Modeling State-Level Attitudinal Measures Related to COVID-19 Vaccination of Children Using Survey Data and Administrative Data

N. Ganesh[1], Shalima Zalsha[1], Vicki Pineau[1]
Elizabeth Allen[1], Zachary H. Seeskin[1], Kirk M. Wolter[1], Tammy Santibanez[2], James A. Singleton[2], Michael Chen[2], David Yankey[2], Yi Mu[2], Tianyi Zhou[23], Anurag Jain[23]
[1]NORC at the University of Chicago, 4350 East-West Highway, Bethesda MD 20814
[2]National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30329
[3]Leidos Inc., 2295 Parklake Drive NE #300, Atlanta, GA 30345

Abstract
The National Immunization Survey-Adult COVID Module (NIS-ACM) and National Immunization Survey-Child COVID Module (NIS-CCM) are random-digit dialing cellular telephone surveys of households, sponsored by the CDC, used to generate weekly and monthly COVID-19 attitudinal measures for adults 18 years and older and children 6 months to 17 years. This paper discusses the use of small area estimation models and methods to generate estimates by state for 5 to 17-year-old children for key attitudinal measures (as reported by the parent or guardian) including vaccination status and intent, COVID-19 vaccine safety, importance of getting vaccinated, and mask wearing in past seven days. Cross-sectional, bivariate, and time-series small area models (i.e., linear mixed models) were used to combine (a) direct survey estimates from the NIS-CCM, (b) direct survey estimates from the NIS-ACM, and (c) regression estimates based on auxiliary data sources.

Key Words: COVID-19, composite estimation, National Immunization Survey, small area estimation

The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

1. Introduction

The National Immunization Survey (NIS) is a family of random digit dialing surveys conducted by NORC at the University of Chicago for the Centers for Disease Control and Prevention (CDC) to estimate vaccination coverage for children of various ages. Beginning April 2021, CDC added the NIS-Adult COVID Module[1] (NIS-ACM); in July 2021, the NIS-Child COVID Module[2] (NIS-CCM) was added. Both NIS-ACM and the NIS-CCM

---

[1] https://data.cdc.gov/Vaccinations/National-Immunization-Survey-Adult-COVID-Module-NI/udsf-9v7b
[2] https://data.cdc.gov/Vaccinations/National-Immunization-Survey-Child-COVID-Module-NI/uny6-e3dx

have been conducted continually and are used by CDC for weekly and monthly monitoring of vaccination coverage, barriers to vaccination, vaccine hesitancy, and social attitudes and behaviors associated with COVID-19.

NIS-ACM interviews adults 18+ years and utilizes a sample design that allows for monthly state-level and select local area-level estimation. NIS-CCM was initially designed to survey children 13-17 years, but later expanded in October 2021 to survey children 5-17 years old and then again in December 2021 to survey children 6 months-17 years. Unlike NIS-ACM, NIS-CCM does not have sufficient sample to produce monthly state-level estimates and is designed to produce estimates for ten HHS regions[3] rather than state[4].

While the NIS-CCM survey design does not use a large enough sample to produce state-level estimates, this paper reports on the use of small area estimation (SAE), which combines survey estimation with statistical modeling, to produce state-level estimates of COVID-19 attitudinal measures for 5- to 17-year-old children. Specifically, the objective of this paper is to develop and assess SAE models which utilize the correlation in the estimates between children 5-17 years from NIS-CCM and adults 18+ years from NIS-ACM to generate monthly state-level estimates for 5-17 year old children. The goal of SAE is to combine direct survey estimates based on data available from the geographic area, with a model-based estimate based upon the relationship between area-level estimates and covariates (Rao, 2015). The small area estimate for a basic area-level model is a weighted average of the direct survey estimate and the model-based prediction for an area, where the weights of the two components are proportional to their estimated precision.

We evaluate three SAE models to generate state-level estimates. The paper is organized as follows: Section 2 provides a description of the modeled variables and discusses the data sources that were used for the SAE models. Section 3 provides a description of the three SAE models. An evaluation of the three SAE models is presented in Section 4. Some preliminary results are provided in Section 5. Section 6 provides a summary and discusses some limitations of the methods.

2. Modeled Variables and Data Sources

2.1 Modeled COVID-19 Attitudinal Variables

The response variables considered for the SAE models were state-level direct survey estimates from NIS-ACM and NIS-CCM for the months of January, February, and March 2022. Table 1 presents a list of the modeled variables[5]. Survey questions associated with each modeled variable (including one or more doses of COVID-19 vaccine) were administered in both the NIS-ACM and the NIS-CCM modules.

---

[3] See https://www.hhs.gov/about/agencies/iea/regional-offices/index.html

[4] The NIS-ACM was reviewed by CDC and was conducted consistent with applicable federal law and CDC policy (e.g., 45 C.F.R. part 46.102(l)(2), 21 C.F.R. part 56; 42 U.S.C. §241(d); 5 U.S.C. §552a; 44 U.S.C. §3501 et seq.).

[5] For purposes of evaluating the different SAE models and to compare against an external benchmark (obtained from CDC's vaccine administration data), one or more doses of COVID-19 vaccine (see Section 4) was also modeled.

Table 1: List of Modeled COVID-19 Attitudinal Variables.

Description of Modeled Variable
Lack of confidence in COVID-19 vaccine safety (not at all/somewhat safe)
Importance of getting COVID-19 vaccine (not at all/a little important)
Unvaccinated, definitely/probably/unsure will get COVID-19 vaccine
Unvaccinated, definitely not/probably not get COVID-19 vaccine
Not at all/a little concerned about getting COVID-19
Mask wearing (never/rarely)
Provider recommendation of COVID-19 vaccine
None/some friends and family have received a COVID-19 vaccine
Difficulty getting the COVID-19 vaccine (not at all/a little)
Difficulty getting the COVID-19 vaccine (not at all/a little) among unvaccinated children

2.2 Auxiliary Data Sources

Data sources and state-level covariates used in all SAE models are presented in Table 2 below.

Table 2: List of State-Level Covariates Used in All SAE Models

Data Source	*Variable*
2020 American Community Survey (ACS)	- Proportion of adults 18+ years and children 5-17 years who are Hispanic - Proportion of adults 18+ years and children 5-17 years who are non-Hispanic White - Proportion of adults 18+ years and children 5-17 years who are non-Hispanic Black - Proportion of adults 18+ years and children 5-17 years who live in a rented household - Proportion of adults 18+ years and children 5-17 years who live in poverty - Proportion of adults 18+ years and children 5-17 years who are U.S. citizens - Proportion of adults 18+ years and children 5-17 years who moved in the past year - Proportion of adults 18+ years and children 5-17 years who have health insurance - Proportion of adults 18+ years and children 5-17 years who have private health insurance - Proportion of adults 18+ years and children 5-17 years who have access to the internet - Proportion of adults 18-30 years[6] - Proportion of adults 31-44 years[6] - Proportion of adults 65+ years[6] - Proportion of adults 18+ years who do not have a high school degree[6] - Proportion of adults 18+ years who have a high school degree[6] - Proportion of adults 18+ years who have a college degree or higher[6] - Proportion of adults 18+ years who are unemployed[6]

[6] These variables were only derived for adults 18+ years but included as covariates in the SAE models for children 5-17 years.

	-	Proportion of adults 18+ years who are married[6]
CDC vaccine administration data	-	Proportion of adults 18+ years and children 5-17 years who had received one or more doses of COVID-19 vaccine[7]
2020 Presidential Election Data	-	Proportion of the population that voted for Joseph Biden[6]

3. Small Area Models

SAE models were specified and fit relating direct survey estimates of each COVID-19 attitudinal measure listed in Table 1 to covariates thought to be related to these measures. Three potential SAE models were considered (Fay and Herriot, 1979, Rao and Molina, 2015):

1. Fay-Herriot model[8]
2. Bivariate Fay-Herriot model[9]
3. Bivariate time-series Fay-Herriot model[10]

For the Fay-Herriot model, state-level direct survey estimates from NIS-CCM for each variable listed in Table 1 were modeled using as potential covariates the variables listed in Table 2. On average, for a given month and state, there were 140 completed interviews from NIS-CCM. The NIS-ACM has a much larger sample size, with an average 1,200 completed interviews for each month and state. Given the much larger sample size associated with NIS-ACM and the correlation between state-level adult and child COVID-19 attitudinal measures, bivariate models that jointly model direct survey estimates from NIS-ACM and NIS-CCM are likely to yield more precise estimates for children 5-17 years. Finally, given the continual data collection for both NIS-ACM and NIS-CCM, using bivariate time-series models that jointly model multiple months of NIS-ACM and NIS-CCM direct survey estimates may yield attitudinal estimates for 5-17 year old children with smaller mean squared error compared to the other two models.

A more detailed description is provided for the bivariate time-series Fay-Herriot model below. The bivariate Fay-Herriot model is similar but excludes the autoregressive time-series component, while the Fay-Herriot model is similar but also excludes the modeling of the joint distribution with the adult COVID-19 attitudinal measure.

Typically, when modeling proportions, the direct survey estimates are transformed to preserve the bounds of 0 and 1 for a proportion. A logit-transformation was used, with the transformed direct survey estimate given by:

$$y_{ijt} = log\left(\frac{z_{ijt}}{1-z_{ijt}}\right),$$

[7] These proportions were for the reference dates of January 15, February 15, and March 15, correspond to mid-month for the monthly direct survey estimates from NIS-ACM & NIS-CCM.
[8] Implemented in R using the sae https://cran.r-project.org/web/packages/sae/index.html and msae https://cran.r-project.org/web/packages/msae/index.html packages
[9] Implemented in R using the msae package
[10] Implemented in R using author's code

where z_{ijt} is the direct survey estimate for a given attitudinal variable of interest for state *i*, domain *j* (5-17 years, 18+ years), and month *t* (January, February, March). Since y_{ijt} is undefined when z_{ijt} = 0 or 1, the direct survey estimates were truncated to 0.005 if $z_{ijt} \leq 0.005$ or 0.995 if $z_{ijt} \geq 0.995$. The logit-transformed direct survey estimate was modeled as:

Level 1 (sampling model): $y_{ijt} = \theta_{ijt} + e_{ijt}$
Level 2 (linking model): $\theta_{ijt} = x'_{ijt}\beta + v_{ij} + u_{ijt}$

where

- Level 1 captures the sampling variability associated with the direct survey estimates y_{ijt}, and θ_{ijt} is the true (unknown) mean for the attitudinal variable of interest;
- e_{ijt} is the sampling error with $e_{ijt} \sim N(0, \psi_{ijt})$, where ψ_{ijt} is assumed known and the sampling errors are independent across *i*, *j*, *t*;
- Level 2 links the true mean for the attitudinal variable of interest to a set of covariates x_{ijt}; as potential covariates, the variables listed in Section 2.2 were considered;
- v_{ij} is a random effect that is independent across states

$$\binom{v_{i1}}{v_{i2}} \sim N\left(\binom{0}{0}, \begin{pmatrix} \sigma_{v1}^2 & \rho_v\sigma_{v1}\sigma_{v2} \\ \rho_v\sigma_{v1}\sigma_{v2} & \sigma_{v2}^2 \end{pmatrix}\right);$$

- u_{ijt} is independent across domains and for a given domain, u_{ijt} is assumed to follow a first-order autoregressive model;

$$u_{ijt} = \rho_j u_{ij(t-1)} + \epsilon_{ijt}$$

  where $|\rho_j| < 1$ and $\epsilon_{ijt} \sim N(0, \sigma_{uj}^2)$. For computational convenience, it was assumed that $\rho_1 = \rho_2 = \rho$ and $\sigma_{u1}^2 = \sigma_{u2}^2 = \sigma_u^2$; and
- e_{ijt}'s, v_{ij}'s, and ϵ_{ijt}'s are pairwise mutually independent.

Note that the above model can also be specified as

$$y_{ijt} = x'_{ijt}\beta + v_{ij} + u_{ijt} + e_{ijt}$$

Most of covariates listed in Section 2.2 do not vary by month; the only time-varying covariate is one or more doses of COVID-19 vaccine. Thus, the regression parameters in the bivariate time-series model were allowed to vary by month (to allow for time-varying regression estimates). The variance parameters $(\sigma_{v1}^2, \sigma_{u1}^2, \rho_1, \sigma_{v2}^2, \sigma_{u2}^2, \rho_2, \rho_v)$ were estimated using the restricted maximum likelihood estimator.

Typically, $\theta_{ijt} = x'_{ijt}\beta + v_{ij} + u_{ijt}$ is the parameter of interest in SAE. However, for the specified model, θ_{ijt} is the true logit-transformed proportion. Thus, after deriving the empirical best linear unbiased predictor (EBLUP; see Rao and Molina, 2015) for θ_{ijt}, that estimate was transformed to obtain an estimate for the proportion. This estimated proportion is the final model-based estimate for a given attitudinal variable.

The bivariate Fay-Herriot model for a specific month is given by

$$y_{ij} = x'_{ij}\beta + v_{ij} + e_{ij}$$

where

- y_{ij} is the logit transformed direct survey estimate for a given attitudinal variable; of interest for a given month for state *i* and domain *j* (5-17 years, 18+ years);
- x_{ij} is a vector of covariates;
- e_{ij} is the sampling error with $e_{ij} \sim N(0, \psi_{ij})$, where ψ_{ij} is assumed known and the sampling errors are independent across *i*, *j*; and
- v_{ij} is a random effect that is independent across states.

$$\begin{pmatrix} v_{i1} \\ v_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{v1}^2 & \rho_v \sigma_{v1} \sigma_{v2} \\ \rho_v \sigma_{v1} \sigma_{v2} & \sigma_{v2}^2 \end{pmatrix}\right).$$

Finally, the Fay-Herriot model for a specific month is given by

$$y_i = x_i' \beta + v_i + e_i$$

where

- y_i is the logit transformed direct survey estimate for a given attitudinal variable; of interest for a given month for children 5-17 years for state *i*;
- x_i is a vector of covariates;
- e_i is the sampling error with $e_i \sim N(0, \psi_i)$, where ψ_i is assumed known and the sampling errors are independent across *i*; and
- v_i is a random effect that is independent across states with $v_i \sim N(0, \sigma_v^2)$.

The SAE methodology for the Fay-Herriot model is similar to the methodology previously used by the CDC and NORC to obtain vaccination coverage for 19- to 35-month-old children using National Immunization Survey-Child (NIS-Child) data (see Smith and Singleton, 2008 and 2011, Ganesh et al., 2016, Seeskin et. al., 2020).

4. Evaluation of the Models

Evaluation of the model performance was conducted using SAE models which were developed to predict one or more doses of COVID-19 vaccine for the month of March for children 5-17 years[11]. For the evaluation, the same covariates listed in Section 2.2, except for one or more doses of COVID-19 vaccine, were considered as potential covariates. The model-based prediction from each model for the month of March for 5-17 year old children by state was compared against the external benchmark for one or more doses of COVID-19 vaccination coverage from CDC's vaccine administration data[12]. The results are presented in Table 3. In Table 3:

- "correlation" is the correlation between the model-based prediction from a specific model and the benchmark,
- "mean deviation" is the average difference between the model-based prediction from a specific model and the benchmark (where the average is computed over all states, and the estimate is a proportion),
- "mean absolute deviation" is the average absolute difference between the model-based prediction from a specific model and the benchmark,

[11] Note that monthly NIS-CCM data were calibrated within each HHS Region to the external benchmark for one or more doses of COVID-19 vaccination coverage from CDC's vaccine administration data.

[12] https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc

- "95% confidence interval coverage" is the proportion of states for which the model-based prediction from a specific model and its associated 95% confidence interval includes the benchmark, and
- "mean RMSE" is the average estimated root mean squared error (RMSE) under the specific model for predicting the proportion of children who have received one or more doses of COVID-19 vaccine.

Compared to the Fay-Herriot model, both bivariate Fay-Herriot models produced estimates that were more correlated with the benchmark measure. Estimates from both bivariate models also have a lower mean absolute deviation, though the RMSE for the bivariate time-series model appears to be underestimated as shown by the 95% confidence interval coverage being less than its nominal value.

Table 3: Summary of the Model Evaluation.

Model	*Correlation*	*Mean Deviation*	*Mean Absolute Deviation*	*95% Confidence Interval Coverage*	*Mean RMSE*
Fay-Herriot model	0.910	-0.0034	0.042	0.941	0.048
Bivariate model	0.944	-0.0023	0.037	0.961	0.043
Bivariate time-series model	0.954	-0.0023	0.033	0.843	0.034

5. Results

Parameter estimates from the bivariate time-series Fay-Herriot models are presented in Table 4. The results suggest moderate to strong correlations between child and adult state-level random effects for all modeled variables except for 'unvaccinated, definitely not/probably not get COVID-19 vaccine'. One explanation for this may be that the covariates in the model for 'unvaccinated, definitely not/probably not get COVID-19 vaccine' are explaining the state-to-state variation along with the correlation across child/adult, and thus, the residuals are approximately uncorrelated across child/adult. The variance of the time-varying random effect is estimated to be zero or near zero in all cases such that the time effect can be thought of as a fixed effect (captured in the regression estimate) rather than a random effect.

Table 4: Estimates of the Variance Parameters for the Bivariate Time-Series Fay-Herriot Model.

Variable	$\boldsymbol{\sigma_{v1}}$	$\boldsymbol{\sigma_{v2}}$	$\boldsymbol{\rho_v}$	$\boldsymbol{\sigma_u}$	$\boldsymbol{\rho}$
Lack of confidence in COVID-19 vaccine safety	0.004	0.014	0.426	0.000	0.000
Importance of getting COVID-19 vaccine (not at all/a little important)	0.012	0.032	0.267	0.000	0.000
Unvaccinated, definitely/probably/unsure will get COVID-19 vaccine	0.015	0.067	0.957	0.000	0.000
Unvaccinated, definitely not/probably not get COVID-19 vaccine	0.018	0.171	0.094	0.000	0.000
Not at all/a little concerned about getting COVID-19	0.011	0.007	1.000	0.000	0.000
Mask wearing (never/rarely)	0.208	0.205	1.000	0.030	0.564
Provider recommendation of COVID-19 vaccine	0.005	0.007	0.709	0.000	0.000
None/some friends and family have received a COVID-19 vaccine	0.023	0.023	0.342	0.000	0.000

Difficulty getting the COVID-19 vaccine (not at all/a little)	0.045	0.057	1.000	0.000	0.000
Difficulty getting the COVID-19 vaccine (not at all/a little) among unvaccinated	0.070	0.012	1.000	0.016	0.086

Figure 1 provides a comparison of the state-level estimates between children 5-17 years and adults 18+ years for the attitudinal variable ‘lack of confidence in COVID-19 vaccine safety’. The three plots in the figure illustrate the relationship between the direct survey estimates for adult (from NIS-ACM) and three types of estimates for child: direct survey estimates (from NIS-CCM) in the first plot, the model-based estimates from the bivariate time-series model in the second plot, and the model-based estimates from the bivariate model in the third plot. As shown in the figure, both bivariate models appear to improve the correlation of the estimates between child and adult. This is because bivariate models utilize the correlation between the NIS-CCM and NIS-ACM direct survey estimates.

Figure 1: Comparison of the State-Level Estimates for ‘Lack of Confidence in COVID-19 Vaccine Safety’, March 2022.

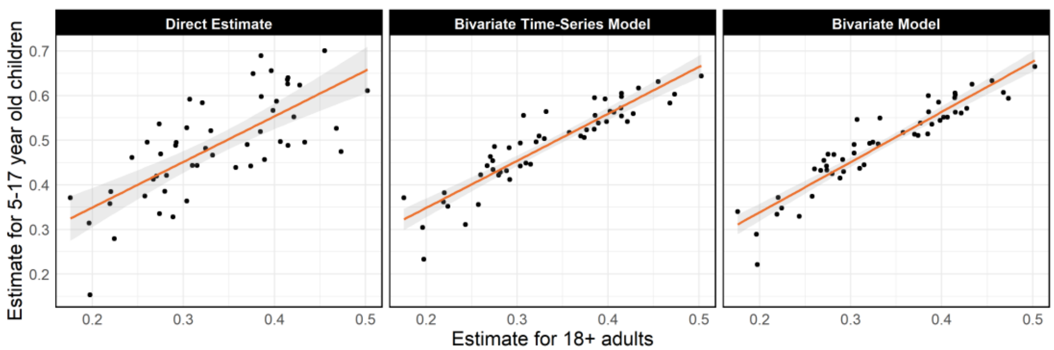
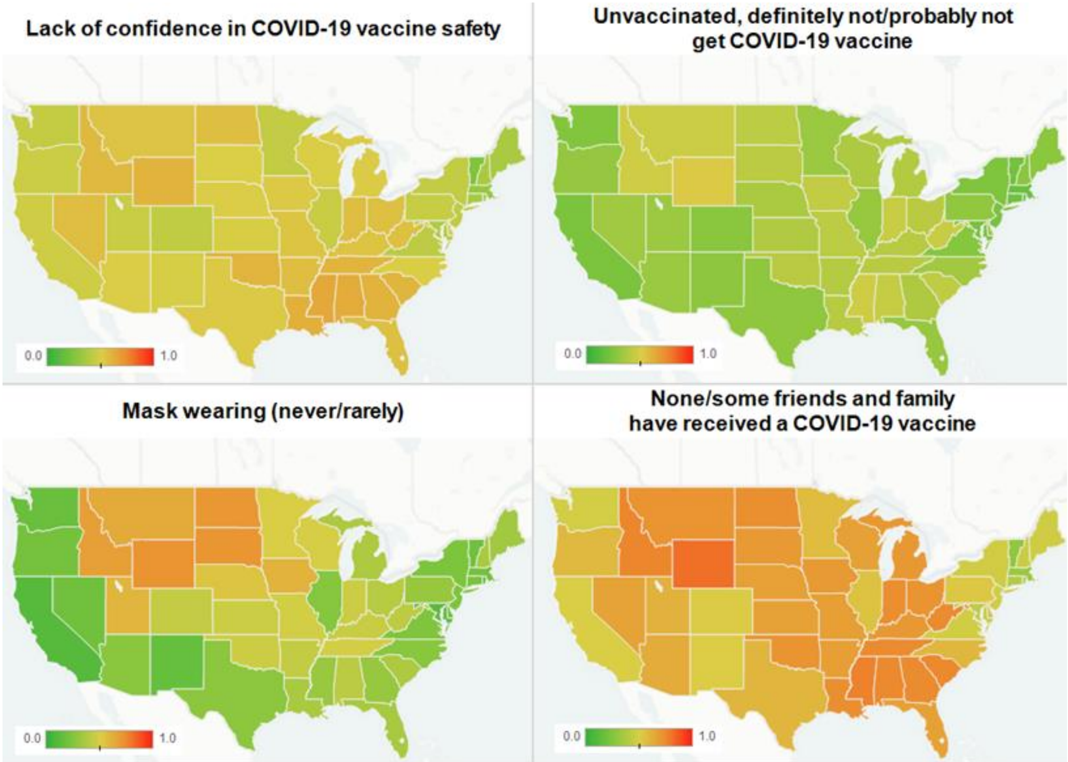


Figure 2. illustrates the state-level estimates for four of the COVID-19 attitudinal measures for children 5-17 years for March, 2022, generated using the bivariate time-series Fay-Herriot model. The results suggest large variation in mask-wearing across states and more moderate variation across states for the other three modeled variables.

Figure 2. State-Level Estimates for Children 5-17 Years from the Bivariate Time-Series Fay-Herriot Model, March 2022



6. Summary and Limitations

State-level estimates of COVID-19 attitudinal measures were produced using SAE models for children 5-17 years. Three potential SAE models were evaluated. The evaluation suggested both bivariate models performed best, though the bivariate model might be preferred over the bivariate time-series model based on the 95% confidence interval coverage. However, there are some limitations associated with these state-level estimates:

1. The SAE models are subject to model error as a result of having an imperfect relationship between the modeled attitudinal variable and the set of selected covariates.
2. There is measurement error in one or more doses of COVID-19 vaccine coverage sourced from the vaccine administration data.
3. Except for the covariate one or more doses of COVID-19 vaccine coverage, all other covariates are not time-varying. This was addressed by allowing for time-varying regression parameters.
4. Random effect variances (i.e., $\sigma_{v1}^2, \sigma_{v2}^2, \sigma_u^2$) were close to 0. Thus, model-based estimates were "weighted" toward the regression estimates.

Finally, a couple of improvements may be considered to refine these SAE models in the future. First, ratio-adjustment would allow the model-based estimates to agree with national or HHS Region level direct survey estimates from NIS-CCM. Second, NIS-ACM direct survey estimates could be incorporated into the SAE models via a measurement error model (Ybarra and Lohr, 2008).

References

Fay, R.E. and Herriot, R.A. 1979. Estimates of income for small places: An application of James Stein procedures to Census data. Journal of the American Statistical Association, 74, 269–277.

Ganesh, N., Chakraborty, A., Copeland, K.R., Wolter, K.M., Santos, K., Liu, L., Smith, P. J., Yankey, D., Jeyarajah, J., Santibanez, T.A., Singleton, J.A., Greby, S., Elam-Evans, L.D., and Smith, C.S. 2016. Small-area estimation methods for county-level vaccination coverage rates using the NIS. Presentation at the 2016 Joint Statistical Meetings.

Rao, J.N.K. and Molina, I. 2015. Small area estimation. New York: Wiley.

Seeskin, Z.H., Ganesh, N., Maitra, P., Herman, P., Wolter, K.M., Copeland, K.R., English, N., Chen, M.P., Singleton, J.A., Santibanez, T.A., Yankey, D., Elam-Evans, L.D., Sterrett, N., Smith, C.S., Gibson, K., and Meador, S. 2020. Estimating county-level vaccination coverage using small area estimation with the National Immunization Survey/-Child. In JSM Proceedings, Survey Research Methods Section, Alexandria, VA: American Statistical Association.

Smith, P.J. and Singleton, J.A. 2008. Vaccination coverage for selected counties: Attainment of the healthy people 2010 goals and association with county-level indices of access to care, economic conditions, and demographic composition. Public Health Reports, 123, 155-172.

Smith, P.J. and Singleton, J.A. 2011. County-level trends in vaccination coverage for selected counties in the U.S., children aged 19-35 months, United States, 1995-2008: attainment of the healthy people 2010 objectives and association with county-level indices of access to care, economic conditions, and demographic composition. Morbidity and Mortality Weekly Report. Surveillance Summaries, 60(4), 1-86.

Ybarra, L.M.R. and Lohr, S.L. 2008. Small area estimation when auxiliary information is measured with error. Biometrika, v. 95(4), 919-931.