A Machine Learning Approach to Predicting Higher COVID-19 Care Burden in the Primary Care Safety Net: Hispanic Patient Population Size a Key Factor

Evan V. Goldstein¹ and Fernando A. Wilson¹,²,³

Abstract
Introduction: The federal government legislated supplemental funding to support community health centers (CHCs) in response to the COVID-19 pandemic. Supplemental funding included standard base payments and adjustments for the number of total and uninsured patients served before the pandemic. However, not all CHCs share similar patient population characteristics and health risks.

Objective: To use machine learning to identify the most important factors for predicting whether CHCs had a high burden of patients diagnosed with COVID-19 during the first year of the pandemic.

Methods: Our analytic sample included data from 1342 CHCs across the 50 states and D.C. in 2020. We trained a random forest (RF) classifier model, incorporating 5-fold cross-validation to validate the RF model while optimizing the model's hyperparameters. Final performance metrics were calculated following the application of the model that had the best fit to the held-out test set.

Results: CHCs with a high burden of COVID-19 had an average of 65.3 patients diagnosed with COVID-19 per 1000 patients in 2020. Our RF model had 80.9% accuracy, 80.1% precision, 25.0% sensitivity, and 98.1% specificity. The percentage of Hispanic patients served in 2020 was the most important feature for predicting whether CHCs had high COVID-19 burden.

Conclusions: Findings from our RF model suggest patient population race and ethnicity characteristics were most important for predicting whether CHCs had a high burden of patients diagnosed with COVID-19 in 2020, though sensitivity was low. Enhanced support for CHCs serving large Hispanic patient populations may have an impact on addressing future COVID-19 waves.

Keywords
community health centers, community health, primary care, prevention, health promotion, COVID-19

Introduction
Federally-funded community health centers (CHCs) provide primary health services and address health disparities experienced by patients living in communities experiencing lower socioeconomic status.¹,² Since 2020, CHCs have served on the front lines of the COVID-19 pandemic, providing COVID-19 screening, treatment, and vaccination services to primarily underserved populations. Like most U.S. health care providers, CHCs faced new operational and financial challenges caused by the COVID-19 pandemic.

The federal government legislated supplemental funding to support CHCs' operations in response to these challenges.³ Provisions included $1.3 billion allocated through the Coronavirus Aid, Relief, and Economic Security (CARES) Act in 2020,³ and additional funding provided through the American Rescue Plan Act (ARPA) in 2021. However, the

¹ Department of Population Health Sciences, Spencer Fox Eccles School of Medicine, University of Utah, Salt Lake City, UT, USA
² Matheson Center for Health Care Studies, University of Utah, Salt Lake City, UT, USA
³ Department of Economics, University of Utah, Salt Lake City, UT, USA

Submitted May 16, 2022. Revised June 23, 2022. Accepted July 7, 2022.

Corresponding Author:
Evan V. Goldstein, Department of Population Health Sciences, Spencer Fox Eccles School of Medicine, University of Utah, Salt Lake City, UT 84108, USA.
Email: evan.goldstein@hsc.utah.edu
amount of supplemental funding provided to each CHC in 2020 and 2021 included a standard base payment plus adjustments for the number of total and uninsured patients served before the pandemic. In other words, supplemental funding provided in response to the pandemic was not necessarily allocated according to the actual burden of COVID-19 at each CHC. Nor was the funding adjusted for patient population factors linked to higher COVID-19 prevalence. This is problematic because not all CHCs share similar organizational characteristics; not all CHC patient populations share the same health risks.

Additional federal support may be needed to help CHCs overcome different factors affecting their ability to provide adequate care during the pandemic. There is little understanding in the scholarly literature about how different patient population characteristics affect the burden of COVID-19 at CHCs. Our objective was to use a machine learning (ML) model to identify the most important factors for predicting whether CHCs had a high burden of patients diagnosed with COVID-19 during the first year of the pandemic. We also aim to develop a practical approach that may be helpful to other Uniform Data System (UDS) data users.

### Methods

#### Data & Sample

We conducted a cross-sectional study. Our primary data source was the UDS (calendar year from January 1 to December 31). The Health Resources and Services Administration (HRSA) collects UDS data annually on patient characteristics, service utilization, and organizational features of all CHCs. Data from the Bureau of Labor Statistics (BLS), U.S. Census Bureau, Ballotpedia, and the New York Times COVID-19 data repository were used to construct several state-level characteristics. Our analytic sample included 1342 CHCs (representing 28,126,105 patients) across the 50 states and D.C. in 2020, the most recent year of data available.

### Variables

Our outcome variable was a binary label indicating whether a CHC had a high burden of patients diagnosed with COVID-19 (ICD-10 U07.1) in 2020. We defined this variable as equal to 1 for CHCs in the 75th-or-higher percentile (ie, top-25%) of CHCs, ranked by the rate of patients diagnosed with COVID-19 per 1000 patients in 2020 and 0 for other CHCs (ie, bottom-75%).

Our ML model used 82 features to predict whether CHCs had a high burden of patients diagnosed with COVID-19 (Table 1). These features included organizational characteristics of each CHC and demographic and health characteristics of each CHC’s patient population extracted from the UDS, which measures and reports these standardized measures for all CHCs annually. We included a state-level measure of per capita COVID-19 prevalence at the end of 2020—aggregated by the New York Times—to account for differences in COVID-19 severity between the states. We used BLS and U.S. Census Bureau data to include measures of unemployment and poverty rates for each state and account for differences in state-level economic circumstances. We included a binary variable equal to 1 if a state enacted a mask-wearing policy in 2020, and equal to 0 if the state did not enact a mask-wearing policy. Before the availability of pharmaceutical interventions for COVID-19, 39 states and the District of Columbia enacted policies requiring individuals to wear masks in indoor or outdoor public spaces statewide in 2020 to mitigate the burdens imposed by the pandemic on health care providers. Mask-wearing policy information was extracted from Ballotpedia, a nonpartisan encyclopedia that catalogs the timing of state-level policy decisions. Geographic indicators for each state were also included.

Data were complete for our state-level characteristics. Data were missing for less than 5% (and typically less than 1%) of

| Table 1. List of Features Used in the RF Model (n = 82). |
|-------------------------------------------------------|
| Feature                                              |
| Percent of patient population non-Hispanic, black     |
| Percent of patient population non-Hispanic, white     |
| Percent of patient population non-Hispanic, Asian     |
| Percent of patient population Hispanic                |
| Percent of patient population female                  |
| Percent of patient population ages 18-64             |
| Percent of patient population ages 65 and older       |
| Percent of patient population living at or below the Federal Poverty Line |
| Percent of patient population veterans                |
| Percent of patient population without health insurance |
| Percent of patient population covered by Medicaid     |
| Percent of patient population covered by Medicare     |
| Percent of patient population covered by private health insurance |
| Medicaid managed care enrollment (member months)      |
| Percent of patient population diagnosed with obesity  |
| Percent of patient population diagnosed with depression or mood disorder |
| Percent of patient population diagnosed with anxiety  |
| Percent of patient population diagnosed with diabetes |
| Percent of patient population diagnosed with heart disease |
| Percent of patient population diagnosed with hypertension |
| Percent of patient population diagnosed with asthma  |
| Percent of patient population diagnosed with tobacco use disorder |
| Total patients; 1000s                                |
| Total annual Health Resources & Services Administration funding |
| Indicator if CHC was Health Care for the Homeless “special populations” provider |
| Indicator if CHC was Public Housing Primary Care “special populations” provider |
| Indicator if CHC served predominantly rural (vs urban) patient population |
| Indicator if state enacted any mask-wearing mandate policy in 2020 COVID-19 cases per capita (state level) |
| Unemployment rate (state level)                       |
| Poverty rate (state level)                            |
| 51 geographic identifier indicators, one for each U.S. state and D.C. |

Notes: The RF model used the features shown in this table to predict whether CHCs had a high burden of patients diagnosed with COVID-19.
all patient population demographic and health characteristics extracted from the UDS. We used multiple imputation by chained equations to replace any missing data.

Analysis

We trained a random forest (RF) classifier model to predict whether CHCs had a high burden of patients diagnosed with COVID-19 during the first year of the pandemic. Instead of taking a hypothesis-driven statistical approach to identifying predictive factors, we used a data-driven approach to perform pattern recognition and construct the predictive model. We chose the RF classifier for its efficiency, adaptability to nonlinearities in the data, and robustness to outliers. All ML modeling was conducted using the scikit-learn library and Python 3.9 on a Jupyter Notebook.

We randomly split our analytic sample into a training set (80%) and a test set (20%). We incorporated K-fold cross-validation (CV) with five folds to validate the RF model while optimizing the model’s hyperparameters to maximize precision using RandomizedSearchCV. CV is a resampling procedure used to evaluate the performance of an ML model on unseen data and develop a more general model. During K-fold CV, our training set was randomly divided into five folds. One fold was used for validation in each of five iterations, while the remaining four folds were used to train the model. The online Appendix provides information about the hyperparameters tuned through this process.

Final performance metrics were calculated following the application of the model that had the best fit to the held-out test set. Accuracy, precision, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) metrics were calculated. We computed feature importance using the Mean Decrease in Impurity method. Bivariate t-tests were used to compare the high- and lower-COVID-19 burden CHCs by the most important features identified by the RF model.

Results

CHCs with a high burden of COVID-19 had an average of 65.3 patients diagnosed with COVID-19 per 1000 patients in 2020. Our tuned RF model had 80.9% accuracy, 80.1% precision, 25.0% sensitivity, and 98.1% specificity in classifying CHCs with high COVID-19 burden (Table 2). The percentage of Hispanic patients served in 2020 was the most important feature for predicting whether CHCs had high COVID-19 burden (Figure 1). The remaining top-10 important features were the percentage of non-Hispanic, white patients served; percentage of non-Hispanic, black patients served; percentage of patients diagnosed with tobacco use disorder; percentage of patients diagnosed with depression; percentage of patients covered by private health insurance; total patients served; percentage of patients covered by Medicaid; percentage of patients ages 65 and older; and percentage of veteran patients.

Compared to CHCs with lower rates of patients diagnosed with COVID-19 in 2020, high-COVID-19-burden CHCs served over one-and-a-half times the percentage of Hispanic patients (36.3% vs 22.9%; \( P < .001 \)). CHCs with a high burden of patients diagnosed with COVID-19 served comparatively fewer non-Hispanic, white (34.2% vs 44.8%; \( P < .001 \)) and non-Hispanic, black patients (15.6% vs 19.5%; \( P = .005 \)). Appended Table A1 describes the mean values for the remaining top-20 important features for CHCs with high and lower COVID-19 burden in 2020.

Discussion

Findings from our RF model suggest patient population race and ethnicity characteristics were most important for predicting whether CHCs had a high burden of patients diagnosed with COVID-19 during the first year of the pandemic. Notably, the state in which a CHC operated appeared to have little importance for predicting whether CHCs experienced high COVID-19 burden. Overall, patient population characteristics seemed to have more predictive importance than state-level characteristics. Our model had high accuracy, precision, and specificity. The model was particularly efficient at correctly identifying true high-COVID-19-burden CHCs and avoiding false-positive predictions of high-COVID-19-burden CHCs. However, our model’s low sensitivity was a significant limitation. Incorrect classification examples were typically false negatives, likely because true negative rates were high in our label of interest.

CHCs may need additional support from policymakers to help address specific risk factors affecting COVID-19 prevalence and treatment needs, especially if other variants emerge. Supplemental COVID-19 funds allocated through the CARES Act and ARPA were adjusted according to the size of each CHC’s total and uninsured patient population sizes. However, our model suggests that enhanced support targeted toward CHCs serving large Hispanic patient populations may have an impact in helping to address future COVID-19 waves. Developing ML algorithms may help policymakers identify which types of CHCs would most benefit from additional support. Our general model may be useful for this purpose; however, additional model development is necessary to overcome our low model sensitivity, including training and testing model performance in different locations.

Table 2. Performance Metrics of the Tuned Random Forest Model.

| Performance metric | Value |
|------------------|-------|
| Accuracy (%)     | 80.9  |
| Precision (%)    | 80.1  |
| Sensitivity (%)  | 25.0  |
| Specificity (%) | 98.1  |
| AUC-ROC          | 78.2  |

Notes: K-fold CV with five folds was used to validate the model while optimizing the model’s hyperparameters. These performance metrics were calculated following the application of the model that had the best statistical fit to the held-out test set.
### 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

The corresponding author is supported by the Department of Population Health Sciences at the Spencer Fox Eccles School of Medicine, University of Utah.

### ORCID iD

Evan V. Goldstein [https://orcid.org/0000-0001-7766-7450](https://orcid.org/0000-0001-7766-7450)

### Supplemental Material

Supplemental material for this article is available online.

### References

1. Goldstein EV. Community health centers maintained initial increases in medicaid covered adult patients at 5-years post-medicaid-expansion. *Inquiry*. 2021;58(1):1-10.
2. Sardell A. *The U.S. Experiment in Social Medicine: The Community Health Center Program, 1965–1986*. University of Pittsburgh Press; 1989.
3. National Association of Community Health Centers. Summary of key CHC provisions in the CARES act. 2020. [https://wsd-nachsparkinfluence.s3.amazonaws.com/uploads/2020/03/CARES-Act-Summary-for-Health-Centers.pdf](https://wsd-nachsparkinfluence.s3.amazonaws.com/uploads/2020/03/CARES-Act-Summary-for-Health-Centers.pdf) (accessed 20 October 2021).
4. Health Resources & Services Administration. COVID-19 Information for Health Centers and Partners. Published 2022. Accessed May 5, 2022. [https://bphc.hrsa.gov/emergency-response/coronavirus-info](https://bphc.hrsa.gov/emergency-response/coronavirus-info)
5. HRSA. Uniform Data System (UDS) Resources. Published 2020. Accessed August 20, 2021. [https://bphc.hrsa.gov/datareporting/reporting/index.html](https://bphc.hrsa.gov/datareporting/reporting/index.html)
6. U.S. Bureau of Labor Statistics. Local Area Unemployment Statistics. Published 2019. [https://www.bls.gov/lau/](https://www.bls.gov/lau/)
7. U.S. Census Bureau. Historical Poverty Tables: People and Families—1959 to 2020. Published 2021. Accessed May 5, 2022. [https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-people.html](https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-people.html)
8. Ballotpedia. State-level mask requirements in response to the coronavirus (COVID-19) pandemic, 2020–2021. Published 2021. Accessed July 29, 2021. [https://ballotpedia.org/State-level_mask_requirements_in_response_to_the_coronavirus_(COVID-19)_pandemic_2020-2021#Footnotes](https://ballotpedia.org/State-level_mask_requirements_in_response_to_the_coronavirus_(COVID-19)_pandemic_2020-2021#Footnotes)
9. The New York Times. Coronavirus (Covid-19) Data in the United States. Accessed January 5, 2022. [https://github.com/nytimes/covid-19-data](https://github.com/nytimes/covid-19-data)

---

**Figure 1.** Relative importance of the top 20 features in predicting whether CHCs had a high burden of COVID-19 during the first year of the pandemic: 2020.

Notes: Feature importance was computed using the Mean Decrease in Impurity method.
10. National Association of Community Health Centers. Health Centers on the Front Lines of COVID-19: $7.6 Billion in Lost Revenue and Devastating Impact on Patients and Staff. 2020. https://www.nachc.org/wp-content/uploads/2020/04/Financial-Loss-Fact-Sheet.pdf

11. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clin Infect Dis*. 2018;66(1). doi:10.1093/cid/cix731.

12. Schonlau M, Zou RY. The random forest algorithm for statistical learning. *Stata J*. 2020;20(1). doi:10.1177/1536867X20909688.

13. Hastie T, Tibshirani R, Friedman J. Springer Series in Statistics The Elements of Statistical Learning—Data Mining, Inference, and Prediction. Vol 2nd. 2009.

14. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning—with Applications in R | Gareth James | Springer. 2013.

15. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12(1):2825–2830.

16. Kluwer T, Ragan-Kelley B, Pérez F, et al. Jupyter notebooks—a publishing format for reproducible computational workflows.

17. Poldrack RA, Huckins G, Varoquaux G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry*. 2020;77(5). doi:10.1001/jamapsychiatry.2019.3671.

**Author Biographies**

**Evan V. Goldstein, Ph.D., M.P.P.** is an Assistant Professor in the Department of Population Health Sciences at the Spencer Fox Eccles School of Medicine, University of Utah.

**Fernando A. Wilson, Ph.D.** is a Professor in the Department of Economics and the Department of Population Health Sciences at the Spencer Fox Eccles School of Medicine, University of Utah. Dr. Wilson is also the Director and Endowed Chair of the Matheson Center for Health Care Studies at the University of Utah.