

Effect of model misspecification on small area estimation of proportions from the National Ambulatory Medical Care Survey data

Vladislav Beresovsky and Donald Malec

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

Emails: vberesovsky@cdc.gov and dmalec@cdc.gov

Abstract

To be able to produce reliable state level estimates, the National Ambulatory Medical Care Survey (NAMCS) modified its design for 2012 and the sample size was substantially increased. Because of limited sample size within small areas, direct estimates may still be inefficient. On the other hand it is known that model-based estimates can be very sensitive to model misspecification. In this simulation study we investigate the effect of misspecification of model covariates on the estimates of proportions in small areas by using logistic and hierarchical logistic-normal models. While estimates by a logistic regression model can be better or worse than direct estimates depending on the degree of model misspecification, estimates by logistic-normal model are robust to misspecification of fixed effects and always, in our investigation, more efficient than direct estimates. Demonstrated robustness of model-based methods in the context of the new NAMCS design may stimulate their practical application for small-area estimation.

Key Words: health care utilization, small area estimation, hierarchical logistic-normal models.

Introduction

NAMCS is a national survey of visits to office-based physicians and selected community health centers conducted by the National Center for Health Statistics (NCHS). It is a component of the National Health Care Surveys which measure health care utilization across a variety of health care providers.

Prior to 2012, NAMCS utilized a multistage design that involved probability samples of geographic primary sampling units (PSUs), physicians within PSUs, and patient visits within practices. Its design has been modified and, beginning in 2012, will include a substantial increase in sample size. With the new design and increase in sample size, precise estimates of visit and physician characteristics are anticipated for all of the largest 36 states and 5 groups of remaining smaller states within census divisions called division remainders. The availability of design-based estimates for selected states and divisions is expected to provide the opportunity for new understanding and analysis of the health delivery system at the small area level. Although this abundance of new information will be sufficient for many analyses, it is anticipated that not all small area analysis will be satisfied, including estimation of proportions for individual states, particularly for health outcomes with very low prevalence, and for testing hypotheses comparing estimates between states. In preparation for this need, a preliminary assessment of the possibility of providing model-based small area estimation is undertaken here.

The availability of small area estimation techniques that “borrow strength” has increased over the years [1]. However, errors caused by model failure are still less understood and, due to lack of small area data, can never be adequately estimated at the small area level. Design-based estimation has similar problems when both estimates and the accompanying estimates of precision are inaccurate but the impact of these errors seems to be better understood. In the following work, proposed small area models for NAMCS are assessed using simulations from realistic models with the aim towards understanding some of the limits of small area estimates for patient visit characteristics. In particular, our goal is to investigate the effect of misspecification of model covariates on small area estimation. By revising and improving proposed model-based small area estimations, based on observed results, the aim is to develop a useful methodology of small area estimation and to provide documentation to assist researchers in understanding the strengths and limitations of the resulting estimates.

First, both the 2009 and 2012 NAMCS designs are summarized. Knowledge of both designs is needed in order to obtain reasonable population models. Since potential outcomes from the 2012 design and sample

need to be predicted, samples from the repeatedly simulated finite populations were drawn using the 2012 design. Next, steps taken to model the population using the 2009 sample, auxiliary information available on the sampling frame and county level socio-economic status covariates from the Area Resource File (ARF), were documented for particular case of modeling the proportion of visits with specified source of payment. Note that for this evaluation, only visits pertaining to office-based physician visits are modeled and evaluated. Visits to community health centers (CHC) are not evaluated, here. The procedure for population simulation, the various small area methods utilized and the means of evaluation were described. Lastly, results from the evaluation are presented for proportion of visits with particular source of payment.

The NAMCS designs and population modeling

The 2009 NAMCS design

Prior to 2012, NAMCS utilized a multistage probability design that involved probability samples of primary sampling units (PSUs), physicians within PSUs, and patient visits within practices. The sampling frame consists of a list of current physicians maintained by the American Medical Association (AMA) and the American Osteopath Association (AOA), including their specialty and mailing address. The first-stage sample included 112 geographical PSUs typically consisting of a county, a group of counties, county equivalents (such as parishes and independent cities), towns, townships, minor civil divisions (for some PSUs in New England), or a metropolitan statistical area (MSA). Because of clustering within geographic PSUs, this design was optimal for national estimates but poorly suited for state-level estimates [2].

Modeling the 2009 NAMCS data:

In order to realistically model population outcomes that were needed in estimation, five outcomes were modeled: proportions of visits with particular source of payment, whether a physician in the frame was in scope, physician non-response, total annual number of patient visits and number of sampled visits per physician. Although the number of sampled visits per physician was controlled by the design, it was decided to model this variability. The number of sampled visits to each responding physician having one of the studied sources of payment was modeled as a binomial count. Logistic mixed model for the corresponding binomial mean depended on the following covariates: 2009 and 2012 NAMCS design variables, physician level variables from the AMA file which was also used as sampling frame, and county level variables from the ARF. State- and physician- level random effects were also included to ascertain the degree of variability needed for simulations. For sampled physicians, in-scope status, non-response, total number of visits and sampled number of visits were modeled using fixed effect models with the same covariates. Non-response and in-scope status were both modeled as a binary random variable with a logistic link. The squared root of annual number of visits was modeled using a linear model. The sample visit count was modeled by the fixed effect linear model using the same available covariates plus the annual number of visits. The strata identifiers were not explicitly included in the models; instead AMA covariates used to construct the original strata were used for modeling.

The 2012 NAMCS design

Starting with 2012, NAMCS has been modified to produce accurate state-level estimates. The sample size has been increased roughly six-fold from the 2009 samples size of 3,000 physicians. In addition, the design was changed to allow for estimation for the largest 36 states and 5 groups of remaining smaller states within the census divisions. The new design stratification was drawn either along state lines or remaining states within census divisions and physician specialty category. Physicians were directly list-sampled within these strata to have an equal number of physicians sampled from each state and a slightly larger number for the 5 groups of smaller states within census divisions.

Undoubtedly, the redesigned NAMCS ensures improvement in the randomization-based state-level estimates as compared to the previous NAMCS design. Conducted simulations estimated efficiency of the direct estimates and allowed for comparison of these estimates with model-based estimates incorporating population covariates from the ARF and the AMA file.

Details of simulation and estimation

Simulation

Most of the information collected by NAMCS is represented by categorical variables. Estimation is generally focused on estimating proportions of visits with various characteristics to physician offices. Model-based methods utilized auxiliary information in the form of population covariates to achieve higher efficiency of estimates. Usually available covariates explained only part of the variability of studied proportions between small areas. This simulation study demonstrated the effect of misspecification of model covariates on mean squared error (MSE) and bias of predicted proportions in small areas. The results were compared with the randomization-based estimates which used only sample data and design information for estimation. These estimates were known to be unbiased but not highly efficient for estimation in domains with few data points.

Proportions of visits to physician offices by patients using different sources of payment were modeled using NAMCS 2009 data. Sources of payment included private insurance (with the national average 57.3%), Medicaid (national average 16.75%) and self-pay patients (national average 6.72%). Proportions of visits p_{ij} to physician i in state j were modeled using a hierarchical logistic regression with random effects at physician and state levels:

$$\text{logit}(p_{ij}) = \mathbf{X}\boldsymbol{\beta} + \theta_{ij} + \theta_j; \theta_{ij} \sim N(0, \sigma_p^2); \theta_j \sim N(0, \sigma_s^2)$$

where \mathbf{X} is a matrix of physician- and county- level covariates and a normal distribution was assumed for physician- and state- level random effects. It is reasonable to treat random effects as representing variability of the outcome variable on different levels of aggregation which cannot be explained by model covariates. Since studied proportions of visits to each individual physician were subject to uncontrollable variations, we found that substantial variability at the physician level existed for any set of model covariates. The situation is different for state-level random effects describing more systematic variability at a higher level of aggregation which can be explained by the proper selection of covariates. First, we included in the model a wider set of covariates and obtained extremely low variance of random effects at the state level. Then the least significant covariates with large p -values were gradually excluded from modeling. At a certain point a smaller set of covariates became insufficient for explaining the state-level variability and, correspondingly, the estimated variance of state-level random effects became significantly greater than 0. Acquired phenomenological experience was applied to simulate the final population and also suggested a scenario to illustrate the effect of model misspecification. The finite population was simulated from the hierarchical logistic-normal model having a minimal set of covariates \mathbf{X}^0 sufficient to explain between-state variability. Consequently, it did not include state-level random effects but included normally distributed physician-level random effects. All model parameters were estimated from the 2009 NAMCS data.

$$\text{logit}(p_{ij}^S) = \mathbf{X}^0 \hat{\boldsymbol{\beta}}_{2009} + \theta_{ij}; \theta_{ij} \sim N(0, \hat{\sigma}_{s,2009}^2);$$

In addition to studied proportions, for each physician i in small area j we simulated the following values: annual volume of visits N_{ij}^S , sampled number of visits n_{ij}^S and two indicator 0/1 variables which become available in practice after a physician has been sampled - indicator S_{ij} to be “in-scope” of the sampling frame and indicator R_{ij} for each sampled physician to respond to field representative by filling out patient record forms (PRF). In the following treatment all these simulated values were considered to be known and uncorrelated with the studied proportions. For each sampled physician, the number of visits by patients identifying a specific source of payment was simulated as binomial random variable $y_{ij}^S = B(p_{ij}^S, n_{ij}^S)$.

The simulated “true” proportions in small area j was calculated by summation over the population of “in-scope” physicians as:

$$P_j^S = \frac{\sum_{i \in U} S_{ij}^S p_{ij}^S N_{ij}^S}{\sum_{i \in U} S_{ij}^S N_{ij}^S}$$

A stratified simple random sample was drawn from the simulated population of physicians in accordance with the accepted rules for 2012 NAMCS. Some of the sampled physicians were excluded from the sample according to their modeled in-scope and response indicators. Sampling weights w_{ij} for the remaining physicians were modified accordingly. Using available information from the sampled physicians it was possible to calculate design-based estimates of proportions in small areas (summation goes over sampled physicians):

$$P_j^D = \frac{\sum_{i \in S} (y_{ij}^s / n_{ij}^s) w_{ij}^s N_{ij}^S}{\sum_{i \in S} w_{ij}^s N_{ij}^S}$$

Model-based estimation

For estimation we used logistic regression models with and without random effects for different sets of covariates. Usually, in the process of selecting a model for estimation it was possible to make the following errors: either failure to include important covariates or interactions, or including those which were not significant, or the combination of the above. We demonstrated consequences for making these errors and the role of accounting for random effects on small area estimation by considering the following estimation scenarios.

- 1) “Correct model”. This model had exactly the same set of covariates \mathbf{X}^0 as the superpopulation model. This and all other models included physician specialty which was used in the sample design to identify strata with unequal probability of selection. That must account for possible informativeness of sample design.
- 2) “Incorrect model”. This model differs from the “correct model” by eliminating the group of the least significant covariates, which were part of the superpopulation model. Its design matrix \mathbf{X}^{0-} had fewer columns than the design matrix of “correct model”. This model was supposed to illustrate the effect of omitting significant fixed effects on model-based estimates in small areas and the role of random effects in correcting it.
- 3) “Incorrect model+”. The design matrix of this model \mathbf{X}^{0-+} was missing some relevant fixed effects but included other covariates which were not in the superpopulation model. This model performs better than the “Incorrect model” if these extra covariates are correlated with missing relevant fixed effects, or worse, if they are not correlated, and just add non-essential noise to the estimated parameters.
- 4) “Correct model+”. The design matrix of this model \mathbf{X}^{0+} included all covariates of the “Correct model” plus covariates which were not part of the superpopulation model and only generated additional noise. The purpose of considering this model was to examine how critical this noise was for the efficiency of estimates.

The following estimation models cover the scenarios described above:

$$\text{logit}(p_{ij}^M) = \mathbf{X}^M \hat{\boldsymbol{\beta}} - \text{fixed effects logistic model}$$

$$\text{logit}(p_{ij}^M) = \mathbf{X}^M \hat{\boldsymbol{\beta}} + \theta_j - \text{random effects logistic-normal model}$$

$$\text{where } \theta_j \sim N(0, \sigma_s^2); \mathbf{X}^M \in (\mathbf{X}^0, \mathbf{X}^{0-}, \mathbf{X}^{0-+}, \mathbf{X}^{0+})$$

If a model predicts for each physician the proportion of visits with given characteristics p_{ij}^M then the small area estimates of the proportion will be:

$$P_j^M = \frac{\sum_{i \in U} S_{ij}^S p_{ij}^M N_{ij}^S}{\sum_{i \in U} S_{ij}^S N_{ij}^S}$$

Since small area proportions of the simulated final population P_i^S were known, it was possible to estimate mean-squared error (MSE) and the bias of direct and various model-based methods as the average over repeated simulations of finite populations, each time drawing a sample and using it for direct estimation and building estimation models from which outcome variables can be predicted for the unsampled part of the population. Relative root MSE and relative bias were used to compare different methods of estimation for small area proportions:

$$RRMSE_j^X = \frac{\sqrt{\frac{1}{R} \sum_{r=1}^R (P_{jr}^X - P_{jr}^S)^2}}{\frac{1}{R} \sum_{r=1}^R P_{jr}^X}; \quad RBias_j^X = \frac{\sum_{r=1}^R (P_{jr}^X - P_{jr}^S)}{\sum_{r=1}^R P_{jr}^X}$$

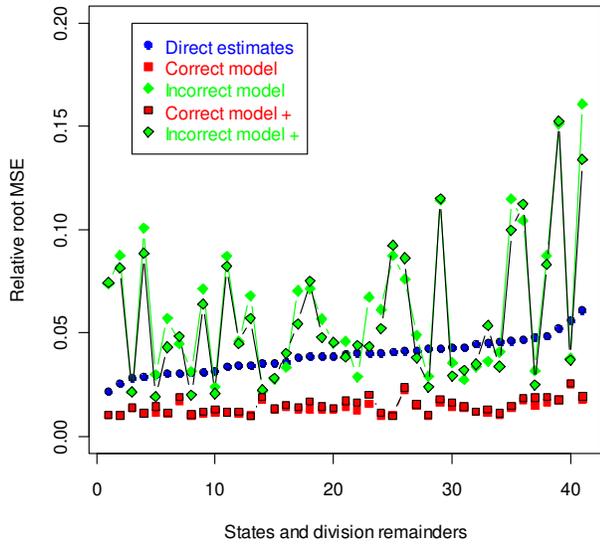
Here $X \in (D, M)$ designated direct or any model-based method of estimation with summation over the R simulated finite populations.

Simulation results

Proportions of visits by patients who specified private insurance, Medicaid or self-payment as a possible source of payment were simulated for every physician in the population from the superpopulation model and then estimated in 36 states and 5 division remainders using design-based and model-based methods described above. Results for the relative root MSE (RRMSE) and relative bias are presented below and sorted by RRMSE of direct estimates.

Figure 1.1. Relative root MSE of estimates of proportions of visits by patients with private insurance in small areas using (a) logistic and (b) logistic-normal models.

(a)



(b)

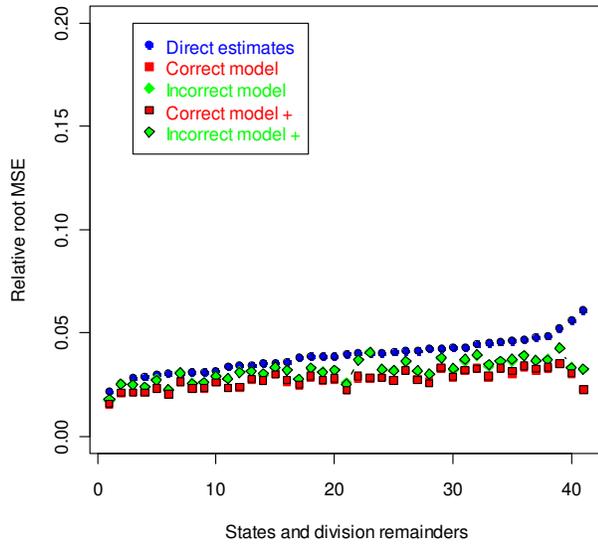
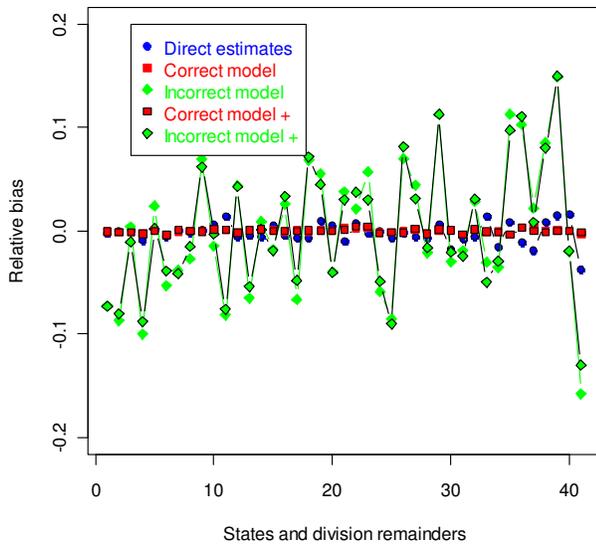


Figure 1.2. Relative bias of estimates of proportions of visits by patients with private insurance in small areas using (a) logistic and (b) logistic-normal models.

(a)



(b)

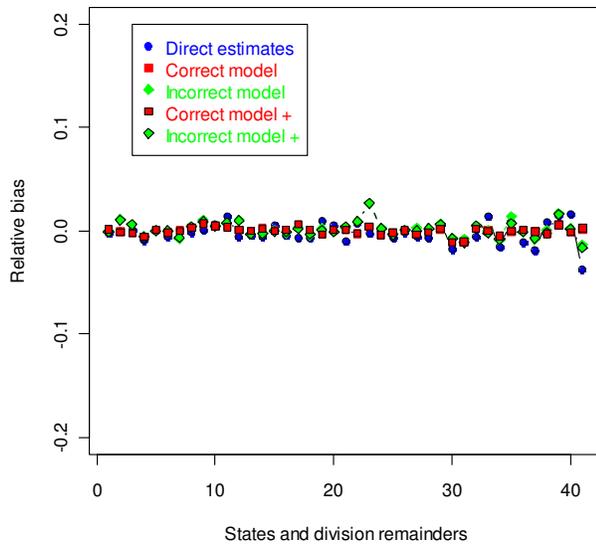


Figure 2.1 Relative root MSE of estimates of proportions of visits by patients with Medicaid in small areas using (a) logistic and (b) logistic-normal models.

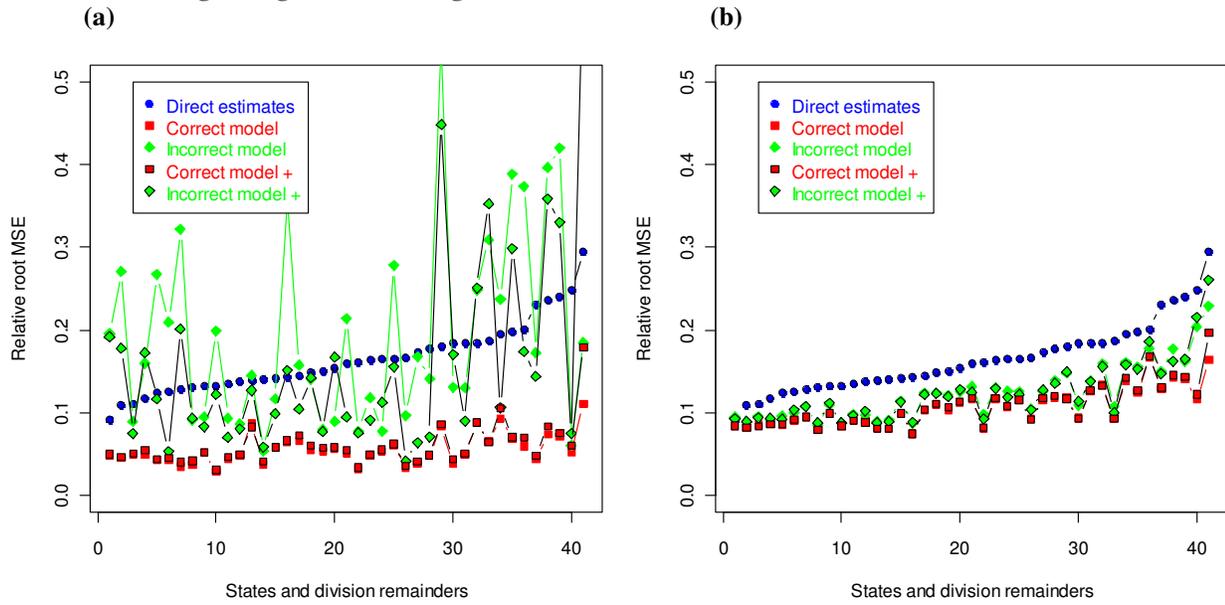


Figure 2.2 Relative bias of estimates of proportions of visits by patients with Medicaid in small areas using (a) logistic and (b) logistic-normal models.

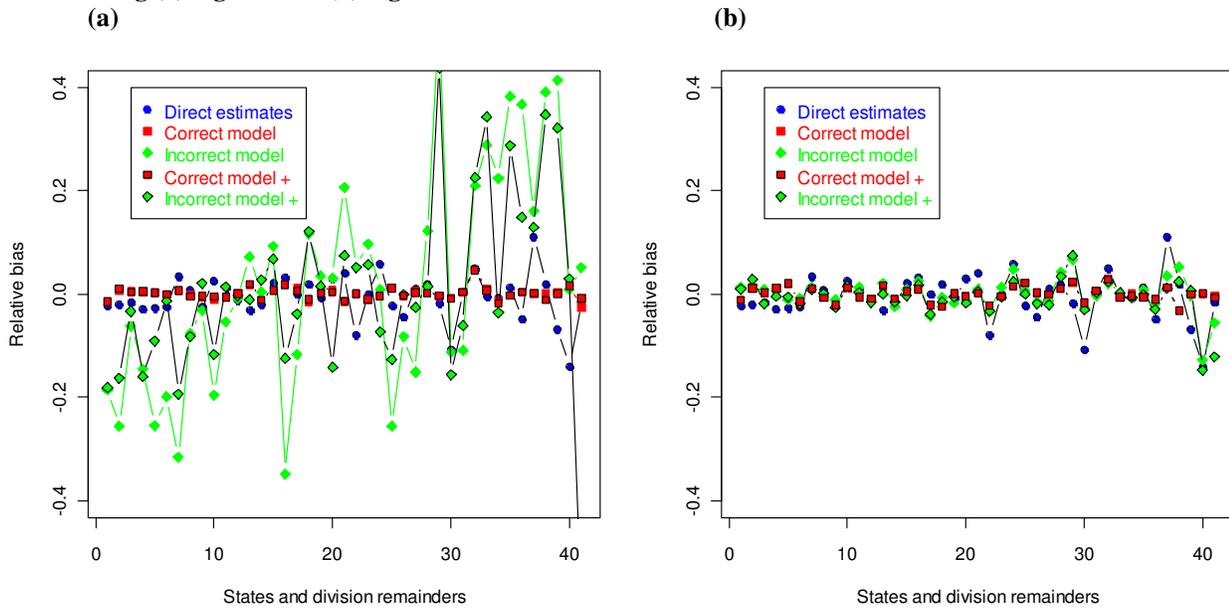


Figure 3.1 Relative root MSE of estimates of proportions of visits by self-pay patients in small areas using (a) logistic and (b) logistic-normal models.

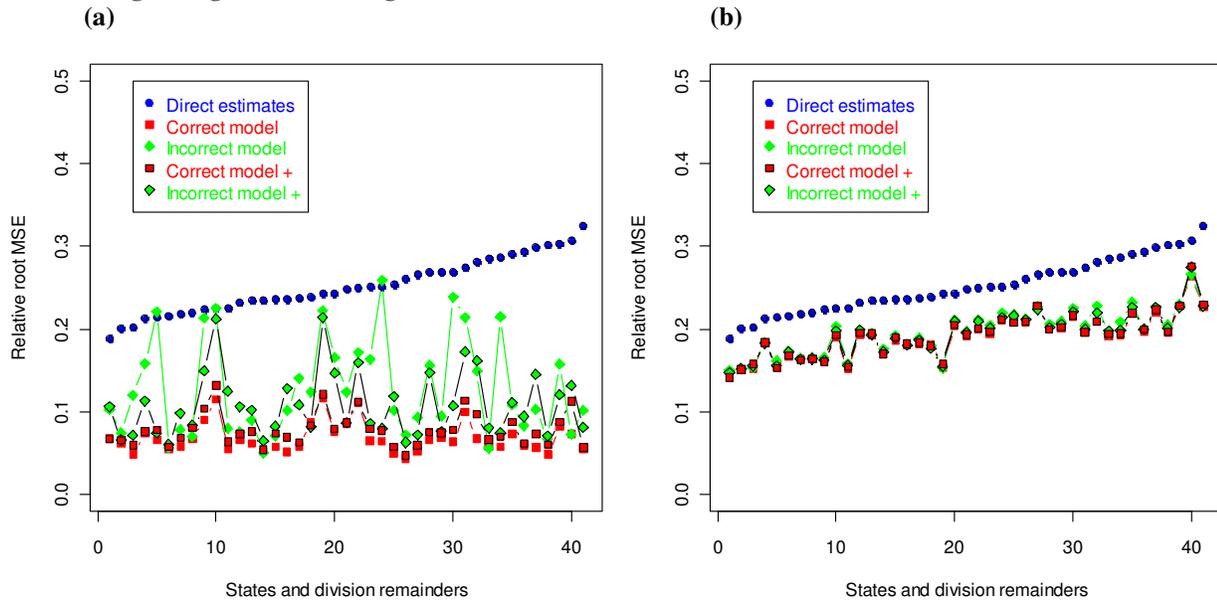
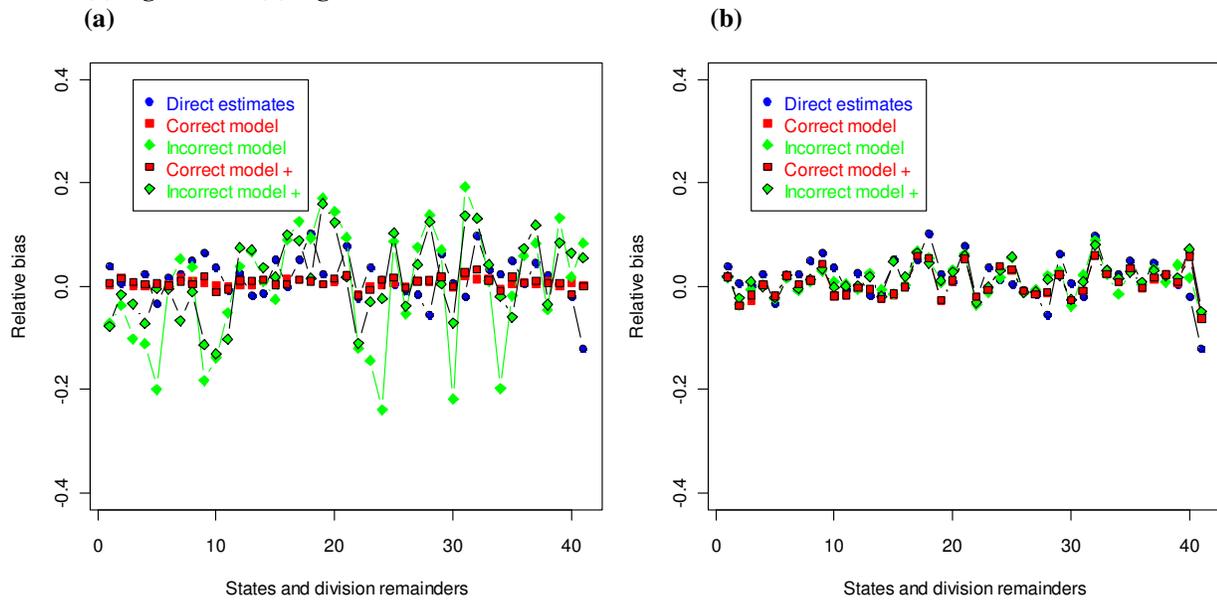


Figure 3.2 Relative bias of estimates of proportions of visits by self-pay patients in small areas using (a) logistic and (b) logistic-normal models.



Comparison of different methods of estimation in small areas were summarized by considering the relative root MSE (RRMSE) and absolute relative bias (ARB) averaged over small areas for different methods of estimation.

Table 1. RRMSE and ARB of model-based estimates in small areas as a percent of direct estimates.

| Method of estimation | Private insurance | | Medicad | | Self-paid | |
|---------------------------------------|-------------------|-----|---------|-----|-----------|-----|
| | RRMSE | ARB | RRMSE | ARB | RRMSE | ARB |
| Direct estimates | 100 | 100 | 100 | 100 | 100 | 100 |
| Logistic model without random effects | | | | | | |
| Correct model | 35 | 18 | 33 | 27 | 27 | 23 |
| Correct model+ | 38 | 18 | 36 | 25 | 31 | 32 |
| Incorrect model | 155 | 710 | 118 | 524 | 51 | 293 |
| Incorrect model+ | 146 | 651 | 97 | 396 | 43 | 200 |
| Logistic model with random effects | | | | | | |
| Correct model | 69 | 35 | 65 | 34 | 76 | 71 |
| Correct model+ | 70 | 35 | 66 | 35 | 76 | 70 |
| Incorrect model | 81 | 72 | 75 | 69 | 78 | 75 |
| Incorrect model+ | 81 | 69 | 75 | 68 | 77 | 76 |

Despite large national average differences between three studied proportions, the small area estimates of these proportions exhibited remarkable similarities. As expected, the most efficient and unbiased estimates were obtained by the logistic model without random effect at the state level using the same covariates as superpopulation model. RRMSE was 27-35% and bias was 18-27% of the corresponding values obtained from direct estimates. Adding extra covariates which were not used for simulation added some noise to the model, slightly reducing efficiency. Estimates of small areas from a model without random effects and missing some of the relevant covariates were substantially inferior to direct estimates. They were much less efficient and demonstrated bias, that was 3-7 times larger than bias of direct estimates. Adding extra insignificant covariates to the estimating model improved performance somewhat, which can be explained by the correlation with missing essential covariates.

At the same time estimates from logistic models with random effects demonstrated robustness toward model misspecification. Omitting important covariates used in the superpopulation model did not substantially affect RRMSE or absolute relative bias. Including extra insignificant covariates marginally improved model performance due to the possible correlation with missing significant covariates. On the other hand, if the model used the “correct” set of covariates, adding state level random effects resulted in an almost two-fold increase of RRMSE and relative bias of small-area estimates.

To summarize, including state level random effect provided protection against possible critical failure of model-based estimates. State level estimates became somewhat less efficient, but remained substantially better than direct estimates.

Conclusions

In this simulation study we developed an approach for using models inferred from the existing NAMCS data to estimate the efficiency of different estimates in small areas under the new sampling scheme adopted for NAMCS 2012. Estimates of standard errors of direct estimates in small areas are known to be very unreliable under a randomization-based approach. Smoothed out estimates inferred from this simulation study provided useful insights on the feasibility of direct estimates in small areas and can potentially identify individual small areas for which estimates are the least efficient.

The critical importance of the correct specification of model covariates on model-based estimation was demonstrated when a logistic regression model without random effects was used to estimate proportions in small areas. The model using the correct set of covariates demonstrated great performance, far exceeding that of direct estimates. But when significant covariates were missing from the estimation model, predictions in small areas rapidly became inefficient and extremely biased.

Including random effects in the model with the correct set of fixed effects significantly reduced the efficiency of the model because clustering is wrongly assumed to exist in the data at the state level, whereas according to the superpopulation model, residuals between model and data represented uncorrelated variability between individual physicians. But when covariates were missing from the estimation model, the aggregated effect on state level estimation was different for states with different socio-economic indicators. Clustering of residuals within states were effectively accounted for by including random effect in the model.

The effect on predictions when insignificant covariates were included in the model was demonstrated for models with and without random effects. If the estimating model had all the covariates from the superpopulation model, including extra covariates introduced additional noise in the model parameter estimates which made predictions in small areas slightly less efficient. When the estimating model was missing some of the significant covariates, adding extra covariates that were correlated with missing ones improved predictions in small areas, particularly for models without random effects. The differences due to including extra covariates were almost negligible for models with random effects.

The results from this conducted simulation study suggest that the following strategy should be followed for efficient model-based estimation of proportions in small areas from NAMCS data:

- Include all covariates which may look significant for estimation. Having extra covariates has little effect, but missing significant covariates may have disastrous consequences;
- Use the maximum likelihood estimate of the state-level random effects variance component to assess the utility of including state-level random effects. If, after including physician-level random effects, the state-level variance component MLE is estimated to be greater than 0, it means that utilized covariates do not sufficiently explain aggregated variability in small areas. In such cases random effects at the state level must be included in the model to provide robustness against probable model misspecification.

In this paper we assumed that random effects are always normally distributed in both superpopulation and estimating models. In the future, we will consider the robustness of this normality assumption in the estimating model. For that purpose we will simulate the final population with state level random effects distributed from the Dirichlet process, or having skewed or heavy-tailed distribution functions, but the estimating model will still use normally distributed random effects.

References

[1] Rao, J.N.K (2003) Small Area Estimation, John Wiley & Sons.

[2] National Center for Health Statistics. Description of the National Ambulatory Medical Care Survey (NAMCS) http://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NAMCS/