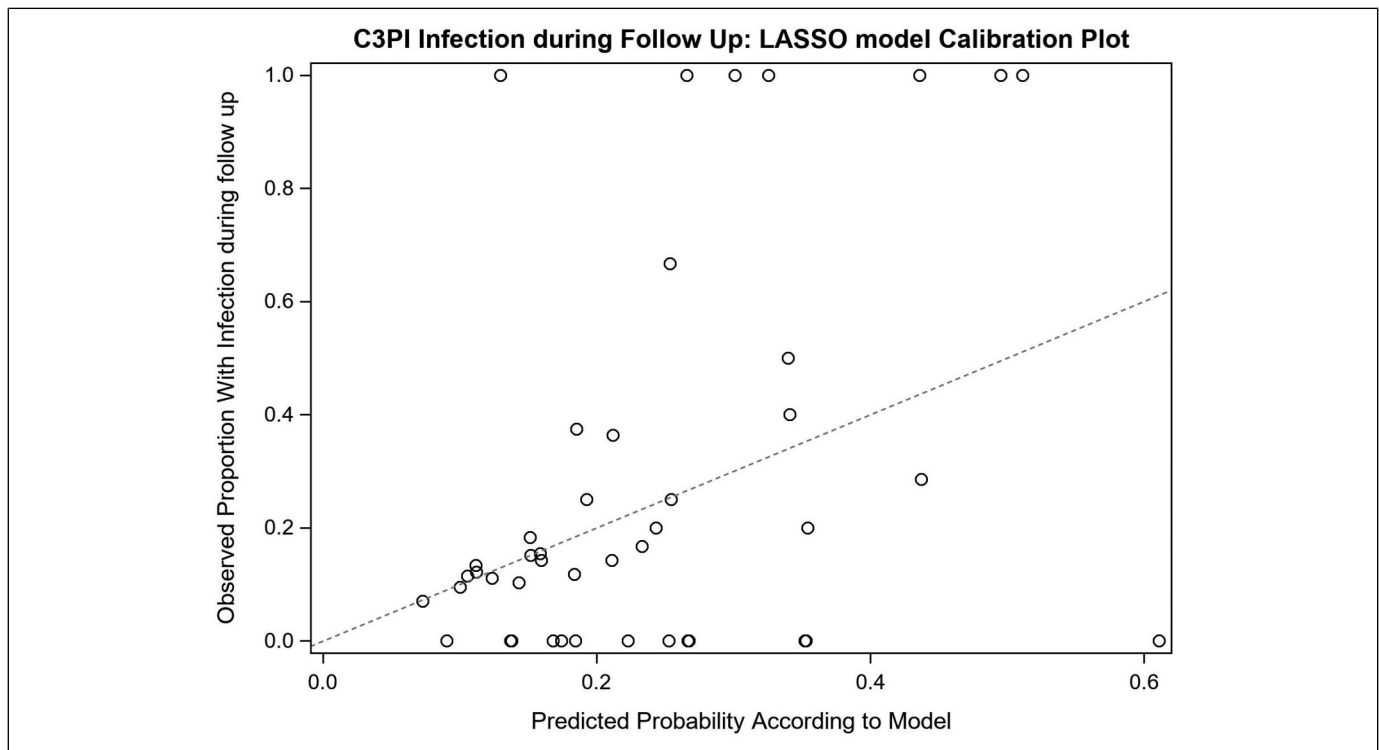
**Table 2.** Multivariable-adjusted prognostic factors assessed at baseline for incident COVID-19 infection during follow-up after penalized least absolute shrinkage and selection operator selection.

	Odds ratio	95% CI
Not having current insurance	2.89	1.46–5.72
Did not frequently work at home (dichotomized) <sup>ψ</sup>	1.42	0.99–2.03
Non-White race	1.61	1.03–5.23
Not frequently avoiding high-risk situations (dichotomized) <sup>ψ</sup>	1.26	0.63–2.55
Number in house $\geq 4$	1.50	1.01–2.23
Not frequently using face mask (dichotomized) <sup>ψ</sup>	1.60	1.04–2.46
c-statistic = 0.631		
Optimism-corrected c-statistic after bootstrap validation = 0.617		

<sup>ψ</sup>CDC recommended mitigating behaviors were dichotomized as: 1, 2, 3 = Not frequently practicing behavior, 4 or 5 = frequently practicing the behavior

<sup>φ</sup>Referent group was White race

Abbreviations: c-statistic, concordance statistic (or area under the receiver operating characteristic curve); CDC: Centers for Disease Control.

**Figure 1.** LASSO calibration plot.

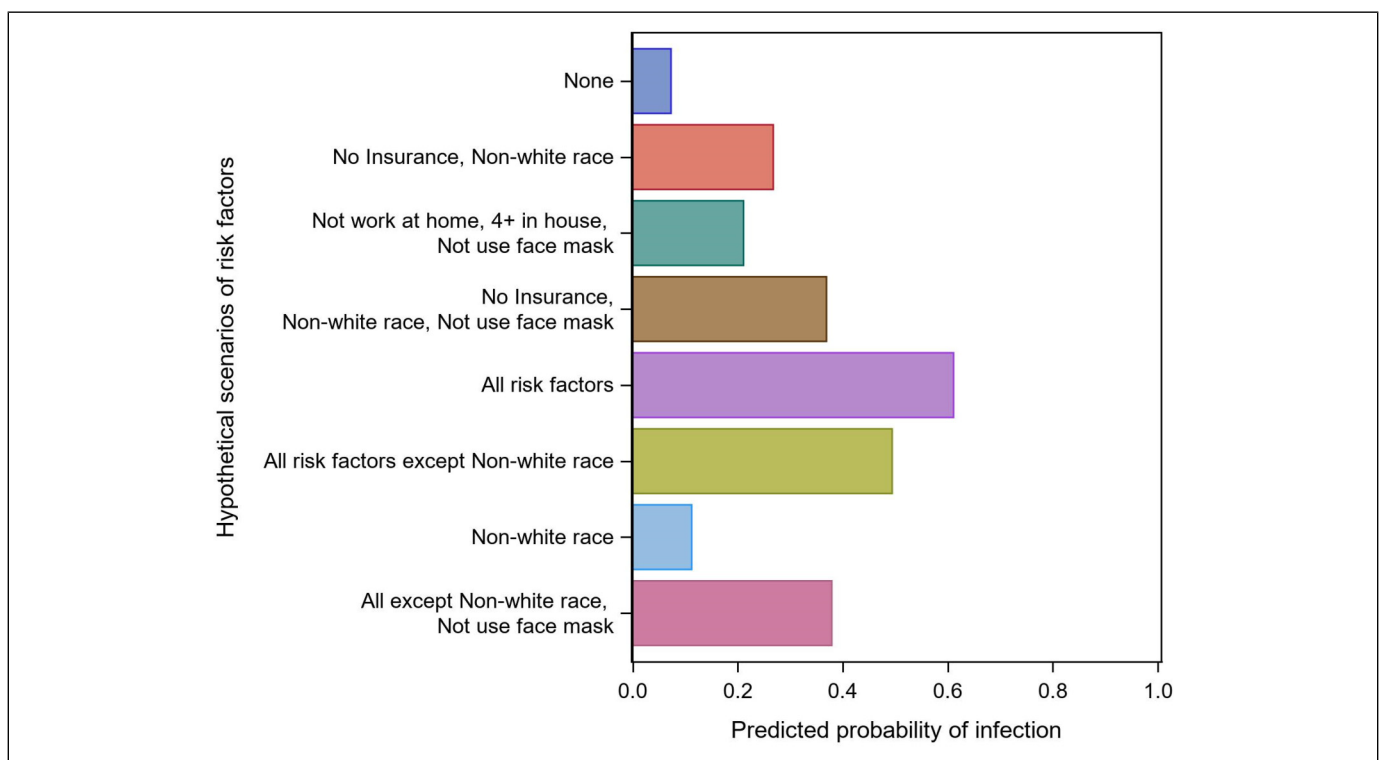
counsel patients as to their risk with respect to modifiable risk factors. On a cumulative basis, the administration of vaccine doses exceeds 13 billion<sup>1</sup> but the global distribution of those doses is highly variable,<sup>40</sup> leaving large groups of individuals as unvaccinated or partially vaccinated. While the vaccines have been shown to be effective<sup>40,41</sup> the antibodies for the Omicron variants can wane substantially over time rendering fully vaccinated and boosted individuals protection against hospitalization and death but vulnerable to asymptomatic and symptomatic infection.<sup>42</sup> For the groups of currently unvaccinated individuals prognostic models similar to the one presented here have value for the clinician offering counseling

regarding modifiable behaviors such as using facemasks, avoiding high-risk situations, working at home, and other nonpharmaceutical behavioral measures. The prognostic models can be revised to estimate risk of Covid-19 infection for the partially and fully vaccinated in the context of waning effectiveness of administered doses of vaccine.

Our final model included variables dichotomized as “yes/no,” which could be assessed using a short questionnaire in a clinician’s office and potentially entered into a web- or computer-based calculator utility. For example, using Table 3 hypothetical patient examples, the model estimates that the non-COVID-19 infected patient with all the risk factors has a

**Table 3.** Hypothetical patient examples with the prediction model's calculated predicted probability of COVID-19 infection.

Not having current insurance	Not work at home	Non-White race	Not avoid high-risk situations	Number in house >= 4	Do not frequently use face mask	Estimated risk of incident COVID-19 infection	95% CI for risk estimate	
							Lower	Upper
-	-	-	-	-	-	0.07	0.06	0.09
-	-	X	-	-	-	0.11	0.07	0.17
-	X	-	-	X	X	0.21	0.14	0.31
X	-	X	-	-	-	0.27	0.15	0.43
-	X	-	-	-	X	0.37	0.20	0.58
X	X	-	X	X	-	0.38	0.19	0.62
X	X	X	X	X	-	0.50	0.28	0.71
X	X	X	X	X	X	0.61	0.38	0.80

**Figure 2.** Predicted probability of infection using hypothetical scenarios from final model.

61% risk of an incident COVID-19 infection at some point in the next year. The patient with all the risk factors except “not frequently using face masks” has about a 50% risk, so all other things being equal, in that case not frequently using face-masks adds about 10–11% risk. Similar types of added risk scenarios can help clinicians counsel their patients in practicing these modifiable risk behaviors. The model had modest discrimination, with a c-statistic of 0.631 (optimism corrected to 0.617).<sup>43</sup> Rozenfeld et al.<sup>43</sup> developed a more discriminating risk factor model (c-statistic 0.78) for COVID-19 infection using a much larger sample of over 34,000, but did not include variables with modifiable risk factors. Among a large cohort of nursing home patients, Mehta et al. found<sup>29</sup> significant

associations of incident COVID-19 infection, including non-White race and other demographic and facility-level factors, but did not construct a multivariable model or risk calculator.

There are a few important limitations to the construction of this prognostic model. First, while the rate of incident infections was in the range reported at the time the database was active, the number of infections was not high enough to provide sufficient statistical power to accurately classify many individuals to the correct true infection status (see the calibration plot, Figure 1). Although it appears that the estimated infection projection below 25% has a modest agreement with the observed infection rate, at higher estimated levels the agreement is low.

With a larger database that offers enhanced discriminatory power it is likely that the calibration between estimated and observed infection rates can be improved. Therefore, the model discrimination, as noted above, was modest. Also, it is common that in the LASSO procedure, variables retained in the final models are developed from being present in at least 50% of the models, while we used 10% as the threshold, mostly for illustrative purposes. Further, most of the incident infections in our analysis were derived from self-report from follow-up surveys rather than from serology databases, which could have led to under- or over-identification of true infections; thus, affecting both power and potentially introducing error into the estimate of associations with infection. However, the surveys did collect the date of diagnosis and method of diagnosis to add stringency to this self-report question. Finally, during the follow-up period there was 15.9% early termination (5.8% in the infected group and 17.2% in the non-infected group) so there are likely incident infection cases that were missed from the final count.

## Conclusion

In summary, the COVID-19 pandemic has persisted since early 2020 and as new variants develop, infections affect patient health status and contribute greatly to healthcare costs and utilization. Our model offers proof of concept that the risk of incident COVID-19 infection can be estimated using a limited number of predictor variables. We believe this prognostic information could provide foundational information to develop tool clinicians and patients can use to assess risk, guide clinical decisions and inform the allocation of healthcare resources.

## Acknowledgments

The authors thank all participants in the C3PI Study for their involvement in and enthusiasm for the study. We also thank Dr Aaron Fleischauer, Career Epidemiology Field Officer, CDC; Jie Liu, Senior Analyst Programmer, Duke Clinical and Translational Science Institute; Angie Wu, Senior Biostatistician, Department of Clinical Research, Cytel Inc. for their engagement and support throughout the C3PI Study.

## Author Contributions

R.S. and C.F.P. contributed to the study conception and design. R.S. performed the data analysis and wrote the initial version of the manuscript. C.F.P. and R.F. provided statistical advice. C.E.N., D.W., and R.F. collaborated on data management and construction of the foundational analytic data file. D.W. and L.K.N. provided critical input on data items for the preliminary variable screening process. D.W., C.W.W., and L.K.N. provided material and administrative support. All authors contributed to the critical revision of the manuscript for important intellectual content.


## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

This work was funded by a gift from the David H. Murdock Institute for Business and Culture and is supported by Duke's NIH National Center for Advancing Translational Sciences (NCATS) Clinical and Translational Science Award (CTSA) (UL1TR002553). The C3PI Study was funded by research grants to Duke University from the NCDHHS and the Center for Disease Control and Prevention (CDC), and the Duke Claude D. Pepper Older Americans Independence Center Grant (5P30AG028716-15).

## ORCID iDs

Richard Sloane  <https://orcid.org/0000-0001-6604-8101>

Coralei E Neighbors  <https://orcid.org/0000-0002-0367-2983>

## Supplemental Material

Supplemental material for this article is available online.

## References

1. WHO. WHO Coronavirus (COVID-19) Dashboard. 29 December 2022].
2. Otto SP, Day T, Arino J, et al. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. *Curr Biol*. 2021;31(14):R918-R929.
3. U HHS. *COVID-19 Vaccines*.
4. CDC. *Coronavirus disease 2019 (COVID-19): how to protect yourself and others*. August 11, 2022 [cited 2020 October]; Available from: <https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html>.
5. WHO. *Considerations for implementing and adjusting public health and social measures in the context of COVID-19: interim guidance, 14 June 2021*. World Health Organization. . 2021 29 December 2022].
6. WHO. *COVID-19: Occupational health and safety for health workers: interim guidance, 2 February 2021*. 2021 29 December 2022].
7. Chan EYY, Shahzada TS, Sham TST, et al. Narrative review of non-pharmaceutical behavioural measures for the prevention of COVID-19 (SARS-CoV-2) based on the Health-EDRM framework. *Br Med Bull*. 2020;136(1):46-87.
8. Fletcher KM, Espey J, Grossman MK, et al. Social vulnerability and county stay-at-home behavior during COVID-19 stay-at-home orders, United States, April 7-April 20, 2020. *Ann Epidemiol*. DEC 2021;64:76-82.
9. Tso RV, Cowling BJ. Importance of face masks for COVID-19: a call for effective public education. *Clin Infect Dis*. 2020;71(16): 2195-2198.
10. Ali ST, Wang L, Lau EHY, et al. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. *Science*. 2020;369(6507):1106.
11. Camacho-Rivera M, Islam JY, Vidot DC. Associations between chronic health conditions and COVID-19 preventive behaviors among a nationally representative sample of U.S. adults: an analysis of the COVID impact survey. *Health Equity*. 2020;4(1): 336-344.

12. Bekalu MA, Dhawan D, McCloud R, Pinnamaneni R, Viswanath K. Adherence to COVID-19 mitigation measures among American adults: the need for consistent and unified messaging. *Health Educ Res.* 2021;36(2):178-191.
13. Czeisler ME, Howard ME, Robbins R, et al. Early public adherence with and support for stay-at-home COVID-19 mitigation strategies despite adverse life impact: a transnational cross-sectional survey study in the United States and Australia. *BMC Public Health.* 2021;21(1):503.
14. Lio CF, Cheong HH, Lei CI, et al. Effectiveness of personal protective health behaviour against COVID-19. *BMC Public Health.* 2021;21(1):827.
15. Lyu W, Wehby GL. Community use of face masks and COVID-19: evidence from a natural experiment of state mandates in the US. *Health Aff.* 2020;39(8):1419-1425.
16. Margraf J, Brailovskaia J, Schneider S. Behavioral measures to fight COVID-19: an 8-country study of perceived usefulness, adherence and their predictors. *Plos One.* 2020;15(12):e0243523. <https://doi.org/10.1371/journal.pone.0243523>
17. Block RJr, Berg A, Lennon RP, Miller EL, Nunez-Smith M. African American adherence to COVID-19 public health recommendations. *Health Lit Res Pract.* 2020;4(3):e166-e170.
18. Alsan M, Stantcheva S, Yang D, Cutler D. Disparities in coronavirus 2019 reported incidence, knowledge, and behavior among US adults. *JAMA Netw Open.* 2020;3(6):e2012403.
19. Badr HA, Oluoyomi L, Woodard X, et al. Sociodemographic and health belief model factors associated with Nonadherence to COVID-19 mitigation strategies in the United States. *Ann Behav Med.* 2021;55(7):677-685.
20. Czeisler ME, Tynan MA, Howard ME, et al. Public attitudes, behaviors, and beliefs related to COVID-19, stay-at-home orders, nonessential business closures, and public health guidance - United States, New York City, and Los Angeles, May 5-12, 2020. *MMWR Morb Mortal Wkly Rep.* 2020;69(24):751-758.
21. Fischer CB, Adrien N, Silguero JJ, Hopper JJ, Chowdhury AI, Werler MM. Mask adherence and rate of COVID-19 across the United States. *PLoS One.* 2021;16(4):e0249891.
22. Hill LM, Davis H, Drewry M, et al. Barriers to and facilitators of COVID-19 prevention behaviors among North Carolina residents. *Health Educ Behav.* 2022;49(2):231-241.
23. Pei S, Kandula S, Shaman J. Differential effects of intervention timing on COVID-19 spread in the United States. *Sci Adv.* 2020;6(49):Pages eabd6370.
24. Hutchins HJ, Wolff B, Leeb R, et al. COVID-19 mitigation behaviors by age group - United States, April-June 2020. *MMWR Morb Mortal Wkly Rep.* 2020;69(43):1584-1590.
25. Talic SS, Shah H, Wild D, et al. Effectiveness of public health measures in reducing the incidence of COVID-19, SARS-CoV-2 transmission, and COVID-19 mortality: systematic review and meta-analysis. *Br Med J.* Nov 17 2021;375:e068302.
26. Wong R, Lovier MA. Relationship between dementia, COVID-19 risk, and adherence to COVID-19 mitigation behaviors among older adults in the United States. *Int J Geriatr Psychiatry.* 2022;37(6):1-9.
27. Tomczyk S, Rahn M, Schmidt S. Social distancing and stigma: association between compliance with behavioral recommendations, risk perception, and stigmatizing attitudes during the COVID-19 outbreak. *Front Psychol.* Aug 11 2020;11:1821.
28. Liang WH, Liang L, Ou B, et al. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Intern Med.* 2020;180(8):1081-1089.
29. Mehta HB, Li S GJ. Risk factors associated with SARS-CoV-2 infections, hospitalization, and mortality among US nursing home residents. *JAMA Netw Open.* 2021;4(3):e216315.
30. Bhattacharya SAA, Dunham MA, Cornish VA, et al. The measurement to understand reclassification of disease of Cabarrus/Kannapolis (MURDOCK) study community registry and biorepository. *Am J Transl Res.* 2012;4(4):458-470.
31. Tenenbaum JD, Christian V, Cornish MA, et al. The MURDOCK study: a long-term initiative for disease reclassification through advanced biomarker discovery and integration with electronic health records. *Am J Transl Res.* 2012;4(3):291-301.
32. Neighbors CE, Wu AE, Wixted DG, et al. The Cabarrus county COVID-19 prevalence and immunity (C3PI) study: design, methods, and baseline characteristics. *Am J Transl Res.* 2022;14(8):5693-5711.
33. Harrell F. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis.* 2 ed. Springer; 2015.
34. Tibshirani R. Regression shrinkage and selection via the LASSO. *Stat Soc Series B Stat Methodol.* 1994 No. 1;58:267-288.
35. Austin PT, Tu JV. Bootstrap methods for developing predictive models. *Am Stat.* 2005;58(2):131-137.
36. Pencina MDA, D'Agostino RB. Evaluating discrimination of risk prediction models: the C statistic. *JAMA.* 2015;314(10):1063-1064.
37. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21(1):128-138.
38. Nash D, Rane MS, Robertson MM, et al. Severe acute respiratory syndrome coronavirus 2 incidence and risk factors in a national, community-based prospective cohort of us adults. *Clin Infect Dis.* 2022 May 27 <https://doi.org/10.1093/cid/ciac423>.
39. Leopold SS. Editorial: beware of studies claiming that social factors are "independently associated" with biological complications of surgery. *Clin Orthop Relat Res.* 2019;477(9):1967-1969.
40. Hoxha I, R. Agahi, A. Bimbashi, M. Aliu, L. Raka, I. Bajraktari, et al. Higher COVID-19 vaccination rates are associated with lower COVID-19 mortality: a global analysis. *Vaccines (Basel).* 2023;11(1):74.
41. Nasreen S, H. Chung, S. He, K. A. Brown, J. B. Gubbay, S. A. Buchan, et al. Effectiveness of mRNA and ChAdOx1 COVID-19 vaccines against symptomatic SARS-CoV-2 infection and severe outcomes with variants of concern in Ontario. *Nat Microbiol.* 2022.7(3):379-385.
42. Andrews N, J. Stowe, F. Kirsebom, S. Toffa, T. Rieckard, E. Gallagher, et al. COVID-19 vaccine effectiveness against the Omicron (B.1.1.529) variant. *N Engl J Med.* 2022;386(16):1532-1546.
43. Rozenfeld Y, Beam J, Maier H, et al. A model of disparities: risk factors associated with COVID-19 infection. *Int J Equity Health.* 2020;19(126).<https://doi.org/10.1186/s12939-020-01242-z>.

## Author Biographies

**Richard Sloane**, MPH is a Senior Biostatistician at the Center on Aging and Human Development's Data Science and Statistics Laboratory.

**Carl F Pieper** is a Professor in the Duke University Medical Center Department of Biostatistics and Bioinformatics. He is director of the Center on Aging and Human Development's Data Science and Statistics Laboratory.

**Richard Faldowski** is a Senior Biostatistician in the Duke Aging Center. He holds a Ph.D. in Quantitative Psychology, with expertise in latent variable methods, multivariate and longitudinal analyses, and psychometrics.

**Douglas Wixted**, MMCI, is senior research program leader in the Duke Clinical & Translational Science Institute and was the operational lead for the C3PI Study.

**Coralei E Neighbors** holds a MSc in global health. Her area of research includes infectious disease surveillance, health economics, and health policy.

**Christopher W Woods** is a professor of medicine, pathology, and global health with a research interest in infectious disease diagnostics.

**L Kristin Newby**, MD, MHS is Professor in the Division of Cardiology, Department of Medicine at Duke University and a senior investigator at the Duke Clinical Research Institute. She was Principal Investigator of the Cabarrus County COVID-19 Prevalence and Immunity (C3PI) Study.