Using a Cloud-Based Machine Learning Classification Tree Analysis to Understand the Demographic Characteristics Associated With COVID-19 Booster Vaccination Among Adults in the United States

Lu Meng1,2,*, Hannah E. Fast1,3,*, Ryan Soele,3 Jennifer Zelt4, Bhavini Patel Murthy3, Neil Chandra Murthy1, Peng-Jun Lu1,5 Lauren Shaw1, Lynn Gibbs-Scharf1, and Terence Chorba1,5

1CDC COVID-19 Response Team, US Centers for Disease Control and Prevention (CDC), Atlanta, Georgia, USA, 2Federal Civilian Division, General Dynamics Information Technology, Inc., Falls Church, Virginia, USA, 3Immunization Services Division, National Center for Immunization and Respiratory Diseases, CDC, Atlanta, Georgia, USA, 4Stat-Epi Associates, Inc., Ponte Vedra Beach, Florida, USA, and 5Division of Tuberculosis Elimination, National Center for HIV, Viral Hepatitis, STD, and TB Prevention, CDC, Atlanta, Georgia, USA

Keywords. COVID-19; COVID-19 vaccination; booster dose; coronavirus.

Coronavirus disease 2019 (COVID-19) booster vaccination increases protection against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, including the recently predominant Omicron variant (B.1.1.529), and reduces COVID-19-associated hospitalization and death [1]. During August–November 2021, a series of Emergency Use Authorizations and recommendations, including those for an additional primary dose for immunocompromised persons and a booster dose for persons age ≥18 years, were approved by the Food and Drug Administration [2]. In the United States, as of April 2022, all adults (age ≥18 years) were eligible to receive a booster dose ≥2 months after vaccination with the 1-dose Johnson & Johnson/Jansen (J&J) primary series or ≥5 months after the second dose of the Pfizer-BioNTech or Moderna 2-dose mRNA primary series [2]. Certain populations may have also chosen to receive a second booster dose using an mRNA COVID-19 vaccine ≥4 months after the first booster dose [2].

As of March 2022, ~47% of persons age ≥18 years who were eligible to receive a booster dose after completing a primary series of COVID-19 vaccine had not yet received a booster [3]. Disparities in COVID-19 vaccine booster uptake have been related to socioeconomic status, insurance status, disability, and social demographic factors, including age, education level, race/ethnicity, and residency in rural or urban areas [4–7]. In the present study, we applied machine learning methods in the form of a classification tree algorithm to identify and describe relationships and interactions of demographic factors associated with the receipt or nonreceipt of a COVID-19 booster vaccine among eligible persons age ≥18 years in the United States.

METHODS

Over 152 million COVID-19 primary vaccine completion records (administered from 12/14/2020 through 09/15/2021) and 81 million first booster dose records (administered through 03/15/2022) reported to the Centers for Disease Control and Prevention (CDC) from 49 states and the District of Columbia (DC) were analyzed using the cloud-based data platform Microsoft Azure DataBricks (Azure Databricks | Microsoft Azure). Texas had data-sharing restrictions on information reported to the CDC; its data were not available for inclusion. Vaccine records from US territories were not included in the present study. Recipients’ primary series and booster dose records were matched. A classification tree model was built to examine factors contributing to receiving a booster dose, with Gini impurity as the classification tree splitting metric [8]. Input variables included primary series vaccine product (Moderna, Pfizer-BioNTech, J&J), age group (18–24, 25–34, 35–44, 45–54, 55–64, ≥65 years), sex (male, female), race/ethnicity (Hispanic/Latino, non-Hispanic Black [Black], non-Hispanic White [White], other/multiracial/unknown [other/unknown]), region (South, Midwest, Mountain, Pacific, Northeast [South Region includes AZ, NM, OK, AR, LA, MS, AL, TN, KY, GA, SC, NC, WV, MD, VA, FL, DE, & DC; Midwest Region includes ND, SD, NE, KS, MN, IA, MO, IL, WI, IN, MI, & OH; Mountain Region includes NV, UT, CO, WY, MT, & ID; Pacific Region includes WA, HI, AK, OR, & CA; Northeast Region includes PA, NY, VT, NH, ME, MA, RI, CT, & NJ]), urbanicity (large central metro, large fringe metro [large fringe metro counties are counties in Metropolitan Statistical Areas of ≥1 million population that

Received 08 July 2022; accepted 27 August 2022; published online 1 September 2022

*Co-first authors

Correspondence: Terence Chorba, MD, CDC/DDID/NCHHSTP/DTBE, 1600 Clifton Rd NE (US 12-4), Atlanta, GA 30329 (tlc2@CDC.GOV).

Open Forum Infectious Diseases®

Published by Oxford University Press on behalf of Infectious Diseases Society of America 2022. This work is written by (a) US Government employee(s) and is in the public domain in the US. https://doi.org/10.1093/ofid/ofac446
do not qualify as large central; for more information regarding urbanicity classification, please see Gaffney et al. [7], medium metro, small metro, micropolitan, noncore [9]), and CDC/ATSDR Social Vulnerability Index (SVI) of zip code of residence (low, medium, and high). Factors affecting SVI scores include socioeconomic status, household composition, disability, minority status, housing type, and transportation. A lower SVI score means the zip code of residence is less socially vulnerable [10–11]. All the input variables were derived from vaccine records. Gender identity was not available. Descriptive analyses were performed for input variables, and feature importance of input variables and prediction rate of each end node were reported. This study was reviewed by the CDC and conducted in accordance with applicable federal law and CDC policy.

RESULTS

As shown in Figure 1, the classification tree model had a depth of 5 branches, with 17 end nodes and 32 nodes in total. Detailed information (eg, sample sizes, prediction rates, etc.) about the 17 end nodes is listed in Supplementary Table 1. All the input variables were derived from vaccine records. Gender identity was not available. Descriptive analyses were performed for input variables, and feature importance of input variables and prediction rate of each end node were reported. This study was reviewed by the CDC and conducted in accordance with applicable federal law and CDC policy.

Figure 1. Classification tree diagram depicting demographic characteristics associated with COVID-19 booster vaccination among adults completing the primary series before September 15, 2021, by social demographic factors, March 15, 2022, United States. Detailed information (eg, sample sizes, prediction rates, etc.) about the 17 end nodes is listed in Supplementary Table 1. Abbreviations: AI/AN, American Indian/Alaska Native; COVID-19, coronavirus disease 2019; J&J, Johnson & Johnson/Janssen; OPI, other Pacific Islander; SVI, Social Vulnerability Index.

Table 1 presents results from descriptive analyses of COVID-19 vaccine booster dose status by social demographic factors. Lower booster coverage was observed among J&J

Table 1 presents results from descriptive analyses of COVID-19 vaccine booster dose status by social demographic factors. Lower booster coverage was observed among J&J
primary series recipients, younger age groups (eg, 18–34 years), residents of areas that are more socially vulnerable, people from racial and ethnic minority groups, and residents of the South.

**DISCUSSION**

This study used 233 million COVID-19 vaccination records to construct a classification tree model that assessed demographic characteristics associated with receipt or nonreceipt of COVID-19 booster vaccination among US adult populations. The classification tree model provides a framework to consider the impact of each input variable on vaccination outcomes within specific subpopulations; it would be prohibitively time-consuming to investigate outcomes at this granularity using other analytical approaches.

Age group was the most important characteristic, with a feature importance score of 0.739, and persons age 18–34 years in all regions were less likely to have received a booster vaccination. Previous studies have identified attitudes and beliefs corresponding to low intent to receive primary series vaccination and low primary series coverage among young adults age 18–39 years [13].

The South had lower booster coverage than the other 4 regions and was split by the model from all other regions to form its own branches. SVI and urbanicity were important predictors of booster status in the South. Southerners residing in less socially vulnerable areas or large fringe metro areas were more likely to have received a booster dose. Residents within these areas report higher household income, which has been linked with higher COVID-19 vaccine uptake [9, 14]. In addition, marginalized populations within rural or socially vulnerable areas may have limited transportation options, less paid time off, and reduced ability to access vaccination providers [15–16]. Our finding that SVI is an important predictor of booster dose status among Southerners age 35–54 years is consistent with the observation of greater income-associated health disparities in the South than in other regions [17]. Among non-Southerners, age, primary vaccine type, race/ethnicity, and urbanicity determined the outcome. For persons age 35–54 years who received a primary series of Moderna or Pfizer, the tree model identified non-Hispanic White persons as more likely to be boosted; however, this pattern of race and ethnicity was not found among persons in other age groups or in residents of the South.

Regardless of age or region, recipients of a J&J primary series were less likely to have received a booster dose. Given lower vaccine effectiveness of a J&J primary series compared with an mRNA vaccine primary series, this population would particularly benefit from the increased effectiveness conferred by a booster dose [18]. More information is needed to understand factors contributing to low booster uptake among J&J recipients. Some J&J recipients may have chosen the 1-dose primary series because they were less likely to complete a 2-dose mRNA vaccination series, whether due to vaccination-related anxiety (eg, needle aversion), to concerns about mRNA vaccines due to health conditions or personal beliefs, or to barriers to accessing health care or vaccine providers (eg, transportation, limited time off, reduced availability of specific vaccines in certain geographic areas) [19–21].

These findings are subject to at least 3 limitations. First, Texas data were not included in this analysis, and given

---

**Table 1.** COVID-19 Vaccine Booster Dose Status for Adults Completing the Primary Series Before September 15, 2021, by Social Demographic Factors, March 15, 2022, United States

| Booster Dose Status | Not Boosted, No. | % | Boosted, No. | % | Total, No. |
|---------------------|-----------------|---|--------------|---|------------|
| Total               | 71 056 071      | 46.71 | 81 060 169 | 53.29 | 152 116 240 |
| Primary series completion dose vaccine product | | | | | |
| Pfizer-BioNTech     | 37 512 813      | 48.88 | 42 272 769 | 51.12 | 79 885 582 |
| Moderna             | 26 062 915      | 43.48 | 33 886 214 | 56.52 | 60 949 129 |
| Johnson & Johnson   | 7 680 343       | 61.04 | 4 901 118  | 38.96 | 12 581 529 |

**Age group**

- 18–24 y: 8 953 407, 64.28% to 4 975 323, 35.72% to 13 928 730
- 25–34 y: 13 177 374, 60.43% to 8 628 063, 39.57% to 21 805 437
- 35–44 y: 12 453 119, 53.82% to 10 685 458, 46.18% to 23 138 577
- 45–54 y: 11 763 997, 48.61% to 12 438 296, 51.39% to 24 202 293
- 55–64 y: 11 470 661, 40.69% to 16 717 198, 59.31% to 28 187 859
- ≥65 y: 13 237 513, 32.40% to 27 615 831, 67.60% to 40 853 344

**Sex**

- Male: 34 863 037, 48.91% to 36 418 962, 51.09% to 71 281 999
- Female: 36 193 034, 44.77% to 44 641 207, 55.23% to 80 834 241

**Urbanicity**

- Large fringe metro: 22 259 023, 46.06% to 26 065 686, 53.94% to 48 324 709
- Large central metro: 18 549 122, 45.38% to 23 224 584, 54.62% to 40 873 706
- Medium metro: 15 313 549, 47.85% to 16 888 091, 52.15% to 32 001 640
- Small metro: 6 127 185, 47.80% to 6 691 771, 52.20% to 12 818 956
- Micropolitan: 5 413 237, 49.11% to 5 608 530, 50.89% to 11 021 767
- Noncore: 3 993 955, 47.97% to 3 881 507, 52.03% to 7 875 462

**Social Vulnerability Index**

- High: 22 625 719, 50.05% to 25 832 831, 49.95% to 45 829 550
- Medium: 28 684 159, 48.96% to 32 396 900, 51.04% to 60 881 059
- Low: 19 746 193, 43.09% to 26 079 438, 56.91% to 45 825 631

**Race/ethnicity**

- Hispanic: 10 212 284, 58.70% to 7 186 269, 41.30% to 17 398 553
- Non-Hispanic Black: 5 890 660, 52.48% to 5 325 513, 47.52% to 11 206 173
- Non-Hispanic American Indian/Alaska Native: 571 057, 57.20% to 427 329, 42.80% to 998 386
- Non-Hispanic Asian/OPI: 3 214 046, 38.12% to 5 218 318, 61.88% to 8 432 364
- Non-Hispanic White: 30 949 706, 42.83% to 41 305 029, 57.17% to 72 254 735
- Other/Unknown: 20 228 318, 48.36% to 21 597 711, 51.64% to 41 826 029

**Region**

- South: 26 809 572, 53.52% to 23 284 388, 46.48% to 50 093 960
- Midwest: 13 871 561, 41.70% to 19 396 829, 58.30% to 33 268 390
- Mountain: 3 481 795, 45.87% to 4 108 301, 54.13% to 7 590 096
- Pacific: 12 059 872, 41.07% to 17 307 608, 58.93% to 29 367 480
- Northeast: 14 833 271, 46.66% to 16 963 043, 53.35% to 31 796 314

Abbreviations: COVID-19, coronavirus disease 2019; OPI, other Pacific Islander.
Texas’ large population size, lack of data from Texas could have impacted these findings. Second, the booster status of a small portion of individuals may have been misclassified if the booster dose record was not able to be linked to the primary series completion record, such as if vaccinations were received in different jurisdictions. Third, the current tree model yields a 61.5% prediction rate, which may limit the application of these findings. A single classification tree model is often reported to have relatively low prediction accuracy; we found during the process of model selection that replacing a single tree with a random forest of trees or growing the tree model to a depth of >5 branches could improve prediction rates but would dramatically reduce interpretability [12].

The classification tree diagram is a novel approach to analyzing public health vaccination data. One advantage of the classification tree approach is its use of a splitting metric to identify partitions in input variable responses, which describes variability across a population in a way that is easy to understand. By structuring certain demographic characteristics into paths, the classification tree was able to describe the relationships structuring certain demographic characteristics into paths, portion of individuals may have been misclassified if the bootstrap dose of COVID-19 vaccine in Peru. Vaccines (Basel) 2022; 10:1183.

Yoshida M, Kobashi Y, Kawamura T, et al. Factors associated with COVID-19 vaccine booster hesitancy: a retrospective cohort study, Fukushima Vaccination Community Survey. Vaccines (Basel) 2022; 10:515.

Lee RC, Hu H, Kawaguchi ES, et al. COVID-19 booster vaccine attitudes and behaviors among university students and staff in the United States: the USC Trojan Pandemic Research Initiative. Prev Med Rep 2022; 28:101866.

Gaffney A, Himmelstein DU, McCormick D, Woolhandler S. Disparities in COVID-19 vaccine booster uptake in the USA: December 2021–February 2022. J Gen Intern Med 2022; 37:2918–21.

Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Routledge; 2017.

Ingram DD, Franco SJ, 2013 NCHS urban–rural classification scheme for counties. National Center for Health Statistics. Vital Health Stat 2014; 2(166):1–73.

Hughes MM, Wang A, Grossman MK, et al. County-level COVID-19 vaccination coverage and social vulnerability - United States, December 14, 2020-March 1, 2021. MMWR Morb Mortal Wkly Rep 2021; 70:431–6.

Centers for Disease Control and Prevention. CDC/ATSDR Social Vulnerability Index (SVI). Available at: https://www.atstd.cdc.gov/placeandhealth/svi/index.html. Accessed April 25, 2022.

Sharma A. Decision tree vs. random forest – which algorithm should you use? Analytics Vidhya (blog). May 12, 2020. Available at: https://www.analyticsvidhyab.com/blog/2020/05/decision-tree-vs-random-forest-algorithm. Accessed April 25, 2022.

Baack BN, Abad N, Yankey D, et al. COVID-19 vaccination coverage and intent among adults aged 18–39 years—United States, March–May 2021. MMWR Morb Mortal Wkly Rep 2021; 70:928–33.

Gertz A, Rader B, Sewalk K, Brownstein JS. Emerging socioeconomic disparities in COVID-19 vaccine second-dose completion rates in the United States. Vaccines (Basel) 2022; 10:121.

Anderson EL. Recommended solutions to the barriers to immunization in children and adults. Mo Med 2014; 111:34–8.

Rodriguez-Diaz CE, Guiral-Hernandez V, Mena L, et al. Risk for COVID-19 infection and death among Latinos in the United States: examining heterogeneity in transmission dynamics. Ann Epidemiol 2020; 52:46–53.e2.

Oates GR, Jackson BE, Partridge EE, Singh KP, Fouad MN, Bae S. Socio-demographic patterns of chronic disease: how the Mid-South Region compares to the rest of the country. Am J Prev Med 2017; 52:S31–9.

Natarajan K, Prasad N, Dascomb K, et al. Effectiveness of homologous and heterologous COVID-19 booster doses following 1 Ad26.COVID.S (Janssen [Johnson & Johnson]) vaccine dose against COVID-19–associated emergency department and urgent care encounters and hospitalizations among adults—VISION Network, 10 states, December 2021–March 2022. MMWR Morb Mortal Wkly Rep 2022; 71:495–502.

Hause AM, Gee J, Johnson T, et al. Anxiety-related adverse event clusters after Janssen COVID-19 vaccination—five U.S. mass vaccination sites, April 2021. MMWR Morb Mortal Wkly Rep 2021; 70:685–8.
20. Curley B. Why some people still prefer the Johnson & Johnson COVID-19 vaccine. Healthline. May 4, 2021. Available at: https://www.healthline.com/health-news/why-some-people-still-prefer-the-johnson-johnson-covid-19-vaccine. Accessed April 25, 2022.

21. Weinstein ER, Balise R, Metheny N, et al. Factors associated with Latino sexual minority men’s Likelihood and motivation for obtaining a COVID-19 vaccine: a mixed-methods study. J Behav Med. 2022 Apr 27:1–13. doi:10.1007/s10865-022-00315-4. Epub ahead of print.