HHS Public Access

Author manuscript

Stat Methods Appt. Author manuscript; available in PMC 2024 December 11.

Published in final edited form as:

Stat Methods Appt. 2022 December ; 33: 1171–1191. doi:10.1007/s10260-022-00678-7.

Hierarchical Bayes small area estimation for county-level health prevalence to having a personal doctor

Erciulescu Andreea L.[*,1], **Li Jianzhu**[1], **Krenzke Tom**[1], **Town Machell**[2]

[1]Westat, Maryland, United States

[2]Population Health Surveillance Branch, Division of Population Health, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, Georgia, United States

Summary

The complexity of survey data and the availability of data from auxiliary sources motivate researchers to explore estimation methods that extend beyond traditional survey-based estimation. The U.S. Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System (BRFSS) collects a wide range of health information, including whether respondents have a personal doctor. While the BRFSS focuses on state-level estimation, there is demand for county-level estimation of health indicators using BRFSS data. A hierarchical Bayes small area estimation model is developed to combine county-level BRFSS survey data with county-level data from auxiliary sources, while accounting for various sources of error and nested geographical levels. To mitigate extreme proportions and unstable survey variances, a transformation is applied to the survey data. Model-based county-level predictions are constructed for prevalence of having a personal doctor for all the counties in the U.S., including those where BRFSS survey data were not available. An evaluation study using only the counties with large BRFSS sample sizes to fit the model versus using all the counties with BRFSS data to fit the model is also presented.

[*]**Correspondence**: Andreea L. Erciulescu, 1600 Research Blvd., Rockville, Maryland, 20850. AndreeaErciulescu@westat.com.

Author contributions

Dr. Erciulescu managed the development and implementation of the small area estimation models, reviewed and provided guidance on the implementation of weighting adjustments, imputation, survey direct estimation, and variable selection, and prepared the initial draft of this manuscript.

Dr. Li led the raking adjustments to survey weights, imputation, creation of survey estimates at various geographical levels, preparation of model covariates from different data sources, implementation of variable selection for the small area estimation models, as well as reviewed the small area predictions and diagnostics.

Mr. Krenzke managed the project in collaboration with the CDC, provided general guidance to align with available resources, and reviewed the various contributions to the project by Dr. Erciulescu and Dr. Li.

Dr. Town developed the initial scope of work, provided oversight from CDC and reviewed all manuscript revisions and reporting documents. She oversaw the production of data from the BRFSS on which this research is based.

Conflict of interest

The authors declare no potential conflict of interests.

Financial disclosure

The work described in this paper was conducted under contract with the Centers for Disease Control and Prevention (CDC Contract #HHSD2002013M53968B Order #75D30120F09442).

Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Keywords

Behavioral Risk Factor Surveillance System; disaggregation; hierarchical Bayes; multiple data sources; nested levels

1 | INTRODUCTION

The U.S. Centers for Disease Control and Prevention's (CDC's) Behavioral Risk Factor Surveillance System (BRFSS) collects a wide range of health information, including whether respondents have a personal doctor. BRFSS data are used by state health departments to plan interventions, allocate scarce resources, and provide information to the general public. While the BRFSS focuses on state-level estimation, there is demand for county-level estimation of health indicators using BRFSS data. The ability to make the most productive use of the BRFSS data and to improve county-level planning based on BRFSS data is essential.

The overall large sample size of BRFSS allows for county-level survey-based estimation of health indicators. However, survey-based estimates can only be produced for counties with sample survey data available and are subject to high uncertainty for counties with small sample sizes. The sparsity of the survey data in some counties and the availability of data from auxiliary sources motivate researchers to explore estimation methods that extend beyond the traditional survey-based estimation. For example, model-based small area estimation (SAE) provides a principled way to combine survey and auxiliary data, while exploring the relationship between the outcome of interest and the auxiliary information, and accounting for the sources of uncertainty in the model components.

There are two main classes of SAE models: unit-level models, introduced in Battese, Harter, and Fuller (1988) [1], and area-level models, introduced in Fay and Herriot (1979) [2]. The first class of SAE models take as input unit-level survey and auxiliary data, while the second class of SAE models take as input area-level survey estimates and auxiliary data. The area-level SAE models are often preferred because of their practical applicability, while overcoming challenges related to the availability of auxiliary data, linking between the survey unit and the auxiliary source unit, inclusion of survey design effects, and others (for example, definition of unit and computational resources). These models are fit to the set of areas with sample data, but estimates can be constructed for all the areas of interest.

In this paper, an area-level SAE model is developed to combine county-level BRFSS survey data with county-level data available from auxiliary sources, for the purpose of estimating the proportions of individuals who do not have a personal care provider or doctor. The areas of interest comprise of the set of U.S. counties (3,142 counties), but the model is fit only to the set of counties for which BRFSS data on having or not a personal care provider or doctor are available (3,115 counties). For evaluation purposes, the model is also fit to the set of counties with BRFSS sample size of at least 500, also known as the Selected Metropolitan/Micropolitan Area Risk Trends (SMART) counties (213 counties). Model-based county-level estimates are produced for all the U.S. counties.

SAE models have been considered in the past for estimating BRFSS quantities. We first review some of the unit-level models studies. For county-level estimation of prevalence of self-reported diagnosed diabetes, Cadwell et al. (2010)[3] developed a census region-specific model taking as input BRFSS survey data at a level defined by the cross-tabulation of age, sex, race/ethnicity, and county; this model fits under the unit-level modeling framework. For county-level estimation of chronic obstructive pulmonary disease prevalence, Zhang et al. (2014) [4] developed a model taking as input BRFSS survey data at a level defined by the cross-tabulation of age, sex, race/ethnicity, county, and state, with extensions to finer geographies including census blocks and tracts; this model fits under the unit-level modeling framework. The model in Zhang et al. (2014) [4] was then applied to BRFSS data on health status and access indicators, after removing sex and state from the definition of the domain level (see Pierannunzi et al., 2016[5]), to BRFSS data on colorectal cancer screening prevalence (see Berkowitz et al., 2018 [6]), and to BRFSS data on mammography screening rates, after removing sex from the definition of the domain level (see Berkowitz et al., 2019 [7]). Holt et al. (2019) [8] developed a similar model to the one in Zhang et al. (2014) [4] for estimating the number of at-risk community-dwelling adults with a chronic condition (chronic obstructive pulmonary disease), while combining BRFSS and National Hurricane Center (NHC) data. As documented in these studies, the current methods are not capturing the BRFSS complex survey design effects, hence are lacking at accounting for a key source of uncertainty in the survey data. Moreover, the pool of auxiliary data is constrained to BRFSS demographic variables for most of these studies. Unlike these studies, we explore a large pool of auxiliary data and adopt an area-level modeling approach that starts with BRFSS survey estimates and accounts for the sampling variability in the survey data.

Raghunathan et al. (2007) [9] developed a joint county-level model for BRFSS and the National Health Interview Survey (NHIS) current-smoking and mammography screening rates; this model fits under the area-level modeling framework. For estimating county-level smoking and screening for female breast cancer, cervical cancer, and colorectal cancer rates, Liu et al. (2019) [10] extended the model in Raghunathan et al. (2007) 9 that com bines BRFSS and NHIS data. Like in Raghunathan et al. (2007) [9], we also model the arcsine-square-root-transformed survey-based county-level estimates, in order to mitigate unstable survey variances and help support a normality assumption for the survey-based county-level esti mates. As a result, the sampling variances for the transformed survey estimates are functions of the effective sample sizes. For counties with small sample sizes, the authors imputed the design effect and set a lower bound for the effective sample sizes. To mitigate such scenarios, as well as extreme survey-based proportions, we approximate the effective sample sizes using Kish's approximation (see Kish, 1965 [11]).

Multi-fold SAE models have been studied in the literature to further account for the nested data structure of the survey data and allow for reliable estimation at various levels of aggregation. Area-level multi-fold SAE models were initially introduced in an unpublished manuscript by Fuller and Goyeneche (1998)[12], as two-fold models, and then considered in Torabi and Rao (2014)[13]. Both of these studies employed frequentist estimation of model parameters. Erciulescu, Cruze, and Nandram (2020)[14] specified a two-fold SAE model as a hierarchical Bayes model and derived closed-form expressions for model predictions for counties with survey data. Krenzke et al. (2020) [15] applied three-fold SAE models specified

as hierarchical Bayes models to predict U.S. adult competency at the county-level. We consider Bayesian inference to be a straightforward inferential approach for multi-fold SAE models, while allowing for the construction of full posterior distributions for quantities of interest. In addition, we model county-level BRFSS and auxiliary data, while accounting for the nested structure of counties within states and states within census divisions and allowing for reliable estimation at these three levels: county, state, and census division. In this aspect, the hierarchical Bayes SAE model developed in this paper is similar in structure to the ones considered in Krenzke et al. (2020) [15].

The rest of the paper is organized as follows. In Section 2, we describe the data available for the application study and the initial survey estimation and variable selection steps necessary to prepare the model input data. The hierarchical Bayes SAE model is presented in Section 3, along with a framework for model fit, internal validation, estimation and prediction, and comparison. Final selected results are provided in Section 4 and include external validation checks. A general discussion is given in Section 5.

2 | DATA FOR THE APPLICATION STUDY

The survey data for the application consist of micro-level BRFSS data for the reference year 2018, subject to unit-level adjustments described next. Using the BRFSS information on whether or not an individual has a personal care provider or doctor, an indicator variable is constructed having these two categories. The item nonresponse rate for this indicator variable is 0.64%. Missing values are imputed using a hot-deck imputation approach. The survey weights are adjusted using a raking procedure to align the county-level population totals estimated from the survey to corresponding fixed controls available from the 2014–2018 American Community Survey (ACS) Summary File (also known as Detailed Tables). The resulting micro-level data are used to produce county-level survey estimates for proportions of individuals who do not have a personal care provider or doctor. Specifically, Hájek-type point estimators and associated Taylor series variance estimators are constructed for all the counties with available survey data. Let these be denoted by county-level pairs $(\hat{p}_{ijk}, \widehat{V}_{ijk})$ where $i = 1, \ldots, m$ indexes the census divisions, $j = 1, \ldots, m_i$ indexes the states in census division i, and $k = 1, \ldots, m_{ij}$ indexes the counties in state j and census division i. Let the county-level population totals from the 2014–2018 ACS Summary File be denoted by t_{ijk}.

County-level survey estimates and associated variances on the arcsine-square-root scale serve as inputs into the models described in the next section. Let these be denoted by county-level pairs $(y_{ijk}, \sigma^2_{ijk})$ where

$$y_{ijk} := sin^{-1}\sqrt{\hat{p}_{ijk}},$$

and

$$\sigma^2_{ijk} := \frac{1}{4\tilde{n}_{e,ijk}}.$$

The county-level effective sample sizes are approximated as

$$\tilde{n}_{e,ijk} \approx \frac{\left(\sum_{a \in k} w_{ijka}^{f}\right)^2}{\sum_{a \in k} \left(w_{ijka}^{f}\right)^2}$$

where w_{ijka}^{f} is the adjusted survey weight associated with the individual a in county k.

The pool of variables related to the transformed proportion of individuals who do not have a personal care provider or doctor consists of demographic characteristics (i.e., race/ethnicity, age, sex, marital status), socioeconomic status (i.e., poverty, income, employment status, occupation), education (i.e., education, English-speaking ability), location (i.e., urbanicity), immigration status (i.e., length of stay for foreign-born people, migration), health (i.e., insurance, health professions, facilities), and other (i.e., journey to work, housing unit tenure, tax, birth rate, fertility rate, infant mortality rate, crime rate, Federal aid, energy consumption). A list of data sources and auxiliary variables is provided in the Appendix Table 1.

As the initial step in the variable selection process, we identify the most complete and less prone to error county-level auxiliary information. Next, a correlation analysis is conducted to reduce the pool of auxiliary variables to those identified to be highly correlated with the transformed proportion of individuals who do not have a personal care provider or doctor and minimally correlated with other auxiliary variables in the pool. Finally, a small set of auxiliary variables is selected using the least absolute shrinkage and selection operator (LASSO), with 20-fold cross-validation. Two choices are considered for the LASSO shrinkage/penalty parameters, both close to the optimal value that minimizes the mean cross-validated error but resulting in different numbers of selected variables. The two corresponding models constructed with these two sets of covariates will be referred to as the full model and reduced model, respectively. The full model contains a larger number of covariates than the reduced model does. Irrespective of the model, let the matrix of covariates be denoted by x_{ijk}; its rows correspond to counties with survey sample and its columns correspond to an intercept, followed by the selected variables. The exact number of counties with survey sample and the exact number of selected variables used to fit the models are discussed in the next section.

3 | MODELS

The overall structure of the proposed small area model follows closely the structure of the small area model for average adult proficiency presented in Krenzke et al. (2020) [15]. First, a sampling level is specified for the survey estimates on the transformed scale:

$$\text{Sampling level: } y_{ijk} | (\theta_{ijk}, \sigma_{ijk}^2) \sim N(\theta_{ijk}, \sigma_{ijk}^2),$$

where θ_{ijk} are the county-level quantities of interest on the transformed scale, i.e., the transformed proportions of individuals who do not have a personal care provider or doctor. Then, we borrow strength from auxiliary data and across small areas, while accounting for the nested structure of the data, via a linking level:

$$\text{Linking level: } \theta_{ijk}\big|(\beta, c_{ijk}, s_{ij}, d_i) = x_{ijk}\beta + c_{ijk} + s_{ij} + d_i,$$

$$c_{ijk}\big|(\sigma_c^2) \sim N(0, \sigma_c^2),$$

$$s_{ij}\big|(\sigma_s^2) \sim N(0, \sigma_s^2),$$

$$d_i\big|(\sigma_d^2) \sim N(0, \sigma_d^2),$$

where β is a vector of unknown regression coefficients, and c_{ijk}, s_{ij}, and d_i are the county, state, and census division latent effects, respectively, assumed to follow normal distributions with zero means and unknown variances σ_c^2, σ_s^2, and σ_d^2, respectively.

To fully specify the model as a hierarchical Bayes model, we adopt independent weakly informative priors for the unknown model parameters β, σ_c, σ_s, and σ_d:

$$\text{Priors: } \beta \sim N\left(0, 10^4\right), \text{ component} - \text{wise}$$

$$(\sigma_c, \sigma_s, \sigma_d) \sim \text{Cauchy}(0,5), \text{ component} - \text{wise}.$$

These choices of prior distributions ensure that little information about the values of these parameters is provided to the model, the data (likelihood) having the major role in shaping the posterior distributions. See Krenzke et al. (2020)[15], Browne and Draper (2006)[16], Gelman (2006) [17], and Polson and Scott (2012)[18], for some recent discussions about the choice of prior distribution for the scale parameter in a univariate hierarchical Bayes model.

3.1 | Fit

Four models are developed, depending on the number of counties with survey sample used and on the number of variables selected. For evaluation purposes only, we consider two sets of counties with survey data to fit the model presented above: one set comprises of all but 27 counties in the United States and corresponds to all the counties with BRFSS survey data (3,114 counties and the District of Columbia), and the other set comprises of counties with BRFSS sample size of at least 500 and corresponds to all the BRFSS SMART counties (213 counties). The variable selection steps are specific to the model, so the process presented above is conducted twice, each variable selection process corresponding to the two sets of counties used in the model fit. To summarize, we fit a pair of full and reduced models for all the counties with survey sample and another pair of full and reduced models for the SMART counties. The two pairs of full and reduced model fits are compared using the widely applicable information criterion (WAIC); smaller values indicate better model fit.

Each model is specified as an R STAN object and it is fit using Markov chain Monte Carlo (MCMC), with three chains, each chain starting with 20,000 samples, of which 5,000

samples are burned in. To reduce the autocorrelation in the chains used to make inference, we thin each of the chains so that every $10th$ iteration is kept and the rest are discarded. The final set of $R = 4,\ 500$ samples is used for inference.

Internal validation checks are conducted iteratively with the model development process for each model. In particular, we follow Erciulescu and Opsomer (2019) [19] and implement mixing and converge diagnostics for the MCMC sampler, residual diagnostics for the normality assumptions, and posterior predictive checks for the normality and linearity assumptions, as well as the correlation between the response variable and the first covariate. As a result, the tuning parameters used in the model fit are specific to each model. Also, the final model specification ensures that the R posterior samples are valid for inference.

3.2 | Prediction

Define in-sample counties as the set of counties used to fit the model. The rest of the counties of interest comprise the not-in-sample counties. The model predictions for the in-sample counties are composites of survey estimates and model-synthetic predictions based on the hierarchical Bayes model presented above. Survey estimates with large associated variance estimates (or small approximated effective sample sizes), relative to all the variance estimates, are smoothed more than the others, towards a common trend prediction that depends on the linear relationship assumed between the survey estimates and the covariates. As a result, the model predictions for such counties deviate from the survey estimates: point estimates may be either smaller or larger, and variance estimates are smaller. At the same time, also a result of the smoothing, the model predictions for counties with small variance estimates (or large approximated effective sample sizes) remain close to the survey estimates; both the point estimates and their associated variance estimates. For example, the model predictions are closer to the corresponding survey estimates for SMART counties, compared to the rest of the counties. Using the R samples $\theta_{ijk,\zeta}, \zeta = 1, \ldots, R$, from the posterior predictive distribution $[\theta_{ijk}|y, x, \sigma]$ posterior summaries such as means, variances, credible intervals, are constructed for the transformed proportion of individuals without a personal doctor, for all the counties.

Compared to the prediction for in-sample counties, prediction for not-in-sample counties relies more heavily on the model structure and on the covariates available for these counties. The contribution of the survey estimates to these predictions is only present via the nested structure of the model, because there is no contribution from the survey (the county-level survey estimates are either not available for these counties, or are not used as is the case with our model fitted only to the SMART counties); see Erciulescu et al. (2020)[14] for derived expressions for the model predictions under a two-fold small area estimation model.

For a not-in-sample county k^*, where the county belongs to a state j already represented in the set of in-sample counties, we generate R samples $\theta_{ijk^*,\zeta}, \zeta = 1, \ldots, R$, from the distribution assumed in the linking model with parameters evaluated at the R posterior samples, $N(x_{ijk^*}\beta_\zeta + s_{ij,\zeta} + d_{i,\zeta}, \sigma^2_{c,\zeta})$. If the not-in-sample county k^* belongs to a state j^* for which no other county component is represented in the in-sample counties, but it belongs to a census division i already represented in the in-sample counties, we generate R samples

$\theta_{ij^*k^*,\zeta}, \zeta = 1, \ldots, R$ in two steps: first, we generate state latent effect samples $s_{ij^*,\zeta}$ from $N(0, \sigma^2_{s,\zeta})$; second, we generate R samples $\theta_{ij^*k^*,\zeta}$ from $N(x_{ij^*k^*}\beta_\zeta + s_{ij^*,\zeta} + d_{i,\zeta}, \sigma^2_{c,\zeta})$. Finally, if the not-in-sample county k^* belongs to a state j^* that belongs to a census division i^* for which no other county component is represented in the in-sample set of counties, we generate R samples $\theta_{i^*j^*k^*,\zeta}, \zeta = 1, \ldots, R$ in three steps: first, we generate census division latent effect samples $d_{i^*,\zeta}$ from $N(0, \sigma^2_{d,\zeta})$; second, we generate state latent effect samples $s_{i^*j^*,\zeta}$ from $N(0, \sigma^2_{s,\zeta})$; third, we generate R samples $\theta_{i^*j^*k^*,\zeta}$ from $N(x_{i^*j^*k^*}\beta_\zeta + s_{i^*j^*,\zeta} + d_{i^*,\zeta}, \sigma^2_{c,\zeta})$. Note that this final scenario is included here for methodology completion, but it was not encountered in the application modeling process.

To construct model predictions on the original scale, we back-transform the posterior samples using a squared-sine transformation, $\theta^f_{ijk,\zeta} := (sin(\theta_{ijk,\zeta}))^2$, for all the in-sample and not-in-sample counties. Then, county-level posterior summaries such as means, variances, credible intervals are constructed using the R samples $\tilde{\theta}^f{}_{ijk,\zeta} \in \left\{\theta_{ijk,\zeta}, \theta^f{}_{ijk,\zeta}\right\}$.

State-level, census division-level, and nation-level model predictions are constructed as aggregates of the county-level model predictions, with ACS county-level population totals t_{ijk} serving as aggregation weights. On the original scale, the state-level model predictions are constructed using the corresponding R samples. The state-level samples are defined as $\theta^f{}_{ij,\zeta} := \sum_{k \in j, j \in i} t_{ijk}\tilde{\theta}^f{}_{ijk,\zeta}(\sum_{k \in j, j \in i} t_{ijk})^{-1}$, the census division-level samples are defined as $\theta^f{}_{i,\zeta} := \sum_{k \in i} t_{ijk}\tilde{\theta}^f{}_{ijk,\zeta}(\sum_{k \in i} t_{ijk})^{-1}$, and the nation-level samples are defined as $\theta^f{}_{\zeta} := \sum_k t_{ijk}\tilde{\theta}^f{}_{ijk,\zeta}(\sum_k t_{ijk})^{-1}$.

The predictive power of the models is tested using cross-validation. For this, the in-sample counties are divided into two groups with similar distribution of sample sizes. Each model is fitted to one of the two groups of counties, in turn, and the counties in the other group are considered not-in-sample. The model predictions for the not-in-sample counties are then compared against the survey estimates, for groups of counties with large sample size.

4. | RESULTS

Variable selection and cross-validation results are presented in Tables 1 and 2, for the models fitted to all the counties with sample data and for the models fitted to the SMART counties, respectively. The auxiliary variables selected in the models fitted to all the counties with sample data are available at the county-level from ACS, Small Area Health Insurance Estimates (SAHIE), or Statistics of Income (SOI). The first six variables listed in Table 2 are available at the county level, but the rest are only available at the state level. Compared to the auxiliary data selected for the models fitted to all the counties with sample data, data from three additional government agencies were selected for the models fitted to the SMART counties: the National Center for Health Statistics (NCHS), the National Highway Traffic Safety Administration (NHTSA), and the Energy Information Administration (EIA). Recall Table 1 in the Appendix with the information on the data sources. The full models have better predictive power than the reduced models, having smaller sum of squared

differences between the model predictions and the survey estimates. The results from the model comparison using WAIC are also in favor of the full models, as shown in Table 3.

For the counties with available survey estimates, we compare the distributions of the survey estimates and the model predictions produced using the four models under investigation. The results are illustrated in Figure 1. There are minor differences between the full and reduced models for each of the two pairs considered. Based on these results and the results from the model comparison and cross-validation, we conclude that the reduced models are no longer of interest.

As illustrated in Figure 1, the point estimates are comparable across the different sources considered in this comparison, with exception being the main mode of the distributions: the model predictions fall in a narrower range than the range of the survey estimates, as a result of smoothing, with the model predictions based on the models fitted to the SMART counties being smoothed towards slightly larger values than the other model predictions. The uncertainty in the model predictions based on the models fitted to the SMART counties is noticeably smaller than both the uncertainty in the survey estimates and the uncertainty in the model predictions based on the models fitted to all the counties with sample data. This result is a consequence of constructing a much larger number of not-in-sample model predictions under the approach where only the SMART counties are used in the model fit, while ignoring the survey data available for the non-SMART counties.

To further evaluate the model fits using different sets of counties, we compare the estimates for large geographies: census divisions and nation. Pairwise comparisons of the survey and model predictions for the high levels of aggregation are presented in Table 4. Note that for most of the census divisions, there is closer agreement between the survey estimates and the model predictions based on the model fitted to all the counties with sample data than the model predictions based on the model fitted to the SMART counties. In addition, for the nation and most of the census divisions, the standard errors of the model predictions based on the model fitted to the SMART counties are larger than the survey standard errors. This result is a consequence of constructing a large number of positively highly correlated not-in-sample model predictions under the model fitted to the SMART counties only, which impacts the variance estimates at aggregated levels. Hence, we identify the model fitted to all the counties with sample data to be the preferred or final model. A map of all the county-level model predictions constructed based on the final model is illustrated in Figure 2.

Finally, we note that the covariates selected do not present a very strong linear relationship to our outcome of interest (the proportion of individuals who do not have a personal care provider or doctor); the R squared for a multiple linear regression fitted to the county-level survey estimates and using the set of covariates selected for the model fitted to all the counties with sample data is approximately 0.08. Hence, it is important to use all the available county-level survey estimates in fitting the model, so that the number of not-in-sample predictions is as small as possible.

4.1 | External validation

Various external validation checks are considered for the final model. For the in-sample counties, we compare the distributions of the survey estimates and the model predictions in terms of point estimates and coefficients of variation (CVs). The results are illustrated in Figure 3 for both the transformed and the original scales. Note that the ranges of the model predictions are narrower than the ranges of the survey estimates as a result of smoothing. The main modes of the overall distributions of point estimates overlap, with no noticeable difference between the distributions for in-sample county-level model predictions and all model predictions (for in-sample counties and not-in-sample counties). On the other hand, the main modes of the overall distributions of CVs are shifted to the left for the model predictions, compared to the survey estimates, as a result of model-based decrease in uncertainty.

Next, we investigate the distribution of the differences between the survey estimates and model predictions, and the ratios of survey to model standard errors, both relative to the county sample sizes. The results are illustrated in Figure 4 for both the transformed and the original scales. A dotted line is included in the plots to indicate the group of SMART counties: points to the right of the line correspond to SMART counties. Note that the survey estimates and model predictions are comparable for larger counties, while for counties with smaller sample sizes, the estimates differ: model predictions deviate from the survey estimates because the latter are not sufficiently precise, and the model standard errors/CVs are smaller than the survey standard errors/CVs (on the transformed scale) as a result of model shrinkage. The shrinkage effects are observed on the scale on which the models are fitted, i.e., the transformed proportions using the arcsine-square-root transformation, and are diminished after the back transformation to the original scale, i.e. the proportions. As a consequence, the model standard errors/CVs are smaller than the survey standard errors/CVs for most, but not all of the counties used in the model fitting process.

Pairwise comparisons of the survey estimates and model predictions for the high levels of aggregation are presented in Table 4 Note that there is close agreement between the point estimates and standard errors, with slight reduction in the standard errors for some areas: census divisions Middle Atlantic, West South Central, and Mountain. The agreement in the point estimates suggests there are no necessary changes in the model specification. The reduction in the standard errors is a result of the model shrinkage, and its small magnitude is expected because at these high levels (nation and census division) the survey estimates are precise.

As a final step in the external validation process, we construct the percentage of counties with non-overlapping 95 percent uncertainty intervals for the survey estimates and model predictions defined as

$$y_{ijk} \pm 2\sigma_{ijk} \text{ (transformed scale) or } \hat{p}_{ijk} \pm 2\sqrt{\hat{V}_{ijk}} \text{ (original scale),}$$

and

$$\left(\widetilde{\theta}_{ijk,(0.025)}, \widetilde{\theta}_{ijk,(0.975)}\right),$$

respectively. For large counties, one would expect to see few or no non-overlapping intervals because the survey estimates in these counties are considered sufficiently precise and the model predictions should not deviate significantly from them. Since the survey estimates for smaller counties are not as reliable as the ones for larger counties, one can expect to see more non-overlapping intervals. For the SMART counties, all the uncertainty intervals overlap. Among the counties with sample sizes greater than or equal to 30, there are less than 2 percent non-overlapping uncertainty intervals. Among all the counties with sample, there are less than 4 percent non-overlapping uncertainty intervals. On the transformed scale, most (31 out of 40) of the non-overlapping uncertainty intervals correspond to counties with survey estimates of 0 or 1 on the original scale. This is expected, because the model predictions for counties with smaller sample sizes are shrunk more toward the common trend, than the large counties are, resulting in significantly different estimates from the survey estimates.

5 | DISCUSSION

In summary, we developed model-based small area estimation methodology for the proportion of individuals with no personal care provider or doctor in the 2018 population. The models combine survey data from the BRFSS with data from other sources, while accounting for the error in the survey data and the nested structure of the data. Model predictions were constructed for all the counties in the nation, irrespective of the availability of survey data. For counties where survey estimates are available, the model predictions are more precise than the survey estimates. For evaluation purposes, the models were developed using either all the counties with survey data or only the counties with sample sizes at least 500. The former was preferred because it resulted in more accurate point estimates than the latter.

As part of the model development, we investigated other models and compared their performance to the performance of the model presented in this manuscript. Among these alternative model specifications, we investigated area-level models for the BRFSS survey estimates on the original scale or transformed using other transformations: square root and logit. Similarly to the transformation adopted for the proposed model, the square root and the logit transformations help relax a normality assumption for the survey estimates on the original scale. However, the logit transformation is only applicable to survey estimates greater than zero and smaller than one. Moreover, for the original scale of these two alternative transformations, the estimated variances associated with the survey estimates would not be defined for all the counties, either due to sample sizes or due to the transformation. Finally, unlike the transformation adopted for the proposed model, none of these three alternative model specifications would stabilize the sampling variances, so a smoothing step would be necessary, too.

Lacking unit-level auxiliary data for the entire population, we did not investigate any unit-level models. However, we did consider a hybrid-level model, where the sampling level is

specified at the unit level and the linking level is specified at the area (county) level. As an initial step, we ignored the survey design effect and the imputation. That is, we specified the sampling level for the BRFSS unit-level raw data. The computational challenges led us to quickly set aside a nation-wide model, and instead fit a model to all the data available for one state. As part of the internal validation process, we noticed residuals with heavy-tailed distribution, so the model specification indicated there was room for improvement. Also, the state-level (aggregated) model prediction was significantly smaller than the state-level survey estimate, again indicating the model specification can be improved.

Future investigations may include closer attention to the hybrid-modeling and extensions to the model presented in this manuscript: (1) multivariate specifications, for example for modeling the BRFSS proportion of individuals who do not have a personal care provider or doctor and the proportion of individuals without health insurance jointly; and (2) measurement error specifications, to account for the uncertainty in the county-level covariates.

ACKNOWLEDGMENTS

This work was conducted under a CDC-Westat project. The authors thank Carol Pierannunzi, the CDC's main contact for the project, for helpful discussions and comments.

AUTHOR BIOGRAPHY

Andreea L. Erciulescu. Dr. Andreea Erciulescu is a Senior Statistician at Westat, working on the interface between survey statistics and other areas of statistics, including combining multiple survey and non-survey sources to answer complex analytic questions. Her areas of expertise include Bayesian statistics, hierarchical models, measurement error, official statistics, resampling methods, small area estimation, statistical data integration, and survey statistics. Dr. Erciulescu is an Elected Member of the International Statistical Institute.

Jianzhu Li. Dr. Li is a Senior Statistician in Westat's Statistics and Evaluation Sciences Unit and has 20 years of experience. She leads and works on survey sampling, small area estimation and data confidentiality and disclosure protection for several large scale household and establishment surveys for government agencies.

Tom Krenzke. Mr. Tom Krenzke is a Vice President and Associate Director in Westat's Statistics and Evaluation Sciences Unit and has 30 years of experience. Mr. Krenzke leads research in survey sampling, statistical confidentiality, small area estimation and other statistical topics. Mr. Krenzke is a Fellow of the American Statistical Association (ASA) and a Westat Senior Statistical Fellow.

Machell Town is the Branch Chief for the Population Health and Surveillance Branch of the Division of Population Health at the CDC. She has 30 years of experience in population estimation and statistics.

APPENDIX

TABLE 1

Auxiliary data pool

Source	Auxiliary variable
Census Bureau	5-year estimates (2013–2017) of socioeconomic
American Community Survey (ACS)	demographic, and housing characteristics
Census Bureau	small area estimates of selected
Small Area Income and Poverty Estimates (SAIPE)	income and poverty statistics
Census Bureau	health insurance coverage status
Small Area Health Insurance Estimates (SAHIE)	by selected economic and demographic characteristics
U.S. Department of Agriculture (USDA) Economic Research Service	classification of counties into metro and non-metro (OMB subdivided counties into three metro and
	six non-metro categories)
Bureau of Labor Statistics (BLS)	monthly and annual employment, unemployment,
The Local Area Unemployment Statistics (LAUS) program	and labor force data
Bureau of Economic Analysis (BEA)	estimates of personal income for local areas
Centers for Disease Control and Prevention	updated statistics about diabetes
Division of Diabetes Translation (DDT)	
Centers for Medicare & Medicaid Services (CMS)	utilization and quality of health care services for the Medicare fee-for-service population
Geographic variation public use file	income and tax data such as number of tax returns,
The Internal Revenue Service The Statistics of Income (SOI) Data	returns with unemployment compensation, returns with taxable Social Security benefit,
	adjusted gross personal income,
	personal unemployment compensation amount, and
	personal taxable Social Security benefit amount, etc.
Health Resources and Services Administration	data on health care professions, health facilities,
Area Health Resources Files (AHRF)	population characteristics, economics,
	health professions training, hospital utilization,
	hospital expenditures, and environment
National Center for Health Statistics (NCHS)	birth rate and infant mortality rate
Federal Bureau of Investigation (FBI)	crime rates
National Highway Traffic Safety Administration (NHTSA)	traffic fatalities
U.S. Energy Information Administration (EIA)	energy consumption height

Abbreviations:

ACS	American Community Survey
BRFSS	Behavioral Risk Factor Surveillance System
CDC	Centers for Disease Control and Prevention
EIA	Energy Information Administration

MCMC	Markov chain Monte Carlo
NCHS	National Center for Health Statistics
NHTSA	National Highway Traffic Safety Administration
SAE	small area estimation
SAHIE	Small Area Health Insurance Estimates
SMART	Selected Metropolitan/Micropolitan Area Risk Trends
SOI	Statistics of Income
WAIC	widely applicable information criterion

References

1. Battese G, Harter R, Fuller W. An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association 1988; 83: 28–36.
2. Fay R, Herriot R. Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association 1979; 74(366a): 269–277.
3. Cadwell B, Thompson T, Boyle J, Baker L. Bayesian small area estimation of diabetes prevalence by U.S. county, 2005. Journal of Data Science 2010; 8: 173–188.
4. Zhang Z, Holt J, Lu H, et al. Multilevel regression and poststratification for small-area estimation of population health outcomes: A case study of chronic obstructive pulmonary disease prevalence using the Behavioral Risk Factor Surveillance System. American Journal of Epidemiology 2014; 179(8): 1025–1033. [PubMed: 24598867]
5. Pierannunzi C, Xu F, Wallace R, et al. A methodological approach to small area estimation for the Behavioral Risk Factor Surveillance System. Preventing Chronic Disease 2016; 13(E91): 150480.
6. Berkowitz Z, Zhang X, Richards T, Nadel M, Peipins L, Holt J. Multilevel small-area estimation of colorectal cancer screening in the United States. Cancer Epidemiology, Biomarkers and Prevention 2018; 27(3): 245–253.
7. Berkowitz Z, Zhang X, Richards T, et al. Multilevel regression for small-area estimation of mammography use in the United States, 2014. Cancer Epidemiology, Biomarkers and Prevention 2019; 28(1): 32–40.
8. Holt J, Matthews K, Lu H, et al. Small Area Estimates of Populations With Chronic Conditions for Community Preparedness for Public Health Emergencies. American Journal of Public Health 2019; 109(S4): S325–S331. [PubMed: 31505141]
9. Raghunathan T, Xie D, Schenker N, et al. Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. Journal of the American Statistical Association 2007; 102: 474–486.
10. Liu B, Parsons V, Feuer E, et al. Small area estimation of cancer risk factors and screening behaviors in U.S. counties by combining two large national health surveys. Preventing Chronic Disease 2019; 16(E119): 190013.
11. Kish L Survey Sampling. New York: John Wiley & Sons, Inc. 1965.
12. Fuller W, Goyeneche J. Estimation of the state variance component. Unpublished manuscript 1998.
13. Torabi M, Rao J. On small area estimation under a sub-area level model. Journal of Multivariate Analysis 2014; 127: 36–55.
14. Erciulescu A, Cruze N, Nandram B. Statistical challenges in combining survey and auxiliary data to produce official statistics. Journal of Official Statistics 2020; 36(1): 63–88.
15. Krenzke T, Mohadjer L, Li J, et al. Program for the International Assessment of Adult Competencies (PIAAC): State and County Estimation Methodology Report. Tech. Rep.

NCES2020225, U.S. Department of Education; Rockville, MD: Westat: 2020. Available at https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2020225

16. Browne W, Draper D. A comparison of Bayesian and likelihood based methods for fitting multilevel models. Bayesian Analysis 2006; 1(3): 473–514.
17. Gelman A Prior distributions for variance parameters in hierarchical models (Comment on an article by Browne and Draper). Bayesian Analysis 2006; 1(3): 515–534.
18. Polson N, Scott J. On the Half-Cauchy Prior for a Global Scale Parameter. Bayesian Analysis 2012; 7(4): 887–902.
19. Erciulescu A, Opsomer J. A model-based approach to predict employee compensation components. In: Joint Statistical Meetings Proceedings. Government Statistics Section. American Statistical Association.; July 27 – August 1, 2019; Alexandria, VA: 1601–1623.

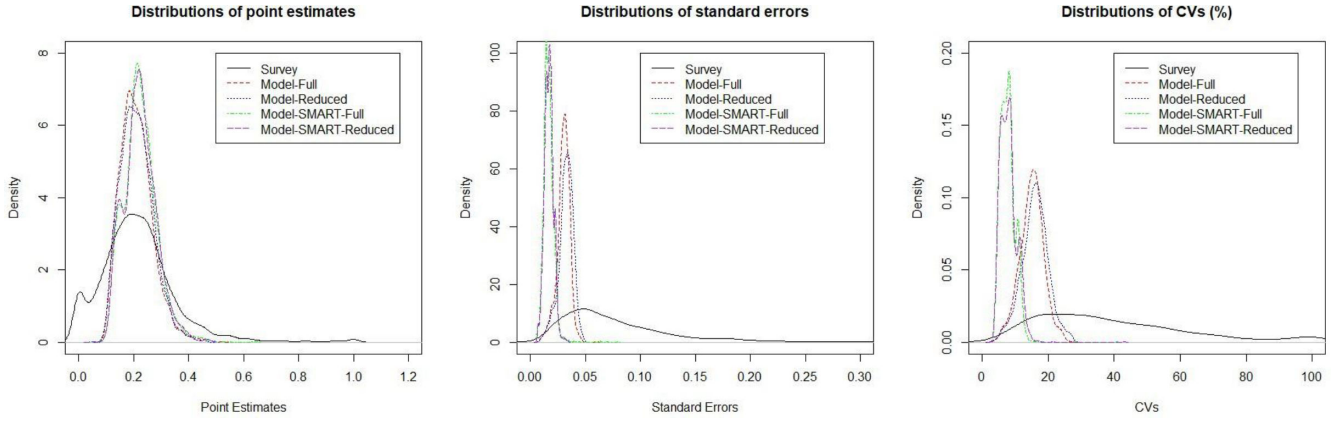

FIGURE 1.

Comparison of county-level point estimates, standard errors, and CVs; models fitted to all the counties with sample data, models fitted to the SMART counties, and survey

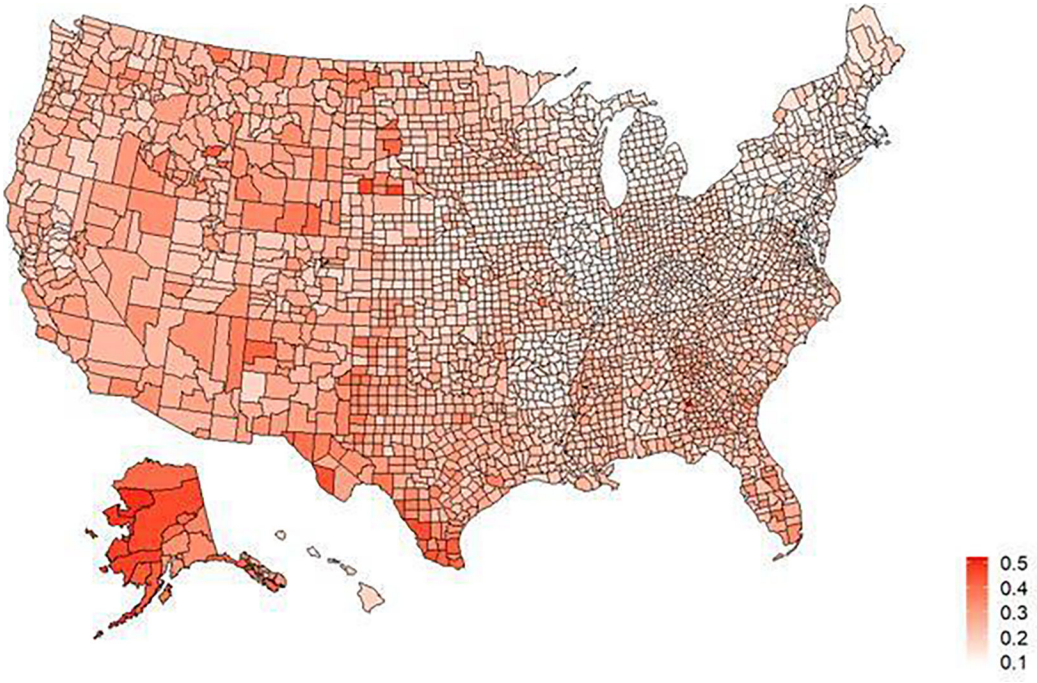

FIGURE 2.

Map of county-level model predictions: proportions of individuals with no personal care provider or doctor

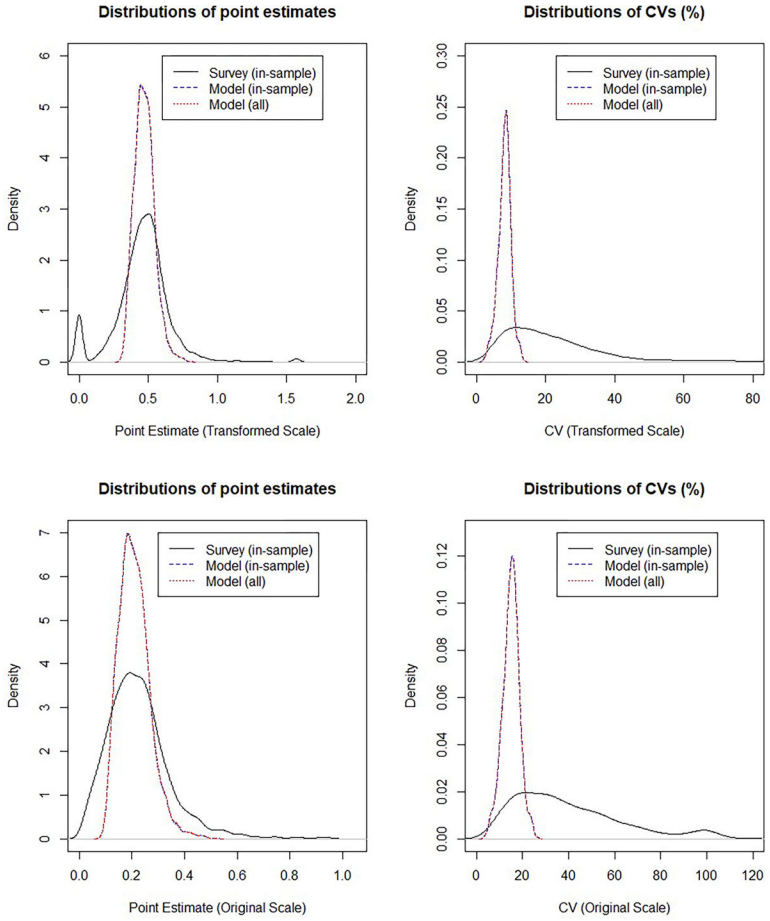

FIGURE 3.

Comparison of county-level model predictions and survey estimates: point estimates and CVs; transformed and original scales

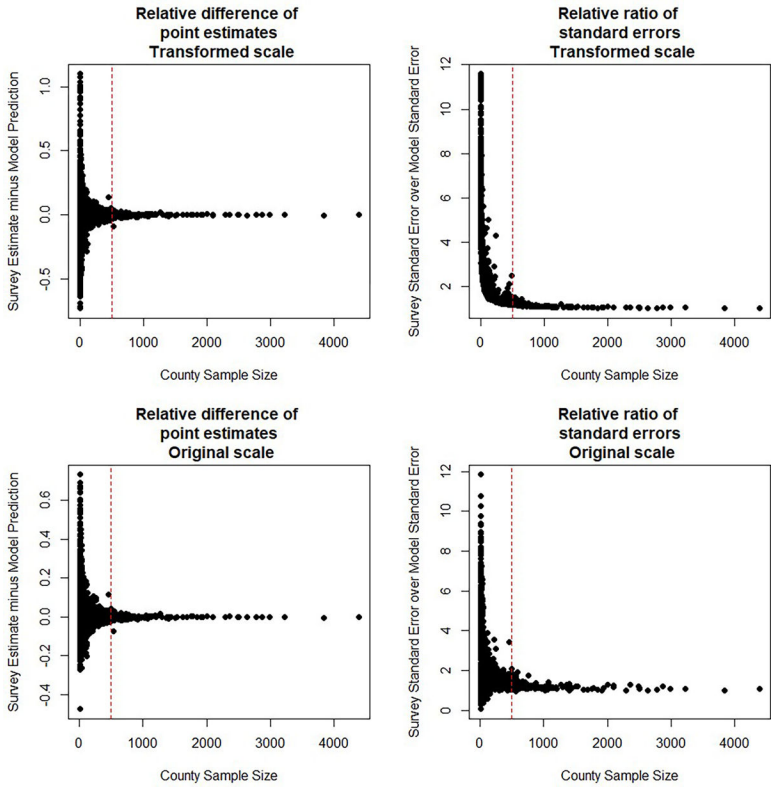

FIGURE 4.

Comparison of county-level model predictions and survey estimates: difference in point estimates and ratio of standard errors, relative to the sample size; transformed and original scales

TABLE 1

Variable selection and cross-validation results for the models fitted to all the counties with sample data

Variable	Source	Full model	Reduced model
Proportion of population age 25+: with high school diploma, no college	ACS	X	X
Proportion of Hispanics	ACS	X	X
Proportion of Non-Hispanic Whites	ACS	X	X
Proportion of Native Americans	ACS	X	
Proportion of owner occupied housing unit	ACS	X	X
Proportion of population 18+: in different house in the past year	ACS	X	
Proportion of population 18+: in different state in the past year	ACS	X	
Proportion of uninsured population	SAHIE	X	X
Proportion of returns with taxable Social Security benefits per person	SOI	X	
Sum of squared differences between model predictions and survey estimates			
213 counties with sample size ≥ 500		0.350	0.416
963 counties with sample size ≥ 100		4.092	4.244

The variables included in the model are indicated by X.

TABLE 2

Variable selection and cross-validation results for the models fitted to the SMART counties

Variable	Source	Full model	Reduced model
Proportion of Hispanics	ACS	X	X
Proportion of Non-Hispanic Whites	ACS	X	X
Proportion of population 55–64 years old	ACS	X	X
Proportion of population 18+: in different state in the past year	ACS	X	
Proportion of uninsured population	SAHIE	X	X
Proportion of returns with taxable Social Security benefits per person	SOI	X	
Birth rate (live birth as a proportion of total population)	NCHS	X	X
Per capita energy consumption	EIA	X	
Traffic fatalities per 100 million vehicle miles	NHTSA	X	X
Sum of squared differences between model predictions and survey estimates			
56 counties with sample size ≥ 1, 000		0.094	0.109
213 counties with sample size ≥ 500		0.409	0.483

The variables included in the model are indicated by X.

TABLE 3

Model comparison results

Model	WAIC
All counties with sample data - full	−0.7198
All counties with sample data - reduced	−0.7177
SMART counties - full	−2.0346
SMART counties - reduced	−2.0117

TABLE 4

Comparison of nation-level and census division-level estimates, standard errors, and CVs; models fitted to all the counties with sample data (Model Nation), models fitted to the SMART counties (Model SMART), and survey

Nation/Census Division	Model Nation		Model SMART		Survey	
	Point estimate	Standard error	Point estimate	Standard error	Point estimate	Standard error
Nation	0.228	0.001	0.229	0.003	0.230	0.001
New England	0.144	0.003	0.140	0.003	0.143	0.003
Middle Atlantic	0.182	0.003	0.181	0.006	0.185	0.004
West North Central	0.224	0.003	0.222	0.005	0.228	0.003
South Atlantic	0.240	0.003	0.240	0.008	0.242	0.003
East South Central	0.221	0.004	0.241	0.014	0.229	0.004
West South Central	0.289	0.006	0.306	0.008	0.295	0.007
Mountain	0.280	0.003	0.277	0.004	0.280	0.004
Pacific	0.252	0.004	0.241	0.005	0.252	0.004