A Search for Robust Model-based Small Area Estimation Methods for the National Ambulatory Medical Care Survey (NAMCS)

Vladislav Beresovsky*

*National Center for Health Statistics, 3311 Toledo Rd, Hyattsville, MD 20782

Abstract

NAMCS has been recently redesigned to allow for reliable direct estimates for larger states and remaining smaller states grouped within Census divisions. When covariates correlated with the outcome variable are available for every unit in the population it is possible to gain efficiency of estimates in small areas by using model-based or model-assisted estimators. On the other hand these methods are sensitive in varying degree to possible misspecification of model assumptions with respect to the superpopulation model. In this simulation study we assess robustness of model-based estimators to possible misspecifications of the estimating models and compare their efficiency to direct and model-assisted estimators.

Key Words: direct estimator, model-assisted estimator, model-based estimator, hierarchical logistic-normal model, small area estimation, health care utilization.

1. Introduction

The National Center for Health Statistics (NCHS) conducts the National Ambulatory Medical Care Survey (NAMCS) - a national survey of visits to office-based physicians and selected community health centers. It is a component of the National Health Care Survey which measures health care utilization across a variety of health care providers. Prior to 2012, the NAMCS utilized a multistage design optimal for producing national estimates, that involved probability samples of geographic primary sampling units (PSUs), physicians within PSUs, and patient visits within physician practices. The 2012 NAMCS was redesigned to provide data for estimates at the state level for as many states as possible and at the Census division level for the remaining states. These estimates are expected to provide an opportunity for new understanding and analysis of the health delivery system at the small area level. In this paper we consider geographical areas within Census region (larger states or remaining smaller states, aggregated to Census division) to be small areas.

Although this abundance of new information could be sufficient for many small area analyses, it is possible that some analyses will not be satisfied, including estimation of proportions for health outcomes with very low prevalence for individual states and for testing hypotheses comparing estimates between states. When design-based estimates are inefficient, model-based and model-assisted methods utilizing population covariates to "borrow strength" across small areas are expected to improve the reliability of small area estimates if model assumptions are correct, see Rao (2003). However, misspecifications of models used in practice tend to increase errors of small area estimates. Depending on the model assumptions, some methods can be more robust to such misspecifications than others. The simulation study presented in this paper is an attempt to compare errors of design-based and model-based estimates under various model misspecifications from the expected 2012 NAMCS sample data.

"The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention."

In Section 2 we will provide additional details about the NAMCS sample design prior to 2012 and redesigned sample for 2012. We will also mention modeling of a few variables necessary for realistic simulation of the NAMCS finite population, sample selection process and calculation of survey weights. Estimation of model parameters for the proportion of *visits by patients with private insurance* under various degrees of misspecification of fixed effects will be described in Section 3. Superpopulation models used to simulate NAMCS 2012 finite populations will be presented in Section 4. In addition to misspecified fixed effects, these models will incorporate distributional assumptions usually ignored in estimation models, such as deviation from normality of random effects and variability of $\beta-$coefficients between small areas. In Section 5 we will present model-based and model-assisted estimators of physician-level proportions and a way of aggregating them to the small area level. Results of simulations will be presented in Section 6. Relative root mean squared errors (RRMSE) for considered estimators will be compared for various kinds of misspecifications in the estimation models. The effect on estimates due to sample size and ratios of variability in the studied proportions between and within small areas will be considered. Finally, we will summarize simulation results and draw conclusions about efficiency of different estimators and their robustness to considered misspecifications of model assumptions. We will also mention other possible sources of misspecifications which may increase the RRMSE of the model-based estimators.

2. Simulation of NAMCS population and samples

The NAMCS sampling frame consists of physicians classified as office based in Master files maintained by the American Medical Association (AMA) and the American Osteopathic Association (AOA). Prior to 2012, NAMCS utilized a multistage probability design that sampled 112 PSUs out of about 1,900 geographically defined PSUs (usually counties, groups of counties or their equivalents) that covered the 50 States and the District of Columbia and were stratified by four geographic regions and Metropolitan Statistical Area (MSA) status. The total 2009 sample included 3,319 physicians, but after removing ineligible (”out-of-scope”) physicians, non-respondents and physicians who saw no patients during their reporting week, the final sample available for analysis included 1,293 physicians who completed patient record forms (PRFs), see NAMCS (2009).

The 2009 NAMCS sample designed for national estimates was poorly suited for state-level estimates because of sample clustering within geographic PSUs. The 2012 NAMCS sample was redesigned to produce state-level estimates. The sample size includes approximately 16,000 physicians. Geographic PSUs are no longer part of the sampling process. Instead, a list sample of physicians was selected from strata defined by the largest 34 states and groups of the remaining smaller states in each of the Census divisions (division remainders). To insure proportional selection of physicians of all specialties within geographically defined strata, physicians were also stratified by primary/non-primary care status and then sorted within strata by their specialties. Systematic samples of 383 physicians were drawn from each targeted state and of 431 physicians from each division remainder.

To estimate efficiency of the design-based estimates in small geographical areas and compare them to model-based estimates, we simulated the NAMCS physician population and sample selection process according to the new design. For that purpose the following five outcomes were modeled from the 2009 data and simulated for each physician i in small area j in the sampling frame: p_{ij}- proportions of visits by patients with *private insurance* (with the national average 57.3%); V_{ij}- total annual number of visits; v_{ij}- number of sampled visits per physician; $I_{ij} = 0/1$ and $R_{ij} = 0/1$ - respectively ”in-scope” and response indicators for a sampled physician. Using these variables the sampled number

of visits by patients with private insurance was simulated as the binomial count $y_{ij} = Bin(p_{ij}, v_{ij})$.

Comparing different methods of modeling proportions p_{ij} of visits to physicians and using them for estimating proportions in small areas is the main focus of this paper. Response R_{ij} and "in-scope" I_{ij} indicators were modeled as binary random variables with a logistic link. A linear model was used for the squared root of the annual number of visits V_{ij} to physician offices. The outcome of this model was used to model sampled numbers of visits v_{ij}. All models utilized county-level covariates from the Area Resource File (ARF), see ARF (2009), and physician-level covariates from the combined AMA/AOA master files. Proportions of visits by patients with private insurance p_{ij} and the number of such visits y_{ij} to physician offices were drawn from the inferred distributions for each simulation of the final population. Other variables were simulated just once and the same fixed values were used for all simulated populations. This allows to eliminate the influence of possible correlations between dependent variable y_{ij} with other simulated variables on the considered estimators of proportions in small areas and focus our analysis only on the variability of the outcome variable y_{ij}.

For each simulated final population, systematic samples were drawn within each stratum defined according to the 2012 NAMCS design described earlier in this Section. Physicians with "in-scope" and "response" indicators equal to 0 were excluded from the sample. Physician sampling weights w_{ij} are identical for all physicians i within a stratum S_j^c defined by small area j and primary/non-primary care status c and were calculated using the simulated "in-scope" and response indicators as follows:

$$w_{ij} = w_j^c = \frac{N_j^c}{n_j^c} \frac{\sum\limits_{i \in S_j^c} I_{ij}}{\sum\limits_{i \in S_j^c} I_{ij} R_{ij}} = \frac{N_j^c}{n_j^c} \frac{n_j^{c,I}}{n_j^{c,IR}} \tag{1}$$

Here N_j^c is the total number of primary or non-primary care physicians in the area's population, n_j^c is the total sample size, $n_j^{c,I}$ is the total number of sampled "in-scope" physicians and $n_j^{c,IR}$ is total number of "in-scope" and respondent physicians with completed PRFs.

3. Models with misspecified covariates

Realistic simulation of the NAMCS population requires utilizing veritable parameters of the superpopulation model. These parameters can be inferred from modeling outcome variables using data collected in previous years (in this study, 2009 NAMCS data were used). The proportion p_{ij} of visits to physician i in small area j by patients using private insurance as a source of payment was modeled by a hierarchical logistic-normal model with random effects at the small area and physician levels:

$$\begin{gathered} \text{logit}\left(p_{ij}\right) = \mathbf{X}_{cnt}\beta_{cnt} + \mathbf{X}_{phys}\beta_{phys} + \theta_{ij} + \theta_j; \\ \theta_{ij} \sim N\left(0, \sigma_p^2\right); \theta_j \sim N\left(0, \sigma_s^2\right) \end{gathered} \tag{2}$$

Two sources of covariates were utilized by the model: county-level covariates $\mathbf{X}_{cnt}$ available from the ARF and physician-specific covariates $\mathbf{X}_{phys}$ available from the AMA database. The covariates $\mathbf{X}_{phys}$ did not change much between small areas and were not particularly helpful for explaining variability of studied proportion between small areas. These covariates were included in the model (2) because they were useful for better explaining variability between individual physicians, reducing the estimated variance of physician-

level random effects σ_p^2 and, consequently, for more realistic simulations of the studied proportion for each physician. Model covariates on county level $\mathbf{X}_{cnt}$ were more instrumental for explaining variability between geographical small areas. When all significant covariates $\mathbf{X}_{cnt}$ were included in the model, the variance of state-level random effects disappeared, that is, $\sigma_s^2 = 0$. On the other hand, if the $\mathbf{X}_{cnt}$ were omitted from the model, there remained substantial unexplained variability between small areas, that is, $\sigma_s^2 = 0.11$.

These observations suggested the following idea for studying misspecification of fixed effects. First, always keep all of the physician level covariates $\mathbf{X}_{phys}$ to insure that the model accounts for sample design information. Second, simulate the effect of misspecification of model covariates by excluding some of the significant county level covariates $\mathbf{X}_{cnt}$ from the design matrix. In the following, $\mathbf{X}_{cnt}^k (k = 1, 2, 3, 4, 5)$ will designate various sets of county-level covariates from the ARF, ranging from complete set $(k = 1)$ to the case with no such covariates $(k = 5)$. Unexplained variability between small areas manifested itself by increasing the variance of random effect $\sigma_s^2 > 0$. At the same time, misspecification of county-level covariates did not cause significant changes in variability between physicians σ_p^2. Table 1, below, presents estimated variances $\sigma_{s,k}^2$ and $\sigma_{p,k}^2$ of random effects in the model expressed in equation (2) for $(k = 1, 2, 3, 4, 5)$.

Table 1: Relation between misspecification of model covariates and estimated variances of random effects. $(k = 1, 2, 3, 4, 5)$ designates various sets of the ARF covariates.

	Less fixed effects ⇒ more random effects				
	$k = 1$	2	3	4	5
	$\sigma_{s,k}^2 = 0$	0.047	0.077	0.09	0.11
ARF covariates $\mathbf{X}_{cnt}$	$\sigma_{p,k}^2 = 1.73$	1.71	1.75	1.81	1.87
HOUSEDENSITY	•	•	•	•	
BEDSNUM	•	•	•	•	
AGE65PCT	•	•	•	•	
SPECIALITS	•	•	•	•	
AGE19PCT	•	•	•	•	
WHITEHOUSENUM	•	•	•	•	
BLACKHOUSNUM	•	•	•		
UNEMPLOYTYP	•	•	•		
LOWEDUC	•	•	•		
OFFICEWRK	•	•	•		
POVERTY	•	•			
SSIBEN	•	•			
RECRTYP	•	•			
NOINSURANCE	•				

Table 1 suggests that different terms of the model expressed in equation (2) explain the variability ("useful signal") of proportions between small areas. This signal was generated by the ARF covariates $\mathbf{X}_{cnt}$ and random effects θ_j and must be extracted from the "background noise" generated by variability of θ_{ij} between individual physicians. The complete set of covariates $(k = 1)$ was sufficient for extracting this signal. If some of the ARF covariates were not available $(k = 2, 3, 4, 5)$, estimation of the random effect θ_j in small areas becomes necessary. Background noise was almost independent of any covariates and complicated efficient estimation of proportions in small areas. With this picture in mind in the rest of the paper, we will consider models with incomplete sets of the ARF covariates representing situations when fixed effects were misspecified to various degrees, which usually happens in practice.

4. Easy-to-misspecify details of superpopulation distributions

In general, models used for model-based estimation can be misspecified in many ways. In this paper we consider misspecification of fixed effects and of distribution of random effects at the small area level. These kinds of misspecifications happen most frequently in practice because: (1) it is never possible to know absolutely a perfect set of covariates and (2) a normal distribution of random effects is usually assumed by the majority of software tools available for fitting mixed models, disregarding other options.

In Section 3, we defined design matrices $\mathbf{X}^k_{cnt}$ of models characterized by various degrees of misspecification for fixed effects and estimated parameters of these models (β^k_{cnt}, β^k_{phys}, $\sigma^2_{p,k}$, $\sigma^2_{s,k}$). These design matrices and parameters were used to simulate NAMCS populations with controlled misspecification of fixed effects. In addition to the normal distribution of small area level random effects, we also assumed the most common deviations from normality, such as skewed and heavy-tailed distributions. A skewed distribution was simulated by a chi-squared distribution with 4 degrees of freedom $\chi^2_{(4)}$ and a heavy-tailed distribution was simulated by two distributions: $t_{(4)}$-distribution with 4 degrees of freedom and a mixture of two normal distributions, one with variance $\sigma^2_{s,k,m1} = \sigma^2_{s,k}$ and another with variance $\sigma^2_{s,k,m2} = 8\sigma^2_{s,k}$ and $p = 0.2$ as the probability of realization. These distributions were normalized to have the same first two moments (0 mean and variance $\sigma^2_{s,k}$) as the original normal distribution of random effects:

$$\theta_j \sim \sqrt{\frac{\sigma^2_{s,k}}{(2\times 4)}}\left(\chi^2_{(4)} - 4\right), \tag{3a}$$

$$\theta_j \sim \sqrt{\sigma^2_{s,k}\frac{(4-2)}{4}}\, t_{(4)}, \tag{3b}$$

$$\theta_j \sim \left(\frac{\sigma^2_{s,k}}{\left((1-p)\,\sigma^2_{s,k,m1} + \sigma^2_{s,k,m2}p\right)}\right)^{1/2} \begin{cases} N\left(0, \sigma^2_{s,k,m1}\right), & Bin\,(1,p) = 0 \\ N\left(0, \sigma^2_{s,k,m2}\right), & Bin\,(1,p) = 1 \end{cases} \tag{3c}$$

The random variability of model coefficients between small areas can be another possible source of misspecification of an estimating model. Robustness of model-based estimators which do not explicitly account for such variability was evaluated by generating a NAMCS superpopulation for model coefficients $\beta^k_{cnt,j} \sim N\left(\beta^k_{cnt}, 0.08\right)$, which vary between small areas with the mean value β^k_{cnt} estimated from the sample data.

The described features of the superpopulation model were likely to affect the efficiency of model-based, but not design-based, estimators. However, both design-based and model-based estimators were expected to be dependent on the noise-like variability between physicians and sample size. Since the author's interest was to evaluate the relative efficiency of model-based methods in comparison with design-based methods, simulations were conducted for nominal, above and below nominal levels of variability between physicians, that is, $\sigma^2_{noise,k} = \sigma^2_{p,k}, 3\sigma^2_{p,k}, 0.1\sigma^2_{p,k}$, and two sample sizes: the first was equal to the projected 2012 NAMCS sample size and the second was $\frac{1}{4}$ of the 2012 sample size. The resulting

superpopulation models used in simulations can be formulated in general form as:

$$\begin{gathered} \text{logit}\left(p_{ij}^{k}\right) = \mathbf{X}_{cnt}^{k}\beta_{cnt,j}^{k} + \mathbf{X}_{phys}\beta_{phys}^{k} + \theta_{ij} + \theta_{j} \\ \beta_{cnt,j}^{k} = const = \beta_{cnt}^{k} \text{ or } \beta_{cnt,j}^{k} \sim N\left(\beta_{cnt}^{k}, 0.08\right) \\ E\left(\theta_{j}\right) = 0;\ Var\left(\theta_{j}\right) = \sigma_{s,k}^{2} \\ \theta_{ij} \sim N\left(0, \sigma_{noise}^{2}\right) \end{gathered} \tag{4}$$

where θ_j has either normal or one of the distributions in (3).

For each simulated population a stratified systematic sample was drawn according to the 2012 NAMCS design and sampling weights w_{ij} for the physicians with PRFs were calculated, as described in Section 2.

5. Design- and model-based estimators of proportions in small areas

The proportion of visits to physician offices by patients with private insurance was estimated by standard design-based, model-based and model-assisted methods for all finite populations generated from the superpopulation models described above. Design-based methods did not use any distributional assumptions and estimated the proportions directly from simulated data for small areas. Model-based methods utilized simplified and misspecified versions of the superpopulation model to produce estimates that were more efficient than the design-based small area estimates. Each method had its advantages and weaknesses, depending on the degree and kind of misspecification of the model used for estimation. The goal of this study was to identify the most efficient and robust method for estimation of small area proportions from the simulated 2012 NAMCS sample by comparing the RRMSEs of estimates from different model-based methods with the RRMSE of the design-based estimates and with each other.

Design-based (D) estimates of the proportions in small area j were calculated as weighted average of physicians i in that area as following :

$$P_{j}^{D} = \frac{\sum\limits_{i \in S_j} w_{ij} p_{ij}^{D} V_{ij}}{\sum\limits_{i \in S_j} w_{ij} V_{ij}} \tag{5}$$

where S_j designates sampled, "in-scope" responding physicians with completed PRFs in area j; w_{ij} is physician weight defined in equation (1); $p_{ij}^{D} = y_{ij}/v_{ij}$ is the simulated proportion of visits by patients with private insurance and V_{ij} is the simulated visit volume (see Section 2 for the definition of variables).

Model-based methods for estimating proportions in small areas used available population covariates and could vary significantly by their efficiency and robustness to model misspecification. First, the simple *logistic regression model (M1)* was used to estimate the proportion for every physician in the finite population. The *M1* model was formulated as:

$$\text{logit}\left(p_{ij}^{k,M1}\right) = \mathbf{X}_{cnt}^{k}\beta_{cnt}^{k} + \mathbf{X}_{phys}\beta_{phys}^{k} \tag{6}$$

This model did not include state-level random effects θ_j found in the superpopulation model in equation (4) and thus might be strongly misspecified if their contribution in small areas was substantial. Without explicitly accounting for the data in small areas, predictions from model *M1* critically depended on the ability of covariates $\mathbf{X}_{cnt}^{k}$ to explain variability between small areas.

The *logistic regression model (M2)* expanded model *M1* by including different inter-

cepts for each small area *j*. This model was formulated as:

$$\text{logit}\left(p_{ij}^{k,M2}\right) = \alpha_j + \mathbf{X}_{cnt}^{k}\beta_{cnt}^{k} + \mathbf{X}_{phys}\beta_{phys}^{k} \tag{7}$$

Directly accounting for the data in small areas greatly improved robustness to misspecification of model assumptions but might have had a negative impact on efficiency, particularly when there was not enough data from small areas and fixed effects happened to be good predictors.

The *logistic-normal model (M3)* used random effects to account for variability not explained by fixed effects in small areas. Model *M3* was formulated as:

$$\text{logit}\left(p_{ij}^{k,M3}\right) = \theta_j + \mathbf{X}_{cnt}^{k}\beta_{cnt}^{k} + \mathbf{X}_{phys}\beta_{phys}^{k}, \theta_j \sim N\left(0, \sigma_k^2\right) \tag{8}$$

Using random, instead of fixed, intercepts in this model is expected to improve efficiency when limited amount of data is available in small areas.

All of the above estimating models *M1-M3* were misspecified relative to the certain distributional assumptions of the superpopulation model. They ignored possible variability of the β_{cnt}^{k}- coefficients between small areas and deviation of random effects θ_j from normality. Simulation results presented in Section 6 demonstrate the importance of these misspecifications for the considered model-based estimators.

Model-based estimates of proportion P_j^{M1-M3} in small area j can be calculated by averaging proportions p_{ij}^{M1-M3} predicted for all physicians i in the finite population of that area U_j, as following:

$$P_j^{M1-M3} = \frac{\sum\limits_{i \in U_j} I_{ij} p_{ij}^{M1-M3} V_{ij}}{\sum\limits_{i \in U_j} I_{ij} V_{ij}} \tag{9}$$

where I_{ij} are 0/1 "in-scope" indicators and V_{ij} are annual visits volumes. In practical situation these characteristics of office-based physicians are not available for the entire NAMCS sampling frame and therefore cannot be used in the model-based estimators of proportion in small areas, similar to expression (9). In this study they were simulated (see Section 2 for details) in order to investigate how different misspecifications of the estimating model affect the RRMSE of model-based and model-assisted estimators. *Model-assisted (MA)* estimators from survey data were described in detail in Sarndal and Lundstrom (2006). We considered a variation of the widely used regression estimator adjusted for estimating proportions instead of totals. This is a composite estimator utilizing both simulated proportions for sampled physicians p_{ij}^{D} and the model-predicted proportions p_{ij}^{M1} for all physicians in the finite population which may be formulated as follows:

$$P_j^{MA} = \frac{\sum\limits_{i \in S_j} w_{ij}\left(p_{ij}^{D} - p_{ij}^{M1}\right) V_{ij}}{\sum\limits_{i \in S_j} w_{ij} V_{ij}} + \frac{\sum\limits_{i \in U_j} I_{ij} p_{ij}^{M1} V_{ij}}{\sum\limits_{i \in U_j} I_{ij} V_{ij}} \tag{10}$$

Note, that the *MA* estimator is a combination of weighted averages in small area j over sampled, "in-scope" and respondent physicians S_j and over "in-scope" physicians in the finite population U_j.

6. Simulation results

In this study we conducted $R = 40$ simulations for $J = 41$ small areas, representing 36 larger states and 5 groups of smaller states within census divisions. The fielded 2012 NAMCS covered only 34 states because of insufficient funding to field samples for all 36 states. For every simulated finite population ($r = 1, 2, .., R$), a sample was selected and the "true" finite population proportion P_{rj}^{FP} of visits with private insurance in small areas j was calculated and consequently estimated by P_{rj}^{X} using estimators $X \in \{D, M1 - M3, MA\}$. The RRMSE of estimated proportions averaged over all small areas can be considered a reasonable measure of the average efficiency of estimator X, that is:

$$RRMSE^X = \frac{1}{J}\sum_{j=1}^{J}\frac{\sqrt{\frac{1}{R}\sum\limits_{r=1}^{R}\left(P_{jr}^{X} - P_{jr}^{FP}\right)^2}}{\frac{1}{R}\sum\limits_{r=1}^{R}P_{jr}^{X}} \tag{11}$$

When the β- coefficients in the superpopulation model (4) do not vary between small areas and random effects θ_j have normal distribution, estimating models may still use a misspecified design matrix of covariates $\mathbf{X}_{cnt}^{k}$. Five levels of such misspecification are presented in Table 1, starting with omitting only one covariate ($k = 1$) and gradually excluding all county-level covariates from the model ($k = 5$). The misspecification of fixed covariates increased the variance of random effects from $\sigma_{s,k}^2 = 0$ to 0.11.

The RRMSE of estimators for the proportion of visits by patients with private insurance in small areas for different values of between-physician variability σ_{noise}^2 and sample sizes are presented in Table 2. The absolute value of RRMSE is provided for design-based estimators and relative percent ($\frac{RRMSE^{M1-M3,MA}}{RRMSE^{D}} * 100\%$) is used to present RRMSE of model-based estimators.

The RRMSEs of the design-based estimator were mostly independent from the variance $\sigma_{s,k}^2$ of random effects in small areas but increased with between-physician variability σ_{noise}^2 and decrease in the sample size. The RRMSE of model-based estimators depended on both of these random terms of the superpopulation model expressed in equation (4). The logistic regression estimator *M1* critically depended on random effects θ_j because they were ignored in the estimating model expressed in equation (6). This dependence was greater for low level of "noise" $\sigma_{noise}^2 = 0.1\sigma_p^2$ between physicians and was less for larger σ_{noise}^2 and smaller samples. At the same time, when there was no contribution from the random effects ($\sigma_{s,k}^2 = 0$), estimator *M1* was the most efficient. Other model-based estimators *M2*, *M3* and *MA* demonstrated greater robustness to misspecification of model covariates. Even when covariates were completely misspecified ($k = 5$), these models performed better than design-based estimator. The advantage of robust model-based methods over the design-based estimator was particularly significant for smaller values of physician-level variability, $\sigma_{noise}^2 = 0.1\sigma_p^2$, and less dependent on sample size. The logistic-normal model *M3* always better utilized the explanatory power of county-level covariates than other robust models (*M2,MA*).

The effect on small area estimators of proportion due to ignoring the variability of β- coefficients between small areas in the superpopulation model is illustrated in Table 3 for logistic and logistic-normal models. The case when county-level covariates were absent from the model ($k = 5$) (see Table 1) is not presented. When fixed effects were specified correctly ($k = 1$), the efficiency of estimator *M1* was sharply diminished by misspecified model coefficients β. This effect was less pronounced for misspecified fixed effects ($k = 2, 3, 4$). The estimator based on the logistic-normal model *M3* was robust to

Table 2: RRMSE of design and model-based estimators of proportions of visits by patients with *private insurance* in small areas for different levels of misspecification of model covariates, variability between physicians and sample sizes. Absolute values of RRMSE are presented for design-based estimator and relative percent for model-based estimators. Data were simulated from the superpopulation models (4) with parameters inferred from the 2009 NAMCS sample.

	Less fixed effects ⇒ more random effects				
	$k=1$	2	3	4	5
Estimator	$\sigma^2_{s,k}=0$	0.047	0.077	0.09	0.11
	$\sigma^2_{noise}=\sigma^2_p$				
Direct *D*	0.0396	0.0383	0.0396	0.0401	0.0395
Logistic regression *M1*	21%	141%	175%	183%	202%
Logistic regression *M2*	80%	82%	82%	85%	85%
Logistic-normal *M3*	69%	79%	80%	84%	84%
Regression estimator *MA*	94%	93%	94%	96%	98%
	$\sigma^2_{noise}=0.1\sigma^2_p$				
Direct *D*	0.0235	0.0241	0.0231	0.0204	0.0199
Logistic regression *M1*	14%	246%	333%	405%	461%
Logistic regression *M2*	62%	62%	67%	75%	78%
Logistic-normal *M3*	31%	60%	66%	75%	77%
Regression estimator *MA*	73%	75%	79%	89%	92%
	$\sigma^2_{noise}=3\sigma^2_p$				
Direct *D*	0.0538	0.0554	0.0542	0.0544	0.0545
Logistic regression *M1*	22%	82%	110%	113%	124%
Logistic regression *M2*	85%	82%	87%	86%	87%
Logistic-normal *M3*	78%	82%	84%	84%	86%
Regression estimator *MA*	96%	97%	98%	98%	99%
	Sample size $1/4$ of 2012 NAMCS sample				
Direct *D*	0.0805	0.0811	0.0782	0.0793	0.0809
Logistic regression *M1*	15%	66%	90%	93%	99%
Logistic regression *M2*	81%	82%	85%	84%	87%
Logistic-normal *M3*	70%	75%	80%	79%	82%
Regression estimator *MA*	94%	93%	96%	96%	99%

possible variability of the β coefficients between small areas in all cases.

The RRMSE of the estimators based on the models *M2* and *MA* are not presented in Table 3. Results presented in Table 2 suggest that these models are more dependent on the data in small areas and less dependent on the fixed effects than model *M3*. Therefore, they are expected to be even less sensitive to the ignored variability of β- coefficients.

Among estimators considered in this study, only the estimator based on the logistic-normal model *M3* explicitly assumes normality of the random effects θ_j at the small area level. Table 4 compares the RRMSE of this estimator when this assumption is correct with RRMSEs in cases when distribution of θ_j in superpopulation models (3, 4) was either heavy-tailed, skewed, or a mixture of two normal distributions. All considered distributions of θ_j had mean 0 and the same value of variance $\sigma^2_{s,k}$, $k \in 2, 3, 4, 5$. Results are presented for nominal and reduced variance σ^2_{noise} of random effects between physicians. In all cases misspecification of the distribution of random effects had no noticeable effect on the RRMSEs of the estimated proportion in small areas.

Table 3: RRMSE of estimators of proportions of visits by patients with *private insurance* in small areas by logistic *M1* and logistic-normal *M3* models for fixed and random β-coefficients in superpopulation model and different levels of misspecification of model covariates. Absolute values of RRMSE are presented for fixed β and relative percent for random β. Data were simulated from the superpopulation models (4) with parameters inferred from the 2009 NAMCS sample.

	Less fixed effects ⇒ more random effects			
β-coefficients of the	$k=1$	2	3	4
superpopulation model	$\sigma^2_{s,k}=0$	0.047	0.077	0.09
Logistic model *M1*				
Fixed β^k_{cnt}	0.0085	0.0539	0.0692	0.0734
Random $\beta^k_{cnt,j} \sim N\left(\beta^k_{cnt}, 0.08\right)$	296%	118%	111%	101%
Logistic-normal model *M3*				
Fixed β^k_{cnt}	0.0272	0.0303	0.0318	0.0335
Random $\beta^k_{cnt,j} \sim N\left(\beta^k_{cnt}, 0.08\right)$	105%	105%	100%	98%

Table 4: RRMSE of estimators of proportions of visits by patients with *private insurance* in small areas by logistic-normal model *M3* depending on deviation from normality of random effects θ_j in the superpopulation model for different levels of σ^2_{noise} and misspecifications of model covariates. Absolute values of RRMSE are presented for normal and relative percent for non-normal distributions of θ_j. Data were simulated from the superpopulation models (4) with parameters inferred from the 2009 NAMCS sample.

Distribution	Less fixed effects ⇒ more random effects			
of random effects in the	$k=2$	3	4	5
superpopulation model	$\sigma^2_{s,k}=0.047$	0.077	0.09	0.11
	$\sigma^2_{noise}=\sigma^2_p$			
Normal	0.0303	0.0318	0.0335	0.0330
Heavy tails $t_{(4)}$	102%	102%	101%	97%
Skewed $\chi^2_{(4)}$	100%	103%	102%	93%
Mixture of 2 normals	105%	99%	104%	96%
	$\sigma^2_{noise}=0.1\sigma^2_p$			
Normal	0.0145	0.0153	0.0153	0.0154
Heavy tails $t_{(4)}$	105%	101%	103%	98%
Skewed $\chi^2_{(4)}$	106%	99%	100%	99%
Mixture of 2 normals	97%	99%	98%	99%

7. Conclusions

In this paper we proposed methodology for a realistic simulation of a finite population and replication of sample selection processes according to specifications for the redesigned 2012 NAMCS. This methodology provides an opportunity to assess and compare the efficiency of direct and model-based methods for estimating proportions of categorical outcomes in small areas. The primary focus of this study was to investigate how various misspecifications in the estimation models, relative to the superpopulation model, affected efficiency of the estimates and to identify the most efficient and robust method of estimation.

By analyzing the currently available 2009 NAMCS data, we estimated the variances of random effects at the small area and physician levels for a number of models with design

matrices of county-level covariates of various ranks. The variance of random effects at the small area level was zero for certain large design matrices and increased when rank was reduced. The variance of random effects between physicians did not noticeably change in the process. Models with reduced rank design matrices and positive variances of random effects at the small area level were considered to represent the models with misspecified fixed effects. Other possible misspecifications in estimating models ignored variability of $\beta-$coefficients between small areas and deviation from normality of state-level random effects.

The RRMSE of the design-based estimator was mostly independent of the design matrix of county-level covariates and random effects at the small area level in the superpopulation model but strongly dependent on noise-like variability between physicians and sample size. The estimator utilizing logistic regression with a common intercept for all small areas substantially outperformed all other estimators when fixed effects were correctly specified but quickly became the least efficient for reduced rank design matrices or ignored variability of $\beta-$ coefficients between small areas. Other model-based estimators utilized logistic regression with separate intercepts for each small area, or logistic-normal hierarchical model with random effects at the small area level, or the regression-type model-assisted estimator. All of them were equally robust to misspecification of county-level covariates and ignored variability of $\beta-$ coefficients, but a logistic-normal model outperformed two other estimators for all considered superpopulation models. Although this estimator assumed normality of random effects at the small area level, its RRMSE did not increase for heavy-tailed, skewed, or mixture distribution of random effects in the superpopulation model as long as the first two moments of these distributions were the same as for normal distribution.

Conducted simulations demonstrated some advantage of robust model-based and model-assisted estimators over a design-based estimator for all considered misspecifications of the superpopulation model. But there are still some issues with these estimators which require further analysis. First, the robustness of the normality assumption for random effects was demonstrated for fairly large number of sampled physicians within each small area, which is expected to be the case for the 2012 NAMCS data. It is unclear if and when this robustness property breaks down for the reduced amount of data in small areas. Second, bias of estimates based on model-based methods associated with possible informative sampling may increase the RRMSE. Further work is required in order to investigate the effect of these factors on the efficiency of model-based methods for estimating proportions in small areas.

References

ARF (2009). *2009 Area Resource File, National County-level Health Resource Information Database*. Health Resources and Services Administration. `http://arf.hrsa.gov/overview.htm`.

NAMCS (2009). *2009 NAMCS Micro-Data File Documentation*. National Center for Health Statistics. `ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NAMCS/doc09.pdf`.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology, Hoboken, NJ.

Sarndal, C-E. and Lundstrom, S. (2006). *Estimation in Surveys with Nonresponse*. Wiley Series in Survey Methodology, Chichester, England.