Application of Small Area Estimation Methods to Emergency Department Data from the National Hospital Ambulatory Medical Care Survey

Vladislav Beresovsky[1], Catharine W. Burt[1], Van Parsons[1], Nathaniel Schenker[1]

[1]National Center for Health Statistics, 3311 Toledo Rd, Hyattsville, MD 20782

Abstract

The National Hospital Ambulatory Medical Care Survey (NHAMCS) is designed to produce national-level estimates, but sample sizes are inadequate for direct county- and state-level estimation. To expand upon estimation capabilities, data from a sample of emergency departments (ED) from the NHAMCS were combined with universe county- and hospital-level covariates from the Area Resource File and Verispan Hospital Database to create small-area prediction models for estimating county- and state-level attributes of emergency department visits (rates of ambulance arrival and visits with asthma or injury). Effects of data clustering at the hospital, county or state levels (due to the multi-level NHAMCS sample design) were modeled by introducing random effects into a generalized linear model. The estimation process employed the SAS procedure GLIMMIX. Point estimates were calculated and compared when random effects were applied at different levels. A bootstrap approach to estimating mean squared errors was illustrated as well.

Key Words: health care utilization, small area estimation, logistic regression, random effects, cluster sampling, parametric double-bootstrap

1. Introduction

The proliferation of the use of electronic medical records holds great promise for obtaining healthcare information in small areas. Currently administrative data collected electronically from the billing records of hospital emergency departments (EDs) are available for about 27 states as part of the Nationwide Emergency Department Sample (NEDS) [1]. These billing records carry a subset of ED clinical data such as diagnoses and patient demographic data but do not carry a full range of utilization information related to treatment and throughput. Considering the currently limited level of adoption of electronic medical records, sample record abstraction remains the primary way to gather clinical information about emergency care provided.

In an ideal situation, there would be sufficient funds available to design and conduct a sample survey that would collect a sufficient amount of emergency care data from each county and state in the United States so that reliable direct estimates would be possible in small areas. But that is not presently feasible.

The National Hospital Ambulatory Medical Care Survey (NHAMCS) is an annual national probability sample of visits to the emergency, outpatient, and ambulatory surgery departments of noninstitutional general and short-stay hospitals, excluding Federal, military, and Veterans Administration hospitals, located in the United States. NHAMCS collects data via medical record abstraction and uses a four-stage probability design starting with selection of geographic primary sampling units (PSUs), hospitals within PSUs as secondary sampling units (SSUs), clinics/emergency service areas within outpatient/emergency/ambulatory surgery departments, and patient visits within clinics/emergency service areas. NHAMCS was designed to provide utilization estimates based on the following priority of levels: national, regional, and Metropolitan Statistical Area (MSA) versus non-MSA designated areas [2].

While these traditional estimates are of significant interest and importance, there is also growing interest in estimates for smaller geographical localities, like states and counties. However, the possibility of producing reliable direct estimates in small areas using NHAMCS ED data is an open question because of limited coverage in small areas, small sample sizes within many covered small areas and the high level of data clustering used in the sample design [2]. For instance, 75% of all EDs are located in counties that were not sampled in NHAMCS 2007 and another 15% of EDs are located in counties with only one or two sampled EDs. Coverage for states is substantially better, yet for many states the reliability of direct estimates is questionable.

Apparent limitations of direct estimates in small areas suggest a need to consider model-based methods. The feasibility of model-based small area estimates for states and counties using data from the National Health Interview Survey (NHIS), a complex sample containing about 100,000 persons, has been demonstrated by Malec et al [3]. They estimated the proportion of the population making at least one visit to a physician office within the last 12 months. Although NHAMCS sampling units (ED visits), sample size (~35,000 visits), and SSUs (hospitals) differ from those in the NHIS, both surveys share similar geography through the PSU structure [4].

In Malec et al [3] small area estimates for demographic classes defined by age, race and gender were obtained from a logistic regression model with random coefficients utilizing county-level covariates taken from the Area Resource File (ARF, http://arf.hrsa.gov). Model parameters were estimated using exact hierarchical Bayes and empirical Bayes methods. Predicted posterior means and standard deviations of proportions were calculated using both methods and compared for states and demographic subpopulations within states. Posterior means of proportions were found to be close for both methods. It was demonstrated that using empirical Bayes underestimates posterior standard deviations when inferences are to be made for subpopulations within states, or when inferences are required for entire states but the sample size is small.

The present study is work in progress on using a logistic regression model with random intercept to estimate proportions in small areas for three ED visit characteristics: injury-related diagnosis, arrival by ambulance, and asthma diagnosis. We believe that the reliability of small area estimates depends on the national average propensities of occurrence. Visit characteristics studied in this paper have varying average propensities, estimated from the NHAMCS ED data: injury-related diagnosis (~26%), arrival by ambulance (~ 16%) and asthma diagnosis (~2%).

2. Methods

The production of small area estimates from provider-based surveys such as the NHAMCS involves a number of steps starting with building a probabilistic model from the sample data and selected covariates that can be used to predict attributes of provider encounters for all providers in the universe of interest. For this study, the provider is the general or short-stay hospital that has a 24-hour emergency department. After the attribute of interest for each hospital is predicted, the hospital estimates are aggregated to form the county- and then state-level estimates. Estimates of the error associated with the small area estimates are also made. The modeling steps are described below.

To account for major sources of variation in the attributes of interest, we considered a logistic regression model with random effects. At the first level, the proportion of visits with the characteristic of interest for each hospital was assumed to have a binomial distribution with a selection probability specific for each hospital. At the second level, a multivariate linear regression model used hospital- and county-level covariates to predict proportions for each hospital in the population. The random term of the second-level model accounts for random deviation of the logit of studied proportions for each sampled hospital from the prediction by linear combination of fixed effects. Proportions predicted for the hospitals were aggregated to county- and state-levels by utilizing the number of ED visits to each hospital in the population available from the Verispan (formerly known as SMG) Hospital Database. As stated in Malec et al [3], such aggregation effectively replaces national stratification of the sample with a new set of weights at the local level equal to the number of ED visits (see below). The described model-based approach incorporates data clustering by including random effects. To the extent that the covariates account for the weighting factors, such factors are incorporated as well. For the estimated mean square error (MSE) of the predicted proportions in small areas we used a double bootstrap technique involving replications of simulated data based on the predicted binomial distributions of the propensity of the attribute in each hospital.

2.1 Data

To achieve greater precision in estimating model parameters we combined NHAMCS ED data for 2006 and 2007. Combined data included 362 hospitals from the 2006 sample and 337 hospitals from 2007 sample. This increase in sample of hospitals is expected to reduce the variance of estimation. Some of the hospitals (250) happened to be sampled in both years, but studied proportions may still vary between years due to the seasonality of the data collection process [2]. Data from the repeatedly sampled hospitals were treated by the model as independent observations.

For each sampled ED visit, NHAMCS classifies and codes three provider diagnoses DIAG1-DIAG3 according to the International Classification of Diseases, version 9, Clinical Modification (ICD-9-CM, www.cdc.gov/nchs/icd/icd9cm.htm). If ICD-9-CM codes for injury (800-998) or asthma (493) were recorded in any of the listed diagnoses, then corresponding indicator variables were assigned to 1, if not- to zero. Proportions were defined in relation to the total number of ED visits. Mode of arrival to hospital ED is recorded in the survey database variable “ARRIVE” having the following values: 1='Ambulance; 2='Public service (nonambulance)'; 3='Personal transportation'; -8='Unknown'; -9=’Blank’. Arrivals by ambulance were identified if ARRIVE=1. Proportions were defined in relation to the number of ED visits with known mode of arrival.

County-level covariates for building a model are available from the 2007 Area Resource File distributed by Health Resources and Services Administration (HRSA). If separate values were available for different years, we used corresponding values for each year of survey data although the differences between years were usually not large. We also used hospital-level covariates from the 2006 and 2007 Verispan Hospital Databases. This additional information accounts for fixed effects explaining variability between hospitals and provides for higher sensitivity of the model. Model covariates grouped by their sources and type of information are presented below.

Table 1: County and hospital level model covariates

County covariates: Demographic	Population; Number of births and deaths per 100,000 population; Percent of blacks, Hispanics, and population under 18 years and at least age 65; Population loss typology.
Social	Median income; Percent of people in poverty, unemployed, uninsured, receiving food stamps, eligible for Medicare; Housing and education typology; Medicaid discharges per 100,000 population.
Economic	Economic dependence typology (Farming, services, manufacturing)
Health care	Number of dentists, primary care physicians, pediatricians, hospital visits and admissions, hospitals and hospital beds per 100,000 population.
Hospital covariates:	Numbers of hospital outpatient department visits and beds; hospital ownership; existence of a residency program; affiliation with a medical school; existence of a trauma unit, intensive care unit, and pediatric care unit.
Sample Design covariates:	Geographic region; MSA status; Number of ED visits

All continuous covariates were standardized before being used for modeling by centering and normalizing by standard deviation:

$$X^{STD} = \frac{\left(X - \overline{X}\right)}{StdDev\left(X\right)} \tag{1}$$

In addition to this linear transformation outliers were truncated at the 99th percentile to improve robustness of the model. Standardizing covariates in many cases improves the convergence of numeric algorithms; model parameters become measured on the same scale and more easily interpretable.

2.2 Probabilistic Specification and Methodology

2.2.1 Model description

Logistic regression models with normally distributed random effects, simulating possible clustering in the data distribution, are commonly used to estimate small area proportions [3, 5]. In principal, clustering can be considered at various levels of aggregation. Because hospitals are secondary sampling units in the NHAMCS, it would be natural to assume clustering of visits at the hospital level. If data in sampled hospitals can be considered representative of the entire small area, then the whole small area can be treated as a clustering unit. For general treatment, let *j* denote the level of application of a random effect, which can be hospital, county, or state. Let Y_{ij} denote a random variable equal to the number of visits with a certain characteristic of hospital *i* within clustering unit *j*. If the total number of ED visits equals N_i, Y_{ij} will have a binomial distribution:

$$Y_{ij} \left| p_{ij} \right. \sim \; binomial\left(N_i, p_{ij}\right) \tag{2}$$

Model parameter p_{ij} is a random variable itself, which can be modeled by a column vector of *M* county- and hospital- level covariates $\mathbf{X}_i = \left(X_{i1},..., X_{iM}\right)$ and random effect θ_j applied at clustering level *j*:

$$\operatorname{logit}\left(p_{ij}\right) = \ln\left(p_{ij}\big/\left(1 - p_{ij}\right)\right) = \mathbf{X}_i\boldsymbol{\beta} + \theta_j ; \theta_j \; \sim \; F\left(0, \sigma_a^2\right) \tag{3}$$

Some methods of estimation require the exact specification of the distribution function for the random effect, but others do not. At this point, the random effect is assumed to have a distribution with mean *0* and variance σ_a^2. Later in the study we will use replication methods for MSE estimation. For data simulation purposes we will be assuming a normal distribution of the random effect.

2.2.2 Estimation method

Parameters of the described model $\boldsymbol{\beta}$ and σ_a^2 can be estimated from maximization of the marginal likelihood function:

$$L\left(\boldsymbol{\beta}, \sigma_a^2, y\right) = \int f\left(\mathbf{y} \left| \boldsymbol{\beta}, \theta\right.\right) p\left(\theta\right) d\theta \tag{4}$$

where $f\left(\mathbf{y} \left| \boldsymbol{\beta}, \theta\right.\right)$ is the conditional distribution of the data, and $p\left(\theta\right)$ is the distribution of the random effects. In the case of a linear model with normally distributed components, this integral can be solved in closed form, and the resulting likelihood can be maximized directly. The model under consideration has a logit link function with no closed form solution, and the model parameters can only be found by approximation techniques. After comparing several alternative statistical procedures which gave very similar results, we used the SAS procedure GLIMMIX which uses a linearization of the likelihood function for solving the problem in a nonlinear case [6, 7].

NHAMCS ED data for 2006 and 2007 were used for estimation of parameters of the two-level model (2), (3). The modeling process demonstrated that many of the county- and hospital- level covariates proved to be significant for modeling studied proportions and improved the model fit. Accounting for random effects, particularly at the hospital level, dramatically improved model fit for the sampled hospitals.

2.2.3 Predicting hospital proportions from the estimated model parameters

We used the estimated model parameters and predicted random effects (3) to predict proportions in each hospital in the universe:

$$\hat{p}_{ij} = \left(1 + \exp\left(-\left(\mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\theta}_j\right)\right)\right)^{-1} \tag{5}$$

If clustering unit *j* did not contain any sampled hospitals, the estimated random effect $\hat{\theta}_j$ was set equal to zero. Only the 2007 data values were used for the prediction whereas the model building involved both 2006 and 2007 values.

2.2.4 Aggregating hospital predictions to small area

From the Verispan Hospital Database we know the number of visits N_{ij} to every hospital ED, *i,* within small area, *j*. These numbers can be used as weights for calculating proportions in small areas by aggregating predicted hospital proportions $\hat{p}_{ij}$:

$$\hat{P}_j = \frac{\sum_{i\in j} N_{ij}\hat{p}_{ij}}{\sum_{i\in j} N_{ij}} \qquad (6)$$

Utilizing the numbers of ED visits for each hospital in the United States (from the Verispan Hospital Database) for aggregating predicted hospital proportions to the small area level is a crucial step for producing model-based estimates from the sample data and also to account for frame information which is necessary for unbiased direct estimation. It will be demonstrated below that described approach produces model-based estimates for large populations which are comparable with direct estimates.

2.3 MSE Estimation

Stochastic variability of small area predictions underscores the importance of being able to estimate MSE of predicted proportions. Hall and Maiti [8] proposed a parametric double bootstrap method for calculating bias-corrected MSEs which is based on replications. The technique can be used with a broad range of small-area models. In application to the current model, the double-bootstrap estimator of MSE may be constructed as follows:

1) Fit the model to sampled data y_i (number of visits with a certain characteristic) and find parameter estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_a^2$;

2) Using estimated model parameters draw random effect $\theta_j^* \sim N\left(0,\hat{\sigma}_a^2\right)$ at clustering unit *j* and calculate from (6) the simulated proportion p_i for each hospital *i* of clustering unit *j*;

3) For each sampled hospital, *i,* draw a simulated sample from $y_i^* \sim \ binomial\left(n_i, p_i^*\right)$, where n_i is the number of sampled visits;

4) Repeat steps 1) to 3) using y_i^* instead of y_i. As a result we will obtain estimated model parameters $\hat{\boldsymbol{\beta}}^*$ and $\hat{\sigma}_a^{2*}$, simulated random effect θ_j^{**} and "true" proportion p_i^{**} for all hospitals and simulated data y_i^{**} for each sampled hospital. All values simulated and estimated at the second step are conditional on the sample y_i^* simulated on the first step;

5) Fit the model to simulated data y_i^{**} and find predicted proportions $\hat{p}_i^{**pred}$ for all hospitals;

6) For small area j aggregate simulated “true” $p_{ij}^{**} = y_i^{**} / n_i$ and predicted $\hat{p}_{ij}^{**pred}$ hospital proportions to small area proportions P_j^{**} and $\hat{P}_j^{**pred}$ using (7);

7) Find the estimator of MSE on area j, $\hat{u}_j = \widehat{MSE}_j^* = E\left[\left(\hat{P}_j^{**pred} - P_j^{**}\right)^2 \middle| y^*\right]$, conditional on the data simulated at the first step and then its expectation, $\hat{v}_j = E\left[\widehat{MSE}_j^* \middle| y\right]$, conditional on the original sample data.

Hall and Maiti [8] proposed three different bias-corrected non-negative MSE estimators constructed from $\hat{u}_j$ and $\hat{v}_j$. Our study confirmed that all three estimators produced similar results.

3. Results

County- and state-level inferences for the proportions of ambulance arrivals estimated from the models with random effects applied at hospital-, county- and state-levels and also for the model with only fixed effects are presented and compared on Figures 1 and 2. Because predictions using the fixed effects model are considered to be a baseline for small area estimates, counties and states are ordered in ascending order by predictions of the fixed effects model for easy comparison. Predicted proportions of visits with injury and asthma diagnoses demonstrate similar behavior compared to those for ambulance arrival and will not be presented in this paper for the sake of brevity.

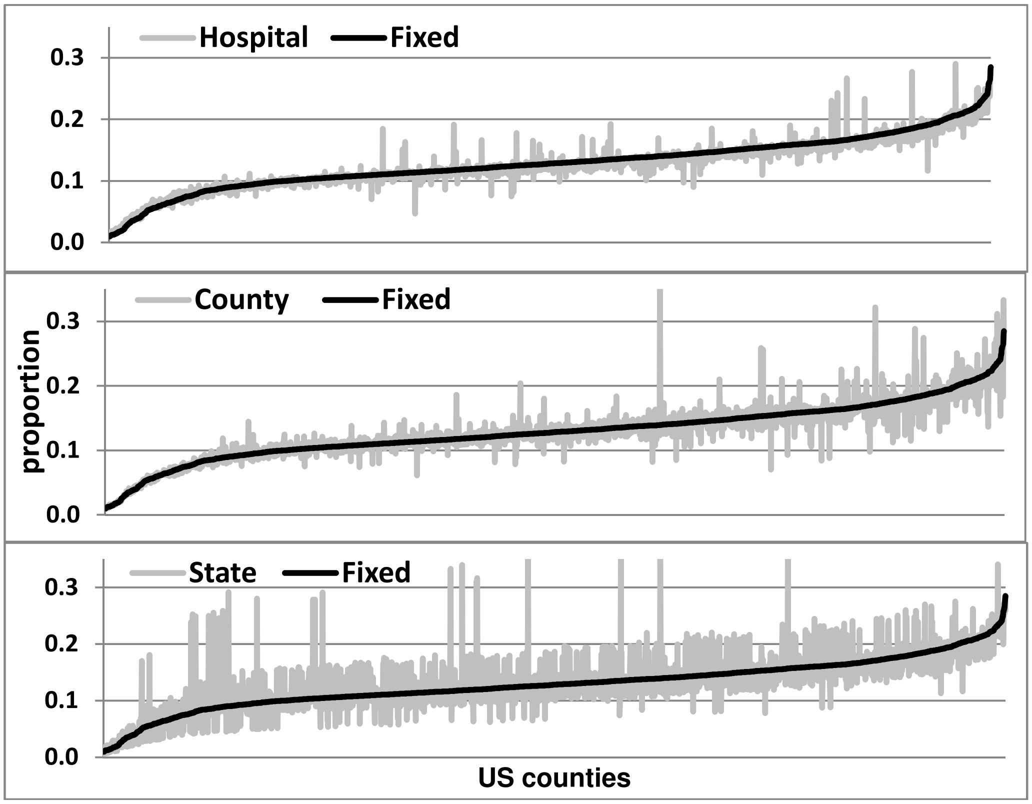

Figure 1: Proportions of ambulance arrivals to ED predicted by models with random effects applied on hospital, county and state levels and aggregated to county-level compared to predictions using a model with fixed effects only.

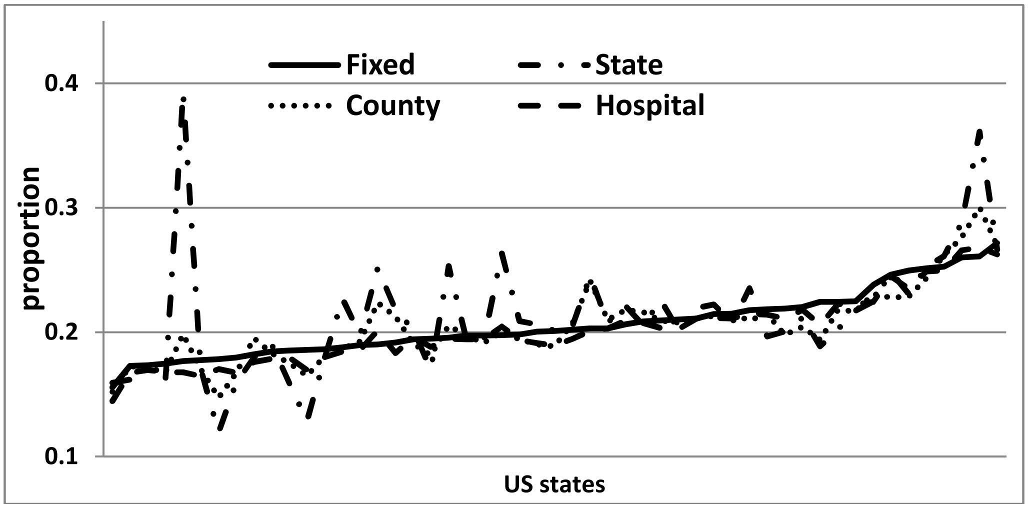

Figure 2: State-level proportions of ambulance arrivals to ED predicted by models with random effects applied on hospital, county and state levels compared to predictions using a model with fixed effects only.

Estimated county- and state- level proportions consistently demonstrate that predictions using the model with random effects at the hospital level are closer to predictions using the fixed effects model than were the other models. Deviation from the fixed effects model increases when the application level of random effects becomes coarser, from hospital to county and to state. These phenomena are consequences of the use of zeroes for the estimated random effects $\hat{\theta}_j$ for clustering units that did not contain any sampled hospitals.

The estimated MSE of model-based predictions depends on the estimated variance of random effect $\hat{\sigma}_a^2$ and error of estimation of model parameters $\hat{\boldsymbol{\beta}}$. The average estimated MSE of predicted proportions of ambulance arrivals to EDs using models with different applications of random effects was calculated for sampled and nonsampled counties and states. Because the MSE is easier to interpret when it is related to the point estimate of the proportion, the estimated relative root mean square errors (RRMSE) defined as $RRMSE = \sqrt{MSE}\big/P*100\%$ are presented in Table 2.

Table 2: Average estimated RRMSE of predicted proportions of ambulance arrivals to EDs in small areas depending on application of random effects.

		Average estimated RRMSE (%) by random effect			
Small Area		Hospital	County	State	Fixed
Counties	Sampled	29.7	22.4	18.8	4.2
	Not sampled	50.3	39.7	17.8	7.8
	All	48.7	38.4	17.9	7.6
States	Sampled	9.4	8.6	9.8	2.6
	Not sampled	16.2	17.4	31.7	3.5
	All	10.4	10.0	13.3	2.8

(Note: RRMSEs estimated by different models cannot be compared for making conclusion about preferred model)

While presenting this table we must not make conclusions about the performance of different models by comparing their estimated MSE because it is defined by the stochastic structure intrinsic to each model. However, estimating MSE will be important for model diagnostics when direct estimates in small areas from administrative data become available.

Comparisons of the model-based predictions with direct estimates are demonstrated for national, regional and MSA status- estimates in Table 3. According to the NHAMCS design, such direct estimates can be considered accurate, thus providing another baseline for validating model predictions.

Table 3: Model-based and direct estimates of proportions of ambulance arrivals to EDs within large populations with corresponding RRMSE(%) depending on application of random effect

	Random effect placement									
Aggregate	Hospital		County		State		Fixed		Direct	
area	Prop	RRMSE	Prop	RRMSE	Prop	RRMSE	Prop	RRMSE	Prop	RRMSE
Northeast	0.200	3.8	0.203	3.3	0.206	3.2	0.206	1.6	0.205	6.6
Midwest	0.146	4.3	0.150	3.7	0.151	3.9	0.149	2.0	0.172	8.3
South	0.151	3.7	0.148	3.5	0.155	3.5	0.154	1.6	0.156	7.1
West	0.144	4.5	0.152	3.9	0.156	4.3	0.153	2.1	0.15	7.7
MSA	0.165	2.2	0.166	1.5	0.171	2.9	0.170	0.9	0.174	4.2
Non-MSA	0.128	6.4	0.133	5.7	0.137	5.7	0.131	3.7	0.128	7.5
All US	0.158	2.1	0.160	2.1	0.164	2.7	0.162	0.9	0.167	3.9

The model-based estimates are consistent with the direct estimates with a notable exception for the Midwest region. Also, many of the model-based estimates appear to be slightly lower than direct estimates. These discrepancies were not observed for models of visits with injury and asthma diagnoses (data not shown).

3.1 Discussion

This study demonstrates that model-based small area estimates from sample data of multilevel provider-based surveys strongly depend on the application of random effects. When random effects are applied at the hospital-level, estimates for counties and states are not much different from the estimates by fixed effect model. But when random effects are applied at county- and state-levels, estimates in those areas deviate from the fixed effects model towards the direct estimates. We believe that such behavior is a consequence of low sampling fractions of hospitals (SSU) and counties in NHAMCS. If more counties and hospitals were sampled, then estimated proportions would be less dependent on the level of application of random effects and gradually converge with direct estimates towards "true" proportions in counties and states. This conjecture could be validated using simulated data.

The estimated MSEs for predicted proportions become larger when random effects are included in a model. Fixed effects models only account for variability associated with the lowest units of analysis (visits) and ignore additional variability associated with data clustering at higher levels (hospitals, counties, or states). Such simplification of the stochastic structure of the predictive model leads to underestimation of MSEs and the widths of corresponding confidence intervals. Inclusion of random effects in the models allows for a more realistic estimation of the MSE.

Dependence of estimates in small areas on the application of random effects does not affect the general consistency observed between the aggregated model-based small area estimates and the NHAMCS design-based direct estimates in large areas (Table 3). We believe that using a rich set of county- and hospital- level population covariates and number of ED visits for each hospital in the United States for modeling and aggregation was very important for consistent estimates.

According to general NCHS guidelines, estimates are considered reliable if the standard error is less than 30% of the point estimate. The following table summarizes the reliability of model based small area predictions based on these guidelines.

Table 4: Reliability of small area predictions, based on estimated MSEs with random effects at the hospital level

Predicted proportion	National average (%)	Unsampled counties	Sampled counties	Unsampled states	Sampled states
Asthma	2	Unreliable	Unreliable	Mixed	Reliable
Ambulance arrivals	16	Unreliable	Mixed	Reliable	Reliable
Injury	26	Reliable	Reliable	Reliable	Reliable

4. Conclusions, Implications for Practice and Next Steps

This study presents a model-based approach for small area estimation of proportions for NHAMCS ED data. To build a better model and to avoid bias associated with seasonality of the data collection process, we combined two years of survey data. County- and hospital-level covariates from the Area Resource File and Verispan Hospital Database proved to be significant for modeling the studied proportions. Available numbers of ED visits for each hospital in the United States were used to effectively recalculate national survey weights for each small area. Random effects applied to different levels of clustering have significant influence on the predicted proportions and estimated MSE. In future we will consider using a model that simultaneously includes random effects for multiple levels of clustering - hospitals, counties and states. Estimating parameters of such a model may require using different software and a Markov Chain Monte Carlo (MCMC) algorithm.

Conclusions about reliability of small area estimates are based on the estimated MSE of the predicted proportions. However, the MSE could be underestimated because it does not include the error term associated with model misspecification since the "true" model is not known. Ultimately, predictions for small areas could be validated by comparing them to direct estimates based on administrative data which are considered sufficiently close to true population values. Administrative data for selected states are available from the NEDS databases. We hope to be able to compare our results with NEDS state- and county-level statistics for the available 27 states.

This study is timely and important because it develops methodology, demonstrates the possibility, and shows the limitations of estimating proportions within small areas using nationally designed survey data from health care providers.

References

[1] Agency for Healthcare Quality and Research. National Emergency Department Sample (NEDS). http://www.hcup-us.ahrq.gov/nedsoverview.jsp

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

[3] Malec, D., Sedransk, J., Moriarity, C.L., and LeClere, F.B., Small Area Inference for Binary Variables in the National Health Interview Survey, *JASA* 92 (1997), 815-826.

[4] Massey J.T., Moore T.F., Parsons V.L., Tadros W., Design and estimation for the National Health Interview Survey, 1985-94. National Center for Health Statistics. *Vital and Health Statistics* 2 (110).1989.

[5] MacGibbon, B., and Tomberlin, T.J., Small area estimates of proportions via Empirical Bayes Techniques, *Survey Methodology*, 15 (1989), 237-252.

[6] Wolfinger, R. and O'Connell, M., Generalized Linear Mixed Models: A Pseudo-Likelihood Approach, *Journal of Statistical Computation and Simulation*, 4 (1993), 233–243.

[7] Breslow, N. E. and Clayton, D. G., Approximate Inference in Generalized Linear Mixed Models, *JASA*, 88 (1993), 9–25.

[8] Hall, P. and Maiti, T.,. On parametric bootstrap methods for small area prediction, *J. R. Statist. Soc. B* 68 (2006), 221-238.