HHS Public Access

Author manuscript

Med Decis Making. Author manuscript; available in PMC 2024 April 24.

Published in final edited form as:

Med Decis Making. 2024 February ; 44(2): 175–188. doi:10.1177/0272989X231218024.

Bias-Adjusted Predictions of County-Level Vaccination Coverage from the COVID-19 Trends and Impact Survey

Marissa B. Reitsma,
Department of Health Policy, Stanford University, Stanford, CA, USA

Sherri Rose,
Department of Health Policy, Stanford University, Stanford, CA, USA

Alex Reinhart,
Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA, USA; Delphi Group, Carnegie Mellon University, Pittsburgh, PA, USA

Jeremy D. Goldhaber-Fiebert,
Department of Health Policy, Stanford University, Stanford, CA, USA; Freeman Spogli Institute for International Studies, Stanford University, Stanford, CA, USA

Joshua A. Salomon
Department of Health Policy, Stanford University, Stanford, CA, USA

Abstract

Background.—The potential for selection bias in nonrepresentative, large-scale, low-cost survey data can limit their utility for population health measurement and public health decision making. We developed an approach to bias adjust county-level COVID-19 vaccination coverage predictions from the large-scale US COVID-19 Trends and Impact Survey.

Design.—We developed a multistep regression framework to adjust for selection bias in predicted county-level vaccination coverage plateaus. Our approach included poststratification to the American Community Survey, adjusting for differences in observed covariates, and secondary normalization to an unbiased reference indicator. As a case study, we prospectively applied this framework to predict county-level long-run vaccination coverage among children ages 5 to 11 y. We evaluated our approach against an interim observed measure of 3-mo coverage for children ages 5 to 11 y and used long-term coverage estimates to monitor equity in the pace of vaccination scale up.

Corresponding Author: Marissa B. Reitsma, Department of Health Policy, Stanford University, 615 Crothers Way, Encina Commons, Stanford, CA 94305, USA; (mreitsma@stanford.edu).

Presented at the Society for Medical Decision Making 44th Annual North American Meeting (October 2022).

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Alex Reinhart received salary support from an unrestricted gift from Facebook. Facebook was involved in the design and conduct of US COVID-19 Trends and Impact Survey. All funders, including Facebook, had no role in the analysis and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Supplemental Material
Supplementary material for this article is available online at https://doi.org/10.1177/0272989X231218024.

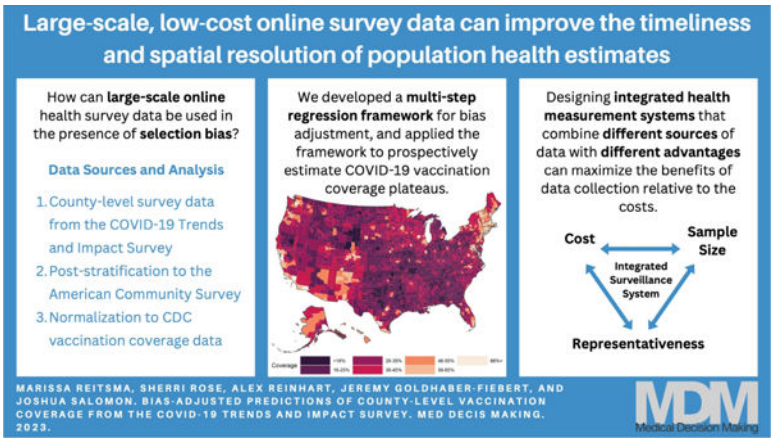
Results.—Our predictions suggested a low ceiling on long-term national vaccination coverage (46%), detected substantial geographic heterogeneity (ranging from 11% to 91% across counties in the United States), and highlighted widespread disparities in the pace of scale up in the 3 mo following Emergency Use Authorization of COVID-19 vaccination for 5- to 11-y-olds.

Limitations.—We relied on historical relationships between vaccination hesitancy and observed coverage, which may not capture rapid changes in the COVID-19 policy and epidemiologic landscape.

Conclusions.—Our analysis demonstrates an approach to leverage differing strengths of multiple sources of information to produce estimates on the time scale and geographic scale necessary for proactive decision making.

Implications.—Designing integrated health measurement systems that combine sources with different advantages across the spectrum of timeliness, spatial resolution, and representativeness can maximize the benefits of data collection relative to costs.

Graphical Abstract



Keywords

COVID-19 vaccination; online survey data; population health measurement; heterogeneity

Background

Local, representative, and timely data can provide immense benefit to public health decision making, but such data are costly to collect repeatedly with samples large enough for county-level estimation in the United States. For example, during the rollout of the COVID-19 vaccine, indicators of people's vaccination intentions could ideally be used to predict subsequent vaccine uptake and to drive targeted efforts to reduce hesitancy and thereby increase achieved coverage. While representative survey data are expensive to collect frequently in large samples, low-cost, online survey data can achieve sufficient samples for county-level estimation but have been shown to yield biased estimates of selected outcomes, with misleadingly small margins of error.[1–3] Although programmatic data offer retrospective reporting of some health indicators at the county level, these data become available too late

to enable prospective planning and decision making, and many important indicators do not have routine reporting systems.[1,4]

Combining data sources with different advantages and limitations can help to balance tradeoffs between time, cost, and representativeness of data collection.[5] Enormous variation in life expectancy across counties in the United States underscores the importance of spatially granular measures of health.[6] Previous studies have combined multiple data sources for retrospective bias correction and small area estimation of health indicators such as smoking prevalence, obesity prevalence, and cardiovascular disease mortality, among others.[7–13] These studies have used a range of statistical techniques, including Bayesian hierarchical models, multilevel regression and poststratification, and propensity score matching. Bayesian hierarchical models and multilevel regression share information across nested units, for example, counties nested within states, and can account for observed covariates. Poststratification matches observed covariates to population reference data, aiming to adjust the sample to better represent the population, while propensity score matching uses the reference data to model the probability that population members are included in the sample. Our study extends this literature by applying regression, poststratification, and secondary normalization methods to a new stream of large-scale online health survey data.

The COVID-19 pandemic catalyzed a new era of massive real-time data collection for public health, exemplified by the US COVID-19 Trends and Impact Survey (CTIS), which ran continuously between April 2020 and June 2022.[2] The US survey had an average of 40,000 responses daily, which is much higher than samples achieved through household health surveys, and a response rate of approximately 1%, which is much lower than household health survey response rates. The size of CTIS allowed for timely small-area estimation of many policy-relevant leading indicators, but its utility has been questioned due to upward bias in estimates of vaccination coverage compared with representative reporting data.[3] Considering the low survey response rate, selection bias—encompassing both sampling and nonresponse biases—may explain the difference between predicted and observed vaccination coverage. Approaches to gain actionable insights in the presence of selection bias from these large-scale online survey data remain relevant to responses to future public health emergencies and to general population health measurement, for which similar large-scale, low-cost surveys could be deployed.

In this study, we present a framework to generate bias-adjusted county-level estimates from a large-scale online survey. As an illustrative case study, we use data from CTIS to prospectively predict county-level vaccination coverage plateaus among children ages 5 to 11 y. Although COVID-19 vaccination has been central to the public health response to the pandemic, coverage has plateaued well below 100%, with wide variation across communities. COVID-19 vaccination intentions have also been an important indicator derived from survey data over the course of the pandemic.[8,14–18] Vaccination intentions can be used to anticipate eventual vaccination coverage for different groups, which can then be used to direct resources and targeted interventions, design policies, deploy additional risk reduction tools, and monitor both the pace and equity of scale up. In the United States,

children ages 5 to 11 y became eligible for COVID-19 vaccination when Emergency Use Authorization (EUA) was extended at the end of October 2021.[19]

Methods

We developed a multistep regression framework (Figure 1) to predict vaccination coverage plateaus among children ages 5 to 11 y. First, we estimated county-level parental hesitancy toward vaccinating their children using a mixed-effects logistic regression model fit to survey data. Next, we used a second logistic regression model to estimate the relationship between county-level parental hesitancy and observed vaccination coverage, for youth ages 12 to 17 y, who became eligible for COVID-19 vaccination earlier than children ages 5 to 11 and therefore provide a reference group. Finally, we combined the results from the 2 regression models to predict county-level vaccination coverage for 5- to 11-y-olds.

Data Sources

We combined individual-level survey responses from wave 11 of CTIS, collected during the period from July 1, 2021, through October 31, 2021, with data from wave 12, collected during the period from December 19, 2021, through February 14, 2022. The survey was managed and implemented by the Delphi Group at Carnegie Mellon University. Participants were recruited through Facebook, and the sampling frame is the Facebook Active User base. Additional information on CTIS has been previously published.[2] The full questionnaire for waves 11 and 12 is available online.[20]

In addition to CTIS, we used individual-level sociodemographic data (age, documented sex, education, race/ethnicity, and household structure) from the 2015 to 2019 American Community Survey (ACS) for poststratification.[21] Individual-level data from ACS are available at the public-use microdata area level. Public-use microdata areas have a minimum population size of 100,000, so some large population counties contain multiple public-use microdata areas, while other small population counties may be combined into a single public-use microdata area. We mapped public-use microdata areas to counties using the Missouri Census Data Center's Geographic Correspondence Engine.[22] Counties are never split between 2 public-use microdata areas. When a single county contained multiple public-use microdata areas, we aggregated public-use microdata areas to the county level. When a single public-use microdata area spanned multiple counties, we assumed the same distribution of sociodemographic characteristics for each county in the public-use microdata area and scaled sample weights by relative county population size.

Finally, we used complete vaccination coverage data for ages 12 to 17 y reported at the county level by the Centers for Disease Control and Prevention (CDC) for the second-stage regression and data from the same source over the first 3 mo after eligibility for ages 5 to 11 y for performance evaluation of coverage predictions.[23]

Estimating County-Level Hesitancy

To estimate county-level parental hesitancy, we fit a mixed-effects logistic regression to survey data on the stated intentions of parents/guardians regarding vaccinating their children. We classified ''No, definitely not'' and ''No, probably not'' as hesitant responses to the

question ‘‘Will you choose to get a COVID-19 vaccine for your child or children when they are eligible?’’ Responses of ‘‘Yes, definitely’’ and ‘‘Yes, probably’’ were considered not hesitant. Consistent with previous analyses, we used the imprecise but available construct of ‘‘reported hesitancy’’ and focused on it principally as an intermediate indicator that would be subsequently mapped to long-run vaccination coverage.

The CTIS survey questionnaire evolved as new information became available over the course of the pandemic. Importantly, although wave 11 asked parents about vaccine hesitancy, it did not ask for the ages of their children. Since wave 12 elicited the age of the parent’s oldest child, we used it to examine differences in parental hesitancy for those whose oldest child was between the ages of 12 and 17 versus ages 5 to 11y.

The first-stage logistic regression modeled the probability of parental hesitancy as a function of fixed effects for documented gender (male, female), age group (18–24, 25–34, 35–44, 45–54, 55–64, 65+ y), education (high school or fewer years of education, some college or a 2-y degree, 4-y degree, graduate degree), and race/ethnicity (Hispanic, non-Hispanic American Indian or Alaska Native, non-Hispanic Asian, non-Hispanic Black, non-Hispanic Native Hawaiian or Other Pacific Islander, non-Hispanic White, non-Hispanic multiracial or other race), and age group of child (unknown, 12–17 y, and 5–11 y) and nested random intercepts on state and county:

$$\ln\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \alpha_{jk} + \beta X_{ijk}$$

$$\alpha_{jk} = \alpha_k + \mu_{jk}; \mu_{jk} \sim N\left(0, \sigma^2_{\mu_{jk}}\right),$$

$$\alpha_k \sim N\left(0, \sigma^2_{\sigma_k}\right), \quad (1)$$

i = individual CTIS responses from parents/guardians; j = counties; k = states

We did not perform a weighted regression to include the CTIS survey weights, instead adjusting for the probability of inclusion and nonresponse through poststratification.[24] Poststratification effectively reweights subgroup-level estimates to be representative of a target population. We combined data from counties with a sample size of 10 or fewer into grouped counties, by state. To generate county-level predictions of hesitancy, including uncertainty around these predictions, from the first-stage regression, we generated 1,000 draws of subgroup-level predicted probabilities of hesitancy for unique combinations of documented gender, age group, education, race/ethnicity, and county using the estimated regression

coefficients, assuming a multivariate normal distribution of the parameters including the fixed and random effects ($\hat{p}_{gj}$), where g corresponds to each unique demographic characteristic combination. We then poststratified county- and subgroup-level predicted probabilities of hesitancy to produce overall county-level hesitancy estimates $\left(\hat{\theta}_j\right)$:

$$\hat{\theta}_j = \frac{\sum w_{gj}\hat{p}_{gj}}{\sum w_{gj}}. \quad (2)$$

In other words, through equation 2 we produced a weighted average of predicted probability of hesitancy by county, in which the component predictions (by group g) are weighted by the population of that group according to a representative data set. Specifically, the weights (w_{gj}) used for poststratification were based on an analysis of individual-level data from the ACS that reflected household structure and incorporated children's sample weights. First, we identified each child's parents/guardians based on the first available of the following: 1) parents directly coded in the ACS, 2) grandparents designated as responsible for 1 or more children directly coded in the ACS, 3) adults (18+ y) in the same household and same family unit, and 4) adults (18+ y) in the same household but different family unit. Next, we assigned the child's sample weight to each of their parents/guardians, dividing the weight by the total number of identified parents/guardians. Finally, we summed the children's sample weights across each subgroup (g), defined by the demographic characteristics of the parents/guardians, and each county (j), resulting in the final weight (w_{gj}) used for poststratification. By assigning children's sample weights to their parents/guardians, our poststratified estimates are representative of the target population of children eligible for COVID-19 vaccination in each county.

Predicting County-Level Vaccination Coverage from Hesitancy Estimates

We used a second logistic regression model to translate county-level hesitancy estimates to county-level vaccination coverage predictions. We trained the model on paired county-level hesitancy and coverage estimates for children ages 12 to 17 y and then projected the predictive relationships onto the hesitancy estimates for children ages 5 to 11 y under the assumption that the same relationships would apply across the 2 age groups. To estimate the regression model for the 12- to 17-y group, we first generated estimates of parental hesitancy for this group using the same regression model specification described above for the 5- to 11-y group, in this case predicting hesitancy for parents of children ages 12 to 17 y and poststratifying estimates based on household structure and sample weights of children ages 12 to 17 y. These county-level hesitancy estimates were used as independent variables in the second logistic regression. For our dependent variable, we used coverage data from February 1, 2022, which was approximately 9 mo after 12- to 17-y-olds first became eligible for vaccination (ages 16–17 y in early April 2021 and ages 12–15 y on May 10, 2021).

For states with at least 10 counties reporting vaccination coverage data for children ages 12 to 17 y on February 1, 2022, with CDC reporting completeness exceeding 80%, we fit state-specific regressions. For all other states and the District of Columbia ($n = 7$), we fit regressions at the census division level. This prevented overfitting to small numbers of counties or low-quality reporting data. Regressions were weighted based on the size of the 12- to 17-y-old population in each county. The second regression was fit to each of the 1,000 draws of county-level hesitancy from the first regression. Uncertainty from the second

regression was captured through 1,000 draws from the multivariate normal distribution of the fixed effects plus the residual variance.

Finally, we used the models fit on the relationship between parental hesitancy and vaccination coverage for children ages 12 to 17 y to predict coverage for children ages 5 to 11 y based on our first-stage estimates of hesitancy for this age group. Final prediction intervals were based on the 2.5th and 97.5th percentiles of 1 million final county-level coverage predictions (1,000 draws from the first regression crossed with 1,000 draws from the second regression).

Performance Evaluation

We compared our estimates of parental hesitancy toward vaccinating children ages 12 to 17 y to estimates on the same indicator produced by the Office of the Assistant Secretary for Planning and Evaluation (ASPE), including comparing correlation coefficients between estimated county-level hesitancy and observed vaccination coverage on February 1, 2022. Our first-stage logistic regression model included only a subset of the predictors incorporated in ASPE's first-stage logistic regression model, so differences in performance mainly reflect the additional information value of county-level data from CTIS, rather than the value of a more sophisticated prediction model.

To evaluate our use of the relationship between hesitancy and coverage for children ages 12 to 17 y, applied to children ages 5 to 11 y, we assessed interim coverage predictions for the 5- to 11-y age group at 3 mo after EUA expansion against observed county-level coverage reported by CDC, based on mean absolute error, the intraclass correlation coefficient, and the percentage of counties for which the 95% prediction interval contained the observed coverage level nationally and at the state level. Since CDC does not report separate county-level coverage estimates for ages 12 to 15 y versus 16 to 17 y, the time since an age group first became eligible for vaccination is an imprecise but best-available approach to this interim performance evaluation.

Monitoring Progress and Equity in Scale up

To monitor the pace of vaccination scale up, we defined a measure of ''3-mo progress'' as follows:

$$\text{progress} = \frac{\text{Observed three month coverage}}{\text{Predicted nine month coverage}} \tag{3}$$

To monitor equity in the pace of vaccination scale up, we used linear regression to analyze associations between this progress measure and the county-level socioeconomic status domain of CDC's Social Vulnerability Index (SVI), which reflects measures of poverty, unemployment, income, and education.[25]

Study Approval and Data Availability

The study was approved by Stanford's Institutional Review Board, under protocol number 56018. All analyses were conducted using the R programming language version 3.6.3. Analytic code is available through GitHub (https://github.com/PPML/CTIS-County-Vaccination-Coverage). CTIS microdata can be accessed through a data use agreement with Carnegie Mellon University, while the ACS data and CDC vaccination data are publicly available.

Role of the Funders

Facebook was involved in the design and conduct of the US COVID-19 Trends and Impact Survey. All funders had no role in the analysis and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Results

Data

Between July 1 and October 31, 2021, a total of 613,460 responses to wave 11 of the US CTIS were collected from parents/guardians of children younger than 18 y with complete demographic information. To allow for variation in parental hesitancy by child age group, we supplemented the analyses with 119,465 responses collected from parents/guardians whose oldest children were between the ages of 5 and 17 y in wave 12, between December 19, 2021, and February 14, 2022. We report exclusions in the sample flowchart (Supplementary Figure S1). Unweighted and weighted distributions of respondents by age, documented gender, education, and race/ethnicity are reported in Table 1. Poststratification to the ACS reduced bias from the nonrepresentative sample. Of 3,142 counties, 1,203 had a sample size of at least 100, while 293 had a sample size between 1 and 10, and 23 had zero respondents. Maps of county-level sample sizes and sample rates are reported in Supplementary Figures S2 and S3. Trends in parental hesitancy by month over the study period are shown in Supplementary Figures S4 and S5.

Hesitancy Estimates

Modeled county-level parental hesitancy toward vaccination for children ages 5 to 11 y ranged from 7% (95% prediction interval: 5%–9%) in San Mateo County, California, to 74% (61%–84%) in Platte County, Wyoming. Although the population-weighted national average hesitancy was 31%, 2,787 counties (89% of all counties) had hesitancy levels exceeding this benchmark. The skewed distribution of county-level estimates versus state- and national-aggregates is largely driven by lower hesitancy in urban areas with large populations and higher hesitancy in rural areas with smaller populations. Across counties, median hesitancy toward vaccination was 21% (interquartile range: 18%–25%) higher for parents of children ages 5 to 11 y, compared with parents of children ages 12 to 17 y. Our estimates of hesitancy among parents of children ages 12 to 17 y using data from CTIS reflected substantially more substate variation in hesitancy, compared with previously published estimates from ASPE (Figure 2). In addition, our estimates showed a stronger correlation with vaccination

coverage on February 1, 2022, among children ages 12 to 17 y (CTIS: 20.78; ASPE: 20.44) (Supplementary Figure S6).

Predicted Coverage Levels

The predicted mean national plateau coverage level among children ages 5 to 11 y by August 2022, 9 mo after EUA, was 46%. There was substantial state-level variation in predicted plateau coverage, ranging from 30% and less in Wyoming, Alabama, Mississippi, Idaho, and Louisiana to 66% and greater in Connecticut, District of Columbia, and Massachusetts. Four of the 5 counties with the highest predicted coverage were in California. Ninety-two percent of counties were predicted to fall short of a 50% coverage benchmark by August 2022 for children ages 5 to 11 y, while 56% of counties were predicted to not reach 30% coverage. Eighty-six percent of counties were predicted to fall short of their state average coverage level, highlighting an urban-rural divide in vaccination. Higher levels of predicted coverage were concentrated in the Northeast, West Coast, and in urban centers across the country (Figure 3).

Model Validation

Figure 4 shows the relationship between predicted 3-mo coverage among children ages 5 to 11 y and observed coverage 3 mo after EUA. The mean absolute error across all counties, weighted by county population size, was 0.059. The unweighted mean absolute error was 0.051. State-level weighted and unweighted mean absolute errors are reported in Supplementary Table S7. The intraclass correlation coefficient for consistency of predicted versus observed 3-mo coverage at the national level was 0.81. Intraclass correlation coefficients were greater than 0.75 for 16 states, between 0.50 and 0.75 for 19 states, and less than 0.50 for 10 states. Intraclass correlation coefficients at the state level are reported in Supplementary Table S8. The prediction interval for 3-mo coverage included the observed coverage level for 81% of counties.

Monitoring Progress and Equity in Scale up

Relative to long-term predicted coverage levels, at the state level, Vermont, Rhode Island, and Maine had the fastest pace of scale up of coverage among children ages 5 to 11 y at 3 mo after EUA, while Louisiana, Alabama, and Mississippi had the slowest pace of scale up (Figure 5). We find that errors in 9-mo coverage predictions among ages 12 to 17 y were not significantly associated with the socioeconomic status domain of the SVI in all states except South Dakota, Nevada, and Montana. As a result, in addition to predicting plateau coverage levels, we can use county-level predicted coverage levels to monitor equity in the pace of vaccination scale up. More vulnerable counties, as measured by the socioeconomic status domain of the SVI generally made less progress toward reaching their plateau coverage levels over the first 3 mo after EUA expansion, compared with less vulnerable counties. There was significantly slower scale-up progress in more vulnerable (higher SVI) counties in 35 of 45 states reporting data on vaccination coverage among children ages 5 to 11 y (Figure 6).

Discussion

We generated bias-adjusted county-level predictions of long-term vaccination coverage for children ages 5 to 11 y that leveraged data on parental vaccination intentions from the US CTIS. To mitigate the impacts of selection bias in the sample, we combined CTIS data with representative sociodemographic data from the American Community Survey and unbiased programmatic data on vaccination coverage. Our approach to estimation and propagation of multiple sources of uncertainty produced prediction intervals that included observed coverage levels for 81% of counties 3 mo after EUA. Our estimation framework can be broadly used to generate actionable indicators on the time scale and at the geographic scale required for proactive decision making in public health emergencies.

Across and within states, our estimates highlight substantial geographic heterogeneity in both parental hesitancy and predicted coverage. When substantial geographic heterogeneity exists within states, county-level data are critical to efficient and effective response. In the case of COVID-19 vaccination, county-level estimates of hesitancy have implications for targeting of efforts to promote vaccination uptake, and expectations for eventual vaccination uptake that are relevant to planning for other protective measures, including masking, testing, and improved ventilation.[26,27]

Despite consistent messaging about the importance of promoting equity in COVID-19 vaccination scale up, we observe a pervasive pattern of slower vaccination scale up in more vulnerable counties, as measured by the socioeconomic domain of CDC's SVI.[28–30] The socioeconomic status domain of the SVI comprises measures of income, poverty, employment, and education. Inequity in vaccination scale up was observed in every phase of the COVID-19 vaccination campaign.[31–34] Moving forward, as organizations prepare for ongoing COVID-19 boosters in the short term and for the next potential pandemic in the long term, more intensive and explicitly proequity policies and programs are required to avoid replicating the disparities in COVID-19 outcomes observed to date.

Large-scale, low-cost surveys offer a promising approach to population health measurement. They offer advantages for rapid and continuous deployment and allow estimation at smaller geographic scales compared with traditional approaches to data collection, including representative household surveys and retrospective reporting data. The value of county-level data from CTIS, compared with the state-level data available from the Census Household Pulse Survey, is evident in their respective performance in capturing substate heterogeneity in vaccination intentions and coverage. The correlation between parental hesitancy and coverage among children ages 12 to 17 y for CTIS was 20.78, compared with a correlation coefficient of 20.44 for ASPE estimates based on the Census Household Pulse Survey.[8] Although we produced estimates using a single snapshot of data, our framework could be adapted to routinely integrate newly available data and regularly update estimates. For example, the large-scale, low-cost survey could be deployed frequently to capture temporal trends, whereas the more expensive unbiased reference could be collected sporadically to update the relationship between the survey and the reference.

Beyond the COVID-19 pandemic, large-scale, low-cost surveys could be deployed to routinely generate and update estimates of geographic heterogeneity in determinants of health, health care access, and health outcomes. Designing integrated health measurement systems that intentionally combine sources with different advantages across the spectrum of timeliness, spatial resolution, and representativeness can maximize the benefits of data collection relative to their costs. Future large-scale, low-cost data collection efforts should ensure sufficient indicators are incorporated in the survey instrument for poststratification as well as availability of appropriate reference indicators for secondary bias adjustment.

The results of our study should be interpreted in the context of several limitations. First, to predict plateau coverage levels for children ages 5 to 11 y, we assumed that the relationship between hesitancy and coverage observed for children ages 12 to 17 y applies to this younger age group. For counties in states that used the division-level regression between hesitancy and coverage, we assumed that the relationship between hesitancy and coverage across all counties in the census division would apply to the states with insufficient data to estimate a state-specific relationship. Our 3-mo validation supported these assumptions, which were necessary for prospective estimation. Second, estimates of hesitancy for children of different age groups became available only in wave 12 of the CTIS survey, and respondents were asked only about intentions to vaccinate their oldest child. Third, we relied on historical relationships between hesitancy and observed coverage, which would not necessarily capture the evolving COVID-19 policy and epidemiologic landscape. Fourth, our analytic framework was designed to capture geographic variation in coverage but not variation by other important population characteristics such as race/ethnicity within small geographic areas. Fifth, our first-stage regression model did not incorporate explicit spatial structure beyond nesting counties within states and did not incorporate additional group-level predictors that may have improved estimates from counties with smaller samples. Further exploration of alternative first-stage models would be warranted if our approach were to be applied in practice. Despite these limitations, our estimates reflect a principled approach to generating bias-adjusted estimates from large-scale, low-cost survey data that can be used to inform decisions and evaluate actual progress against a reference scenario.

Conclusion

A combination of poststratification and secondary normalization to an unbiased reference can reduce bias in large-scale, low-cost survey data. Applying this method to predict long-term county-level COVID-19 vaccination coverage among children ages 5 to 11 y, we found substantial substate geographic heterogeneity and disparities in the pace of scale up. Although direct estimates of vaccination coverage from the COVID-19 Trends and Impact Survey are biased, a multistep regression strategy can result in bias-adjusted actionable predictions on the time scale and geographic scale required for proactive public health decision making.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

MBR is supported by the National Science Foundation Graduate Research Fellowship Program under grant No. DGE-1656518, Stanford's Knight-Hennessy Scholars Program, and the Stanford Data Science Scholars Program. MBR, JDGF, and JAS are supported by the Stanford Clinical and Translational Science Award to Spectrum (UL1TR003142). JDGF, SR, and JAS are supported by funding from the Centers for Disease Control and Prevention and the Council of State and Territorial Epidemiologists (NU38OT000297) and by funding from the Health Equity Research Project Fund from Stanford's School of Medicine. JDGF and JAS are supported by funding from the National Institute on Drug Abuse (3R37DA01561217S1). AR is supported by an unrestricted gift from Facebook.

We would like to thank the Delphi Group at Carnegie Mellon University for their role in managing the COVID-19 Trends and Impact Survey. We appreciate feedback from members of the SC-COSMO group (https://sc-cosmo.org/) and Prevention Policy Modeling Lab (https://ppml.stanford.edu/) on this work.

Data Availability

Survey microdata are not publicly available because survey participants only consented to public disclosure of aggregate data and because the legal agreement with Facebook governing operation of the survey prohibits disclosure of microdata without confidentiality protections for respondents. Deidentified microdata are available to researchers under a Data Use Agreement that protects the confidentiality of respondents. Access can be requested online (https://cmu-delphi.github.io/delphi-epidata/symptom-survey/data-access.html). Requests are reviewed by the Carnegie Mellon University Office of Sponsored Programs and Facebook Data for Good.

References

1. Kreuter F, Barkay N, Bilinski A, et al. Partnering with a global platform to inform research and public policy making. Surv Res Methods 2020;14:159–63.
2. Salomon JA, Reinhart A, Bilinski A, et al. The US COVID-19 trends and impact survey: continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination. Proc Natl Acad Sci U S A 2021;118:e2111454118. [PubMed: 34903656]
3. Bradley VC, Kuriwaki S, Isakov M, Sejdinovic D, Meng XL, Flaxman S. Unrepresentative big surveys significantly overestimated US vaccine uptake. Nature 2021;600:695–700. [PubMed: 34880504]
4. Rosenfeld R, Tibshirani RJ. Epidemic tracking and forecasting: lessons learned from a tumultuous year. Proc Natl Acad Sci U S A 2021;118:e2111456118. [PubMed: 34903658]
5. Blumberg SJ, Parker JD, Moyer BC. National health interview survey, COVID-19, and online data collection platforms: adaptations, tradeoffs, and new directions. Am J Public Health 2021;111:2167–75. [PubMed: 34878857]
6. GBD US Health Disparities Collaborators. Life expectancy by county, race, and ethnicity in the USA, 2000–19: a systematic analysis of health disparities. Lancet 2022;400:25–38. [PubMed: 35717994]
7. Schenker N, Raghunathan TE. Combining information from multiple surveys to enhance estimation of measures of health. Stat Med 2007;26:1802–11. [PubMed: 17278184]
8. ASPE. Parents' intentions to vaccinate children ages 12–17 for COVID-19: demographic factors, geographic patterns, and reasons for hesitancy Available from: https://aspe.hhs.gov/reports/hesitancy-vaccinate-children [Accessed 16 March, 2022].
9. Irimata KE, He Y, Cai B, Shin HC, Parsons VL, Parker JD. Comparison of quarterly and yearly calibration data for propensity score adjusted web survey estimates. Surv Methods Insights Field Epub ahead of print October 2020. DOI: 10.13094/SMIF-2020-00018

10. Ward ZJ, Long MW, Resch SC, et al. Redrawing the US obesity landscape: bias-corrected estimates of state-specific adult obesity prevalence. PLoS One 2016;11:e0150735. [PubMed: 26954566]
11. Zhang X, Holt JB, Yun S, Lu H, Greenlund KJ, Croft JB. Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. Am J Epidemiol 2015;182:127–37. [PubMed: 25957312]
12. Elliott MR, Davis WW. Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. J R Stat Soc Ser C Appl Stat 2005;54:595–609.
13. Vaughan AS, Coronado F, Casper M, Loustalot F, Wright JS. County-level trends in hypertension-related cardiovascular disease mortality—United States, 2000 to 2019. J Am Heart Assoc 2022;11:e024785. [PubMed: 35301870]
14. Daly M, Jones A, Robinson E. Public trust and willingness to vaccinate against COVID-19 in the US from October 14, 2020, to March 29, 2021. JAMA 2021;325:2397–9. [PubMed: 34028495]
15. KFF. KFF COVID-19 vaccine monitor dashboard 2022. Available from: https://www.kff.org/coronavirus-covid-19/dashboard/kff-covid-19-vaccine-monitor-dashboard/[Accessed 16 March, 2022].
16. Office of the Assistant Secretary for Planning and Evaluation. Vaccine hesitancy for COVID-19: state, county, and local estimates Available from: https://aspe.hhs.gov/index.php/reports/vaccine-hesitancy-covid-19-state-county-local-estimates [Accessed 30 September, 2021].
17. Institute for Health Metrics and Evaluation. COVID-19 vaccine hesitancy. Institute for Health Metrics and Evaluation Available from: https://vaccine-hesitancy.healthdata.org/ [Accessed 16 March, 2022].
18. SteelFisher GK, Blendon RJ, Caporello H. An uncertain public—encouraging acceptance of Covid-19 vaccines. N Engl J Med 2021;384:1483–7. [PubMed: 33657291]
19. U.S. Food & Drug Administration. FDA authorizes Pfizer-BioNTech COVID-19 vaccine for emergency use in children 5 through 11 years of age 2021. Available from: https://www.fda.gov/news-events/press-announcements/fdaauthorizes-pfizer-biontech-covid-19-vaccine-emergency-use-children-5-through-11-years-age [Accessed 7 April, 2022].
20. Delphi Epidata API. Questions and coding Available from: https://cmu-delphi.github.io/delphi-epidata/symptom-survey/coding.html [Accessed 16 March, 2022].
21. Ruggles S, Flood S, Foster S, et al. IPUMS USA: Version 11.0 2015–2019 American Community Survey Minneapolis (MN): IPUMS; 2021. DOI: 10.18128/D010.V11.0.
22. Missouri Census Data Center. Geocorr 2018. Available from: https://mcdc.missouri.edu/applications/geocorr2018.html [Accessed 16 March, 2022].
23. Centers for Disease Control and Prevention. COVID-19 vaccinations in the United States, County. Data Available from: https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh [Accessed 16 March, 2022].
24. Gelman A Struggles with survey weighting and regression modeling. Stat Sci 2007;22:153–64.
25. Agency for Toxic Substances and Disease Registry. CDC SVI documentation 2018. Place and health 2022. Available from: https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/SVI_documentation_2018.html [Accessed 16 March, 2022].
26. Centers for Disease Control and Prevention. Operational Guidance for K-12 Schools and Early Care and Education Programs to Support Safe In-Person Learning https://www.cdc.gov/coronavirus/2019-ncov/community/schools-childcare/k-12-childcare-guidance.html. [Accessed 7 April, 2022].
27. Giardina J, Bilinski A, Fitzpatrick MC, et al. Model-estimated association between simulated US elementary school–related SARS-CoV-2 transmission, mitigation interventions, and vaccine coverage across local incidence levels. JAMA Netw Open 2022;5:e2147827. [PubMed: 35157056]
28. The White House. National COVID-19 preparedness plan Available from: https://www.whitehouse.gov/covidplan/ [Accessed 7 April, 2022].
29. Centers for Disease Control and Prevention. Equity in childhood COVID-19 vaccination 2021. Available from: https://www.cdc.gov/vaccines/covid-19/planning/children/equity.html [Accessed 7 April, 2022].

30. Ndugga N, Hill L, Artiga S, Haldar S. Latest data on COVID-19 vaccinations by race/ethnicity. Kaiser Family Foundation 2022. Available from: https://www.kff.org/coronavirus-covid-19/issue-brief/latest-data-on-covid-19-vaccinations-by-race-ethnicity/ [Accessed 7 April, 2022].

31. Faherty LJ, Ringel JS, Williams MV, et al. The U.S. equity-first vaccination initiative: early insights. RAND Corporation 28 January 2022. Available from: https://www.rand.org/pubs/research_reports/RRA1627-1.html [Accessed 7 April, 2022].

32. Wrigley-Field E, Kiang MV, Riley AR, et al. Geographically targeted COVID-19 vaccination is more equitable and averts more deaths than age-based thresholds alone. Sci Adv 7:eabj2099.

33. Dada D, Djiometio JN, McFadden SM, et al. Strategies that promote equity in COVID-19 vaccine uptake for Black communities: a review. J Urban Health 2022;99:15–27. [PubMed: 35018612]

34. Quinn SC, Andrasik MP. Addressing vaccine hesitancy in BIPOC communities—toward trustworthiness, partnership, and reciprocity. N Engl J Med 2021;385:97–100. [PubMed: 33789007]

Highlights

- The COVID-19 pandemic catalyzed massive survey data collection efforts that prioritized timeliness and sample size over population representativeness.
- The potential for selection bias in these large-scale, low-cost, nonrepresentative data has led to questions about their utility for population health measurement.
- We developed a multistep regression framework to bias adjust county-level vaccination coverage predictions from the largest public health survey conducted in the United States to date: the US COVID-19 Trends and Impact Survey.
- Our study demonstrates the value of leveraging differing strengths of multiple data sources to generate estimates on the time scale and geographic scale necessary for proactive public health decision making.

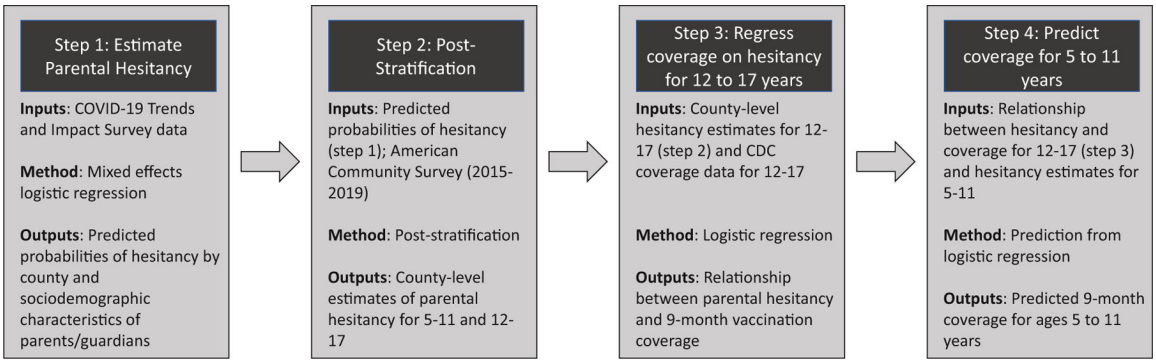


Figure 1.
Methods flowchart.

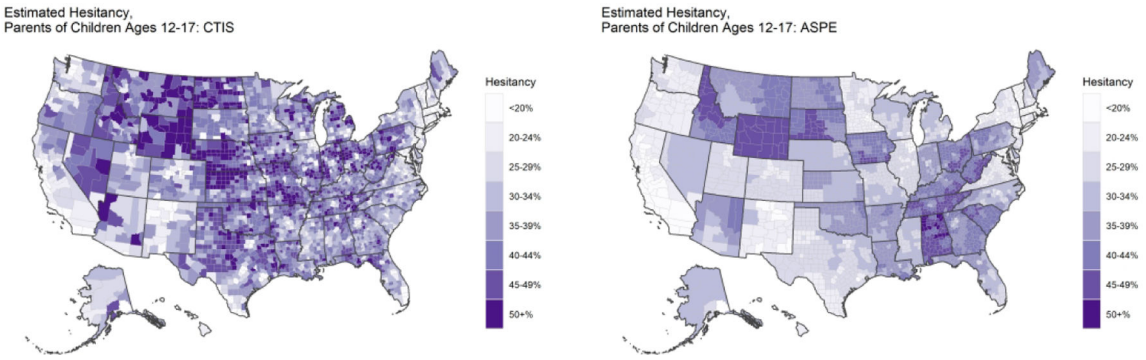


Figure 2.

Comparison of county-level hesitancy estimates for parents of children ages 12 to 17 y. Study estimates using data from the COVID-19 Trends and Impact Survey (CTIS) (left), compared to published estimates produced by the Office of the Assistant Secretary for Planning and Evaluation (ASPE) (right).

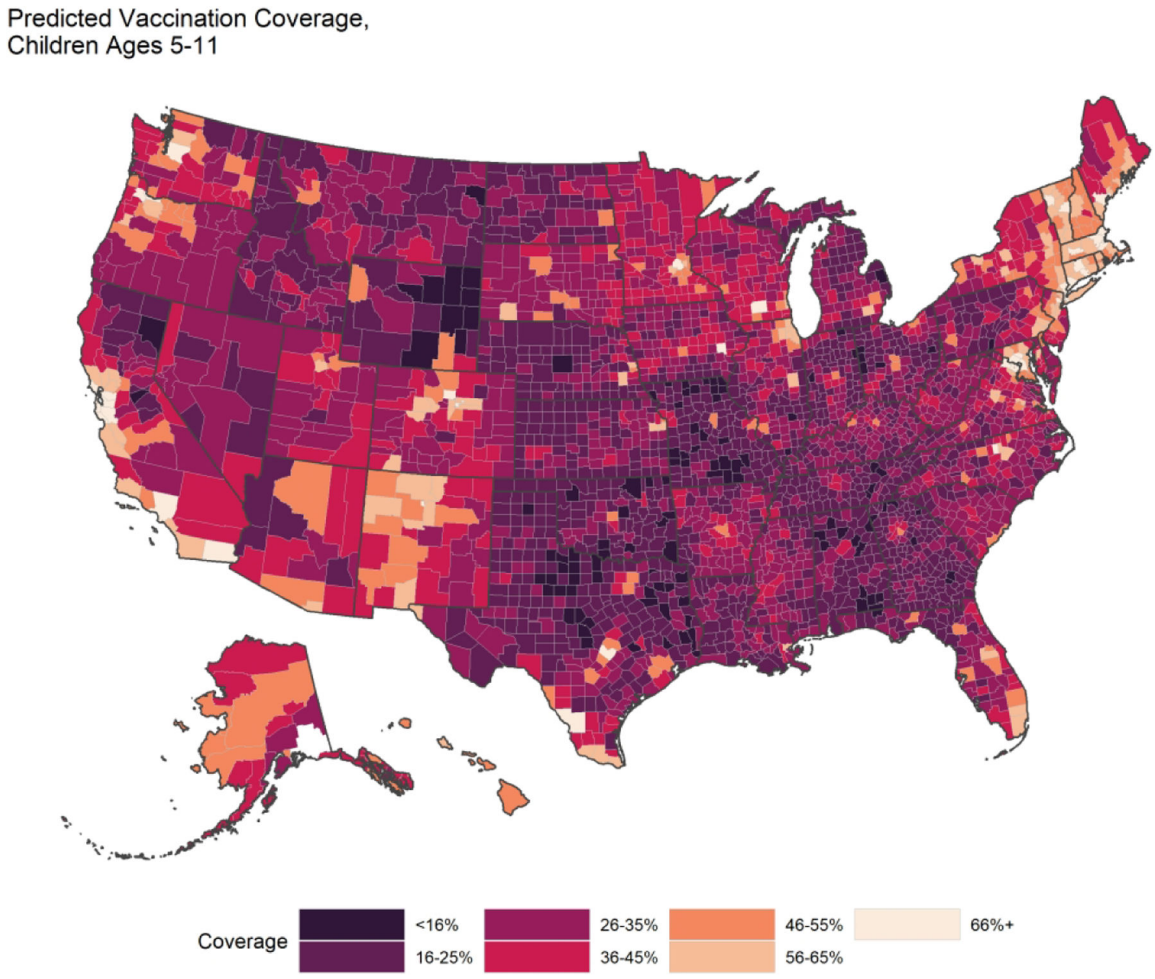


Figure 3.

County-level map of 9-mo predicted plateau complete vaccination coverage levels for children ages 5 to 11 y. The color scale is split at the national average predicted coverage of 46%.

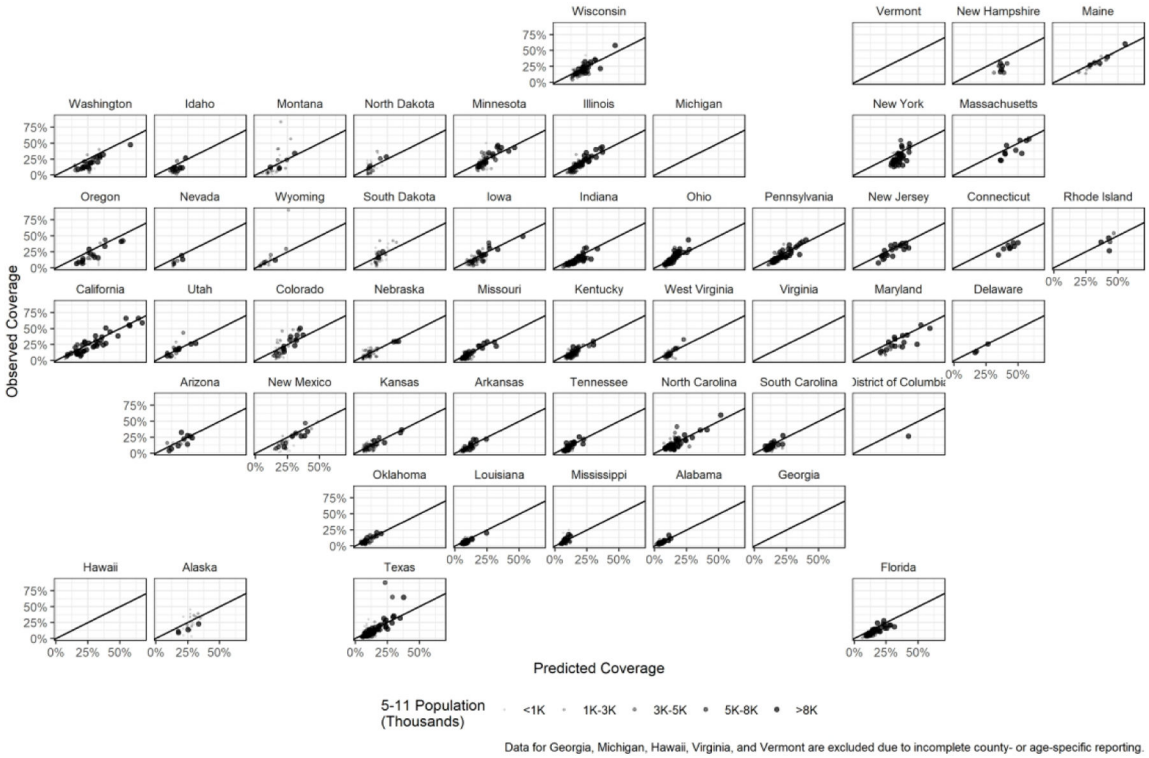


Figure 4.

Three-month validation of county-level predicted versus observed complete coverage among children ages 5 to 11 y.

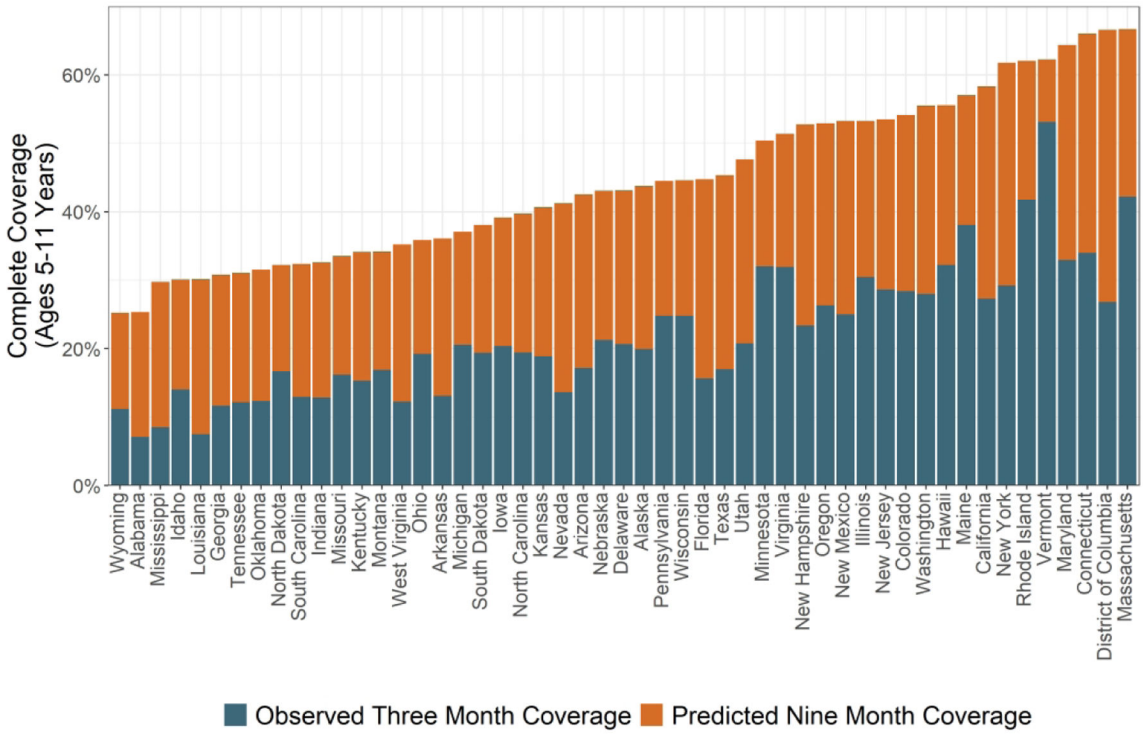


Figure 5.
State-level 3-mo complete vaccination scale-up progress for children ages 5 to 11 y and 9-mo predicted coverage.

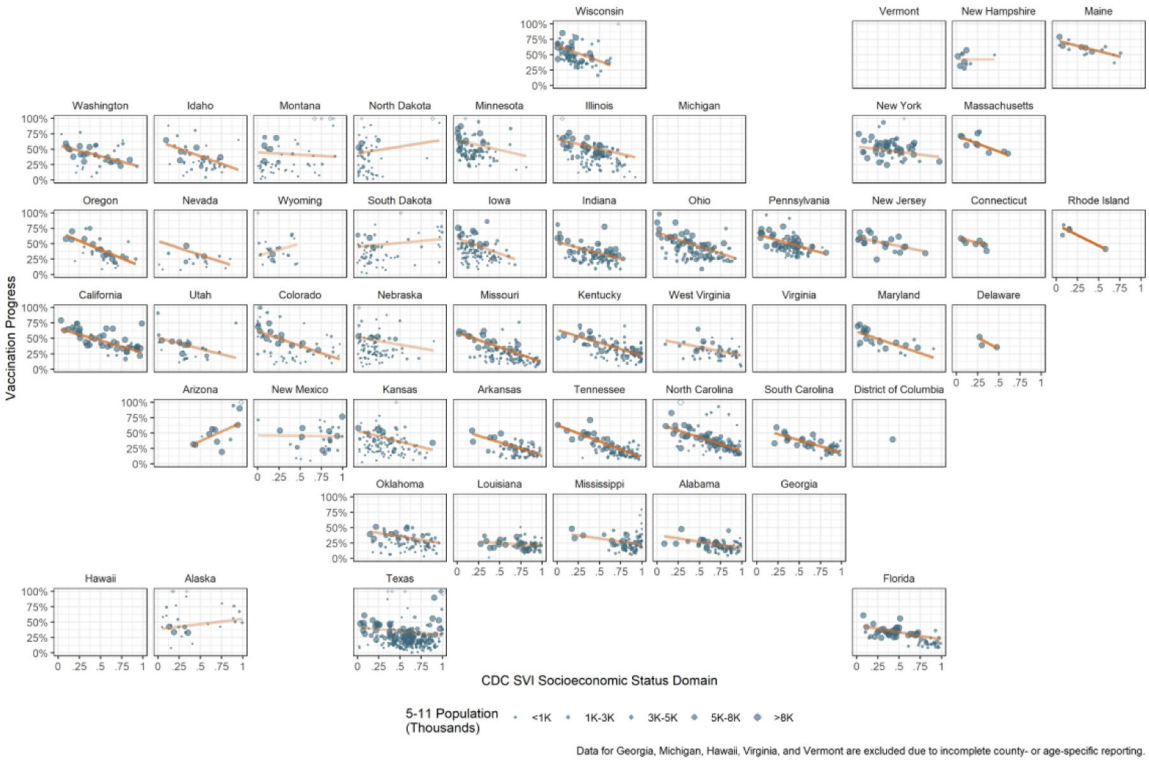


Figure 6.

Association between 3-mo county-level complete vaccination progress for children ages 5 to 11 y and the socioeconomic status domain of CDC's Social Vulnerability Index. The intensity of the trend lines is proportional to the linear regression R2.

Table 1

Unweighted and Weighted Distribution of Sociodemographic Variables of Included Guardians from the COVID-19 Trends and Impact Survey (CTIS), Compared with Distribution of Sociodemographic Variables of Parents/Guardians in the American Community Survey (ACS), Weighted by Children of Target Age Groups

	CTIS Survey Data: Unweighted	CTIS Survey Data: Weighted[a]	Parents/Guardians of Children Ages 12–17 y	Parents/Guardians of Children Ages 5–11 y
Documented Gender %[b]				
Female	65.4	53.9	55.4	55.4
Male	34.6	46.1	44.6	44.6
Age Group, y, %				
18–24	1.9	5.8	0.7	1.4
25–34	13.7	17.7	8.4	29.6
35–44	29.0	28.3	42.1	48.7
45–54	25.6	24.4	39.3	16.8
55–64	16.0	13.6	7.9	2.5
65+	13.7	10.2	1.6	1.0
Education, %				
High school or less	21.7	25.1	37.1	35.9
Some college or 2-y degree	35.9	35.9	29.7	29.8
Four-year degree	23.2	21.7	20.4	20.7
Graduate degree	19.2	17.3	12.8	13.6
Race/Ethnicity, %				
American Indian or Alaska Native	1.1	1.0	0.7	0.7
Asian	2.8	3.4	5.4	5.9
Black	7.1	7.2	11.5	11.5
Hispanic	17.3	21.7	22.0	23.0
Native Hawaiian or Other Pacific Islander	0.3	0.3	0.2	0.2
Other	4.8	5.2	1.8	2.0
White	66.6	61.2	58.4	56.7

[a]Weighted using sampling weights from CTIS, to be representative of the population ages 18 y and older.

[b]CTIS collects information on the respondent's self-reported gender, whereas the ACS collects information on the respondent's self-reported sex.