SMALL AREA ESTIMATION FOR THE U.S. NATIONAL HEALTH INTERVIEW SURVEY

David A. Marker, Westat, Inc.
1650 Research Blvd., Rockville, MD

KEY WORDS: Bayes, synthetic estimation, composite estimation, mean square error

1. Introduction

The National Health Interview Survey (NHIS) is a major source of health statistics for the United States' population. It is based on an area probability sample, with a separate stratum to account for new construction since the last census. Data are collected year-round, with only a one week break at the beginning of each year. Fifty thousand housing units are sampled each year, with data collected on all 125,000 people living at those addresses. Every ten years the NHIS is redesigned based on new Census data and responding to updated priorities for the survey. The two primary goals of the current redesign (to be used by the NHIS for 1995-2004, budget permitting) are to produce accurate estimates for the 50 states plus the District of Columbia, and for blacks and Hispanics nationally.

As part of the plans for the NHIS for 1995-2004, NCHS has worked with the U.S. Bureau of the Census and Westat, Inc., to revise the sample design so as to improve the Center's ability to produce both of these types of estimates. Unfortunately, these two primary goals of the redesign are in substantial conflict. The overwhelming majority of blacks and Hispanics live in the largest states; those for which the sample sizes are already relatively large. Improving the ability of the NHIS to provide minority statistics requires allocating more of the sample to the large states, thereby decreasing the sample sizes in smaller states. Alternatively, improving state-level estimates for the smaller states involves reducing the sample sizes in large states, hence decreasing the accuracy of estimates for blacks and Hispanics.

Given this conflict between the two priorities, a decision was made to oversample blacks and Hispanics to improve estimates of their health and only to use state stratification to improve the ability to provide estimates at the state level. The resulting sample sizes in all but 10 states will be too small to produce direct estimates of the desired accuracy. The alternative is to employ some form of small-area estimation technique to produce state-level estimates using a composite estimator with a model-based component. The remainder of this paper addresses the issues of how to combine a direct design-unbiased state sample estimate with a model-based estimator, and how to estimate the accuracy of the resultant composite estimator.

This paper begins by reviewing the existing small-area estimators that may be appropriate for the NHIS. It then introduces a new, more general, small-area estimator, provides an empirical comparison of alternative estimators that was carried out on the current NHIS, and describes attempts to produce state-specific mean square errors for these estimators.

2. Model-Based Estimators

2.1 Existing Alternatives

Small-area estimates of most health measures face limitations beyond those found for demographic or economic measures. "Truth" is generally not measured, that is, there is no census that asks sufficient health questions that would provide a baseline against which models could be developed or tested. Existing procedures that have potential for the NHIS are synthetic estimation (first developed for use on the NHIS (NCHS, 1968)), and composite estimation combining the direct sample estimator and the synthetic estimator (Schaible 1971).

The form of the classical synthetic estimator of the mean for small area i is:

$$\hat{\bar{y}}_{i..} = \sum_{j=1}^{J} \frac{N_{ij}\,\bar{y}_{.j.}}{N_{i.}}$$

where:

$\hat{\bar{y}}_{i..}$ = synthetic estimator for state i;
N_{ij} = total population for subgroup j in state i;
$\bar{y}_{.j.}$ = sample mean for subgroup j across all states; and,
$N_{i.}$ = total population for state i.

Composite estimation takes a weighted average of the direct design-unbiased estimator and this synthetic estimator, where the weights are inversely proportional to the mean square errors of the two estimators.

For either of these estimators to be unbiased, it is necessary that for any population subgroup the mean is the same across every small area; for example, the average number of doctor visits by 20-44 year old black males is the same in every state. This model assumption is always false. However, if it is approximately correct, the biases will be small, and tolerable.

2.2 Generalized Synthetic Estimator

A more general small-area estimator can be developed that does not make this restrictive assumption of constant averages across all small areas (Marker). Ericson (1969) proposed applying the Bayesian concept of exchangeability to survey sampling. In the following small-area estimator, exchangeability is used at two levels. First, the individuals are assumed to be exchangeable within small area/subgroup; this implies that the values of the characteristic of interest (e.g., the number of doctor visits) of all individuals in a given small area/subgroup are distributed with a common mean and variance. Second, assume that for each subgroup the small area/subgroup means and variances are exchangeable across small areas. Thus, the small area/subgroup means are a random sample drawn from a population with a common subgroup mean. This allows the average number of doctor visits by 20-44 year old black males to vary from state to state, but to have a common expectation and variability about the overall mean for 20-44 year old black males.

Let Y_{ijk} be the value (e.g., the number of doctor visits) of the

i^{th} small area, i = 1, 2,..., I,
j^{th} subgroup, j = 1, 2,..., J, and
k^{th} individual k = 1, 2,..., N_{ij}.

We can now rigorously characterize the prior distribution as follows:

Assumption 1: Conditional on their mean and variance $(\mu_{ij}, \sigma^2_{ij})$, the Y_{ijk} are exchangeable within

small area/subgroups and are independent between small area/subgroups.

Assumption 2: For each subgroup j, the ordered pairs $(\mu_{ij}, \sigma^2_{ij})$ are exchangeable across small areas i, and they are independent between subgroups j with:

$E(\mu_{ij} \mid m_j) = m_j$, (the prior expectation within subgroup j)

$V(\mu_{ij} \mid m_j) = v_j$,

$Cov(\mu_{ij}, \mu_{rj}) = c_j \quad i \neq r$,

$E(\sigma^2_{ij}) = \phi_j$,

To properly represent the assumptions of the classical synthetic estimator, define a second level in the prior distribution:

$E(m_j) = m^*_j$

$V(m_j) = v^*_j$

$P_j = v_j + v^*_j - c_j$

$S_{ij} = (1 - f_i)\,\phi_j / n_{ij}$

P_j and S_{ij} are the prior and sampling variances, respectively, of $\bar{y}_{ij}$. To ensure that weights in the resulting estimator remain positive, it is necessary to have the following restrictions:

either $c_j > 0$, or $c_j < 0$ and $|c_j| \leq \frac{v_j + v^*_j}{(I - 1)}$; and

$$\sum_{i=1}^{I} \frac{P_j}{P_j + S_{ij}} \leq \frac{N^2_j}{\sum_{i=1}^{I} N^2_{ij}},$$

Expanding on the work by Ericson (1981, 1983), it can be shown (Marker) that the posterior expectation given the prior distribution described under these two assumptions is (let (s, y) represent the sample s containing values y):

$$\bar{E}(Y_{i..} \mid (s, y)) = \left[n_{i.}\,\bar{y}_{i.} + \sum_{j=1}^{J} (N_{ij} - n_{ij}) \frac{\hat{\mu}_{.j} \left(\frac{\sum_{i=1}^{I} N^2_{ij}}{N^2_{.j}} P_j + c_j \right) + m^*_j \left(\left(\sum_{i=1}^{I} \frac{1}{S_{ij} + P_j} \right)^{-1} - \sum_{i=1}^{I} \frac{N^2_{ij}}{N^2_{.j}} P_j \right)}{\left(\sum_{i=1}^{I} \frac{1}{S_{ij} + P_j} \right)^{-1} + c_j} \right] / N_{i.} \qquad (1)$$

This Bayes estimator is called the Generalized Synthetic Estimator (GSE). The reason for this name becomes clear if we make three further assumptions.

1. We assume that the mean response for a given subgroup is the same for all small areas, i.e., $\mu_{ij} = \mu_{.j}$ for all j, or

   $V(\mu_{ij} \mid m_j) = v_j = 0$

   for all j. This also implies that the covariance between any two such means for a given subgroup is equal to the variance of one of the means, i.e.,

   $c_j = Cov(\mu_{ij}, \mu_{rj}) = V(\mu_{ij}) = v^*_j$.

   Thus, $P_j = v_j + v^*_j - c_j = 0$.

2. Assume that the prior distribution for m_j is diffuse (at least relative to the S_{ij}) so that

   $V(m_j) = v^*_j = \infty$ then

   $$\bar{E}(Y_{i..} \mid (s, y)) = \left[n_{i.}\,\bar{y}_{i.} + \sum_{j=1}^{J} (N_{gj} - n_{gj})\,\hat{\mu}_{.j} \right] / N_{i.} \qquad (2)$$

   $\hat{\mu}_{.j}$ is the weighted average of the small area/subgroup means $\bar{y}_{ij.}$, this is only equal to the overall sample average for the subgroup $\bar{y}_{.j.}$ if we assume

3. The expected elemental variances of the y_{ijk}, $E(\sigma^2_{ij}) = \phi_j$, are equal for all areas. Then:

   $$\bar{E}(Y_{i..} \mid (s, y)) = \left[n_{i.}\,\bar{y}_{i.} + \sum_{j=1}^{J} (N_{ij} - n_{ij})\,\bar{y}_{.j.} \right] / N_{i.} \qquad (3)$$

   Finally, assuming there are no sampled elements in small-area i ($n_{i.} = n_{ij} = 0$) results in the classical synthetic estimate first introduced by NCHS (1968):

   $$\bar{E}(Y_{i..} \mid (s, y)) = \sum_{j=1}^{J} N_{ij}\,\bar{y}_{.j} / N_{i.} \qquad (4)$$

This commonly used form of the synthetic estimate is therefore appropriate if the following conditions are true:

(a) The two initial assumptions hold.
(b) In any subgroup, the means for all small areas are equal.
(c) Prior knowledge of the value of the common mean for all small areas is nonexistent.
(d) In any subgroup, the expected variances for all small areas are equal.
(e) There are no respondents in the small area of interest.

If only (e) is not true, the synthetic estimator still holds for the total of those not sampled.

Conditions (a) and (b) are discussed above. Condition (c), that there is no prior knowledge of the common value is in most cases not true. Prior information can be available from a number of sources. The same data may have been collected in a previous year.

If the prior distribution on m_j is not diffuse, then the values of the S_{ij}'s and v^*_j's must be considered. S_{ij} is the expected variance in the average response in any small area for members of subgroup j (e.g., the expected variability for the sample average of 20-44 year old black males in state i). v^*_j is the measure of confidence you have in the prior mean for subgroup j. To not assume that the means within a subgroup are all equal, just use $(v_j + v^*_j - c_j)$ instead of v^*_j.

Condition (d), homoscedastic variances within subgroups, goes unmentioned in all other literature on the synthetic estimator. If the variable is a proportion, then the variances are homoscedastic since

$$E(\sigma^2_{ij}) = E(\mu_{ij}(1 - \mu_{ij})) = m_j - m^2_j$$

because the small area/subgroup means are exchangeable. Similarly, if the variable follows a Poisson distribution the variances are homoscedastic since the variance of a Poisson variable is equal to its mean m_j. If the variable is not a proportion or Poisson (e.g., continuous), this is not necessarily true. Variances are likely to vary since certain small areas are going to have more homogeneous subgroups than others. (If the differences in homogeneity

are not large, the impact of this departure from the assumptions may be minimal.) If one part of the sample is significantly more accurate than another, it is logical to weight its average more highly. While the true variances σ^2_{ij} may vary from small area to small area, they must be exchangeable with a common expected variance.

Close examination of the GSE shows that for each subgroup it is a weighted average of the classical synthetic estimate and the prior expected total. If one feels more confident in the prior for a subgroup than in the sample responses due to a large expected sampling variance, then the Generalized Synthetic Estimator will automatically give more weight to the prior than to the classical synthetic estimate. If instead one is more willing to trust the sample data than the prior, the Generalized Synthetic Estimator will automatically adjust in that direction since P_j will be larger than S_{ij}.

3. Empirical Analysis

Two NHIS variables were studied in the empirical analysis, one the mean of a count variable, and the other a proportion. The count variable was the average number of doctor visits in the past 12 months, and the proportion was the percent with a self-perceived poor health status. The exact forms of the questions on the NHIS questionnaire were as follows:

- During the past 12 months, how many times did [—] see or talk to a medical doctor or assistant? (Do not count doctors seen while an overnight patient in a hospital.)
- Would you say [—] health in general is excellent, very good, good, fair, or poor?

3.1 Estimators

A total of 6 small-area estimators for states were used. They are described below.

1. **Direct, design-unbiased inflation estimator.**

$$\bar{y}_{1i.} = \sum_j \sum_k w_{ijk}\, y_{ijk} / \sum_j \sum_k w_{ijk}$$

where

w_{ijk} = weight for the kth respondent, in subgroup j, in state i

2. **Synthetic Estimator.** This is equation (4). Sixteen subgroups were used, 4 age categories (0-19, 20-44, 45-64, 65 or older), 2 race categories (black, nonblack), and 2 sex categories (male, female).
The predictive version of the synthetic estimator was also examined (equation 3), but resulted in insignificant differences between the two forms of the synthetic estimator. Thus, we only report the results from the more common version given above.

3. **Composite Estimator.** This is a linear combination of the first two estimators where the weights are proportional to the mean square errors of the two estimators.

$$\hat{\bar{y}}_{3i.} = t_i\, \bar{y}_{1i.} + (1-t_i)\, \hat{\bar{y}}_{2i.}$$

where

$$t_i = \frac{MSE(\hat{\bar{y}}_{2i.})}{Var(\bar{y}_{1i.}) + MSE(\hat{\bar{y}}_{2i.})}$$

$Var(\bar{y}_{1i.})$ can be estimated directly from the survey data. The calculation of $MSE(\hat{\bar{y}}_{2i.})$ is described in Section 3.2.

4. **GSE With Heteroscedastic Variances.** As mentioned previously, synthetic estimates of proportions where the subgroup means are assumed to be equal across small areas will automatically have homoscedastic variances. For continuous variables it is quite possible that variances will be heteroscedastic. Therefore, this estimator only applies to the doctor visits' variable.

Assume the mean response for a given subgroup is the same for all small areas, i.e., $V(\mu_{ij} \mid m_j) = v_j = 0$. Also, assume the prior distribution for m_j is diffuse (at least relative to S_{ij}) so that

$V(m_j) = v^*_j = \infty$ then

$$E(\bar{Y}_{i.} \mid (s, y)) = \left[n_{i.}\, \bar{y}_{i..} + \sum_{j=1}^{J} (N_{ij} - n_{ij})\, \hat{\mu}_{.j} \right] / N_{i.}$$

For the synthetic estimator we set $\hat{\mu}_{.j} = \bar{y}_{.j}$, but this is only appropriate when the expected elemental variances, $E(\sigma^2_{ij}) = \phi_j$, are equal for all areas. To test the synthetic estimator's robustness to the assumption of homoscedastic variances, we will make a relatively extreme, but plausible, assumption regarding the variances.

Assume that within a subgroup, there are differences in the distributions of doctor visits between those in central cities and those who live elsewhere. In particular, assume the mean and variance of each small area/subgroup is (μ_{ij}, $\sigma^2_{ij}\, z_{ij}\,(1 - z_{ij})$), with the pair ($\mu_{ij}$, σ^2_{ij}) exchangeable, as before.

z_{ij} = the proportion of the state's population living in central cities (capped at .95).

Thus,

$$\hat{\bar{y}}_{4i.} = (n_{i.}\, \bar{y}_{i.} + \sum_{j=1}^{J} (N_{ij} - n_{ij})\, \hat{\mu}_{.j}) / N_{i.}$$

where

$$\hat{\mu}_{.j} = \sum_{i=1}^{I} \frac{n_{ij}\, \bar{y}_{ij}}{z_{ij}\,(1 - z_{ij})} \Big/ \sum_{i=1}^{I} \frac{n_{ij}}{z_{ij}\,(1 - z_{ij})}$$

5. **GSE With Balanced Prior and Sampling Variances (16 subgroups).** If our prior distribution on the subgroup mean is diffuse (relative to the sampling variance), the GSE has been shown to simplify to the synthetic estimator. If the prior distribution is believed to have very little variation, we would use the *a priori* mean and not bother to collect the sample data. This estimator explores the GSE when the values of the prior and sampling variances are both moderate. Bounds on the

relationship between these two variances were specified in Assumption 2.

$$0 \quad \leq \quad \sum_{i=1}^{I} \frac{P_j}{P_j + S_{ij}} \leq \frac{N_j^2}{\sum_{i=1}^{I} N_{ij}^2}$$

When the sampling variance is large, the central term approaches 0, but when the prior variance dominates it approaches the term on the right. This RHS describes how evenly subgroup j is spread across all of the small areas. If the subgroup is only present in a single small area, then the RHS is equal to 1. If the subgroup is evenly spread across all small areas, then it is equal to the number of small areas. Thus, for the NHIS this term can vary from 1 to 51 (50 states plus the District of Columbia).

For this estimator, we balance the two variances by choosing the midpoint in the allowable range, that is, for each subgroup j:

$$\sum_{i=1}^{I} \frac{P_j}{P_j + S_{ij}} = \frac{1}{2}\left(\frac{N_j^2}{\sum_{i=1}^{I} N_{ij}^2}\right) = \frac{1}{2} b_j \qquad (5)$$

Also, we have already examined $c_j = \text{cov}(\mu_{ij},\mu_{rj})$ when it is large and positive (estimators 2 and 4), so for this estimator we let μ_{ij} be a random sample with expectation $\mu_{.j}$, thus

$$c_j \quad = \quad -\left(\frac{v_j + v_j^*}{I - 1}\right)$$

This implies that

$$P_j \quad = \quad v_j + v_j^* - c_j = \frac{I}{I - 1}(v_j + v_j^*) \qquad (6)$$

Substituting (5) and (6) into the GSE (1) provides the following estimator

$$\hat{\bar{y}}_{5i} \quad = \quad \left[n_i \bar{y}_i + \sum_{j=1}^{J} (N_{ij} - n_{ij}) \left[\frac{\hat{\mu}_j\left(\frac{51}{50b_j}\frac{1}{50}\right) + m_j^*\left(\frac{51}{50b_j}\right)}{\left(\frac{51}{25b_j}\frac{1}{50}\right)} \right] \right] / N_i.$$

For this empirical analysis we set

$\hat{\mu}_j$ = $\bar{y}_{.j}$ from the 1988 NHIS, and

m_j^* = $\bar{y}_{.j}$ from the 1987 NHIS.

With 16 subgroups the values for b_j vary from 19.85 to 23.51.

6. **GSE with Balanced Prior and Sampling Variances (32 Subgroups).** This estimator is of the same form as estimator 5. However, to examine the possibility of significant differences between members of a subgroup who live in a central city versus outside the central city, we use 32 subgroups. These are each of the 16 subgroups used previously, subdivided into central city and noncentral city. While the values of b_j used for estimator 5 only vary from 19.85 to 23.51, the b_j used for estimator 6 vary from 14.30 to 25.92.

Tables 1 and 2 provide the estimates derived for each state for each of the estimators for number of doctor visits and perceived poor health status, respectively. The states are sorted in descending order of the estimates from estimator 1. Note that for two of the states, North Dakota (ND) and Nebraska (NE), no sample was collected and thus no direct, design-unbiased estimate exists. The largest and smallest estimate is shaded for each estimator. While the model-based estimators are not consistent with the design-unbiased estimator, it is striking to note the similarities between the different model-based estimators. In particular for doctor visits, the standard synthetic estimator and all three generalized versions found Alaska (AK) to have the smallest estimated numbers. All but estimator 6 found Florida (FL) to have the largest estimated number (Florida was second largest for estimator 6). The different versions of the synthetic estimator had very little differences, generally less than 1 percent. This robustness of the synthetic estimator to heterogeneous variances and inclusion of data from the prior year is surprising and also comforting if this estimator were to be used to produce state-level estimates.

Similar results are observed for self-perceived poor health status. The three versions of the synthetic estimator each found Alaska to have the lowest incidence rate and the District of Columbia (DC) the highest. It is intriguing to note that these two states had direct estimates that were very similar. Again, the robustness of the synthetic estimator for the NHIS is a striking finding.

3.2 Accuracy of Estimators

In the previous section it was noted that the different model-based estimators produced similar point estimates. We now examine their accuracy; in particular, that of estimator 2, the synthetic estimator, and estimator 5, the GSE with 16 subgroups and relatively balanced contributions from prior and sample estimates. While the GSE was developed in a Bayesian framework, and the synthetic estimator is a model-based estimator, practitioners would like to measure their accuracy in a design-based manner. This would allow for comparisons with the variance of the design-unbiased direct estimator.

Gonzalez and Waksberg (1973) introduced the average (across small areas) mean square error as the measure of accuracy for model-based small-area estimates. The average MSE is computed as follows:

$$E(\bar{y}_{1i} - \hat{\bar{y}}_{fi})^2 \quad = \quad \text{aveMSE}(\bar{y}_1) + \text{aveMSE}(\hat{\bar{y}}_f)$$

$$\quad = \quad \text{aveVar}(\bar{y}_1) + \text{aveMSE}(\hat{\bar{y}}_f)$$

where

$\bar{y}_{1i}$ = design-unbiased inflation estimator for state i; and,

$\hat{\bar{y}}_{fi}$ = model-based estimator f for state i.

Thus,

$$\text{aveMSE}(\hat{\bar{y}}_{fi}) \quad = \quad E(\bar{y}_{1i} - \hat{\bar{y}}_{fi})^2 - \text{aveVar}(\bar{y}_{1i}) \qquad (7)$$

Table 1. Average number of doctor visits by state

State	Design-unbiased inflation estimator	Synthetic estimator	Composite estimator	Hetero-scedastic variances	GSE with (16) balanced variances	GSE with (32) balanced variances
VT	6.036	3.889	4.600	3.912	3.904	3.893
DE	5.648	3.856	4.178	3.898	3.895	3.873
CO	4.868	3.831	4.477	3.858	3.851	3.859
DC	4.706	3.852	4.121	3.967	3.987	4.084
MI	4.594	3.852	4.461	3.890	3.888	3.890
CT	4.589	3.926	4.343	3.957	3.948	3.956
AZ	4.562	3.889	4.350	3.915	3.906	3.927
NV	4.530	3.841	4.026	3.871	3.862	3.865
MA	4.481	3.933	4.362	3.960	3.951	3.959
RI	4.424	3.959	4.098	3.984	3.974	3.977
MT	4.394	3.911	4.159	3.933	3.925	3.920
KY	4.333	3.895	4.179	3.925	3.921	3.913
PA	4.204	3.967	4.175	4.000	3.993	3.994
OH	4.201	3.896	4.162	3.930	3.926	3.930
WY	4.156	3.827	3.912	3.849	3.844	3.842
CA	4.111	3.817	4.093	3.848	3.841	3.850
OK	4.104	3.898	4.041	3.927	3.923	3.928
NM	4.084	3.842	3.898	3.865	3.860	3.866
ME	4.075	3.928	3.981	3.949	3.941	3.933
MD	4.045	3.811	3.970	3.863	3.861	3.841
FL	3.981	4.003	3.985	4.041	4.030	4.025
TN	3.950	3.876	3.931	3.916	3.915	3.924
WA	3.869	3.871	3.870	3.896	3.888	3.890
AR	3.840	3.908	3.867	3.949	3.949	3.937
NY	3.828	3.902	3.836	3.943	3.937	3.958
IA	3.821	3.957	3.878	3.980	3.971	3.973
NJ	3.806	3.913	3.828	3.950	3.943	3.930
WV	3.734	3.975	3.888	3.999	3.992	3.981
OR	3.704	3.928	3.800	3.951	3.941	3.943
UT	3.704	3.740	3.721	3.761	3.761	3.760
WI	3.692	3.894	3.732	3.920	3.912	3.920
SC	3.677	3.780	3.730	3.838	3.843	3.801
VA	3.671	3.819	3.704	3.864	3.864	3.859
AL	3.638	3.846	3.698	3.899	3.903	3.893
HI	3.575	3.850	3.796	3.875	3.864	3.869
KS	3.573	3.897	3.690	3.924	3.917	3.922
MO	3.549	3.913	3.655	3.947	3.942	3.940
MS	3.489	3.763	3.674	3.829	3.840	3.793
TX	3.450	3.789	3.482	3.824	3.823	3.838
IL	3.420	3.862	3.483	3.901	3.897	3.908
LA	3.410	3.754	3.502	3.813	3.819	3.810
IN	3.380	3.885	3.516	3.915	3.910	3.920
SD	3.334	3.921	3.702	3.924	3.933	3.932
NH	3.236	3.879	3.798	3.901	3.893	3.893
ID	3.162	3.857	3.616	3.877	3.871	3.864
AK	3.093	3.629	3.542	3.656	3.653	3.665
MN	3.034	3.885	3.258	3.908	3.900	3.899
GA	3.000	3.755	3.190	3.809	3.812	3.787
NC	2.685	3.842	2.949	3.890	3.891	3.874
ND	NA	3.908	3.908	3.931	3.920	3.923
NE	NA	3.913	3.913	3.938	3.931	3.938

Table 2. Percent with perceived poor health status by state

State	Design-unbiased inflation estimator	Synthetic estimator	Composite estimator	GSE with (16) balanced variances	GSE with (32) balanced variances
WV	7.17	2.87	5.28	2.86	2.87
MS	6.28	3.19	5.28	3.18	3.23
AL	5.42	3.13	4.94	3.13	3.14
ME	5.19	2.60	4.01	2.59	2.59
AR	5.10	3.07	4.49	3.07	3.08
KY	5.06	2.70	4.56	2.69	2.70
NC	4.68	2.99	4.48	2.99	3.01
TN	4.46	2.91	4.23	2.91	2.90
OK	3.81	2.75	3.61	2.74	2.74
LA	3.52	2.96	3.44	2.96	2.97
GA	3.44	2.81	3.38	2.81	2.83
SC	3.42	3.03	3.36	3.03	3.07
IN	3.40	2.69	3.33	2.68	2.68
AZ	3.29	2.57	3.18	2.56	2.55
VA	3.29	2.80	3.25	2.80	2.80
TX	3.02	2.49	3.00	2.48	2.47
FL	2.96	3.29	2.97	3.30	3.30
DC	2.95	4.34	3.62	4.35	4.25
AK	2.87	1.76	2.13	1.74	1.74
ID	2.80	2.42	2.66	2.40	2.41
OH	2.79	2.81	2.79	2.80	2.80
MI	2.78	2.76	2.78	2.75	2.75
MO	2.64	2.88	2.66	2.87	2.87
DE	2.61	2.85	2.72	2.85	2.87
OR	2.50	2.65	2.53	2.64	2.64
WY	2.41	2.32	2.36	2.30	2.30
IL	2.40	2.82	2.42	2.82	2.81
CA	2.31	2.45	2.31	2.45	2.44
NY	2.31	2.96	2.33	2.96	2.94
KS	2.21	2.69	2.28	2.68	2.68
MA	2.21	2.71	2.24	2.70	2.69
MT	2.10	2.59	2.27	2.58	2.58
VT	2.09	2.45	2.24	2.44	2.44
NJ	2.03	2.94	2.07	2.94	2.95
SD	2.00	2.63	2.23	2.62	2.62
RI	1.97	2.78	2.19	2.77	2.77
MD	1.84	2.91	1.91	2.92	2.93
UT	1.75	1.99	1.79	1.97	1.97
NV	1.74	2.56	1.92	2.56	2.56
PA	1.69	3.02	1.72	3.01	3.01
NM	1.67	2.36	1.79	2.35	2.35
IA	1.65	2.77	1.76	2.76	2.76
WA	1.44	2.50	1.50	2.49	2.49
WI	1.38	2.64	1.45	2.63	2.62
CT	1.27	2.82	1.38	2.82	2.81
CO	1.20	2.36	1.27	2.35	2.35
MN	1.16	2.51	1.23	2.49	2.49
HI	0.87	2.44	1.08	2.44	2.43
NH	0.68	2.42	0.86	2.40	2.40
ND	NA	2.59	2.59	2.58	2.58
NE	NA	2.67	2.67	2.66	2.66

The average MSE is simple to compute and provides a good overall measure of accuracy. Unfortunately, by averaging across all states, it overstates the error associated with states where the model fits well or for which the sampling error is small (compared to an average state). Similarly, it understates the error associated with states where the model fails or for which the sampling error is large.

It would be far preferable to produce state-specific MSEs for the model-based estimators. This would provide smaller MSEs in states where the model fits well or for which the sampling error is small. It will also provide larger MSEs in states where the model fails or for which the sampling error is large. We therefore develop the following procedure for estimating state-specific MSEs.

$$\text{MSE}(\hat{\bar{y}}_{fi}) = \text{Var}(\hat{\bar{y}}_{fi}) + \text{Bias}^2(\hat{\bar{y}}_{fi}) \quad (8)$$

Unfortunately, the lack of an estimate of the "truth" for these biased estimators requires the use of an average bias, in conjunction with a state-specific variance.

$$\text{aveBias}^2(\hat{\bar{y}}_{fi}) = \text{aveMSE}(\hat{\bar{y}}_{fi}) - \text{aveVar}(\hat{\bar{y}}_{fi}) \quad (9)$$

Combining equations (7) and (9) gives

$$\text{aveBias}^2(\hat{\bar{y}}_{fi}) = E(\bar{y}_{1i} - \hat{\bar{y}}_{fi})^2 - \text{aveVar}(\bar{y}_{1i}) - \text{aveVar}(\hat{\bar{y}}_{fi}) \quad (10)$$

Using this average bias (10), we can produce state-specific mean square errors by replacing (8) with:

$$\text{MSE}(\hat{\bar{y}}_{fi}) = \text{Var}(\hat{\bar{y}}_{fi}) + \text{aveBias}^2(\hat{\bar{y}}_{fi}) \quad (11)$$

$\text{Var}(\hat{\bar{y}}_{fi})$ is calculated using replicated variances (jackknife or balanced repeated replication).

Table 3 shows the results of the six-step process for computing MSEs for estimators 2 and 5 for average number of doctor visits. Unfortunately, the average variance across states is quite small compared with the average bias (less than 2 percent). As a result, the largest and smallest state-specific MSEs are very similar to the average MSE. For the synthetic estimator, the aveMSE = .1703, while the smallest state-specific MSE is .1701 (Georgia) and the largest is .1732 (DC). This result is not very surprising since numerous examples have shown that the bias of the synthetic estimator is bigger than its variability. This is compounded by the fact that the sample size of the NHIS is so large (125,000 respondents) that the sampling variation is quite small. For most surveys, where the sample size is likely to be much smaller, the sampling variation can be expected to play a more major role, increasing the utility of state-specific MSEs.

3.3 State Groupings

Given the large sample size of the NHIS, it is possible to divide the nation into groups of states within which it is anticipated they would have similar biases with respect to the variable of interest. When this is true, the differences in MSEs between these groups can be large; and, for states in groupings where the model fits well, the average bias will be small and the state-specific MSEs will vary. The range of estimated mean values may also increase when the computations are based on data from a restricted subset of states.

Two potential groupings of states were examined. They were based on the percent of the nonblack population living in central cities, and the percent of the black population living in central cities. Figure 1 identifies

which states were classified into the high, medium, and low categories under each of these groupings.

Table 3. State-specific mean square errors for number of doctor visits using estimators 2 and 5

		Synthetic estimator	GSE with (16) balanced var.
1.	$E(\bar{y}_{1i\cdot} - \hat{\bar{y}}_{fi\cdot})^2$	.3777	.3754
2.	$aveVar(\bar{y}_1)$	.2074	.2074
	$aveMSE(\bar{y}_f)$	.1703	.1680
4.	$aveVar(\bar{y}_f)$	.0020	.0003
	$\frac{aveVar(\bar{y}_f)}{aveMSE(\bar{y}_f)}$	1.2%	0.1%
5.	$aveBias^2(\bar{y}_f)$	.1684	.1678
6.	$MSE(\bar{y}_f)$		
	smallest	.1701 (GA)	.1680 (many)
	largest	.1732 (DC)	.1684 (DC)
	$RMSE(\bar{y}_f)$	.412-.416	.410

Figure 1. Groupings of states by common expected biases

Percent of 1990 nonblack population in central city:

0-15%	DE, VT, SC, MS, WV, GA, NJ, MD, ID, ME, KY, MI
16-37%	all 34 others
38-100%	AK, NY, TX, AZ, DC

Percent of 1990 black population in central city:

0-35%	MS, SC, HI, VT, ND, ID, WV, ME, GA, DE, NC
36-77%	AR, MD, MT, UT, FL, NV, NM, NJ, SD, AL, KY, LA, VA, NH, WY, WA, CA, MO, OK, CO, AK, TX, RI, KS, OH, AZ, TN, MN, PA, CT, IA
78-100%	IL, MI, OR, MA, NY, NE, IN, WI, DC

3.3.1 Point Estimates Using Subnational State Groupings

Estimated number of doctor visits was derived for each state for each of the estimators, based only on data from within that state's groups of states, grouped by percent of black or nonblack population in central cities. Again the similarities were striking between the different model-based estimators. With states grouped by black percentage of the population in central cities, the standard synthetic estimator and all three generalized versions found Mississippi (MS) to have the smallest estimated number of doctor visits and the District of Columbia (DC) to have the largest estimated number. The four model-based estimators again agreed on the states with the largest and smallest number of doctor visits when states were grouped by percentage of nonblack population in central cities, but the extreme states were not the same as with the first grouping. Now West Virginia (WV) was always highest and Alaska (AK) lowest.

A major difference between these estimates and those in Table 1 is that the state-to-state range of model-based point estimates is much larger when the states are grouped than when all states used national subgroup estimates. Table 4 demonstrates that the range of model-based estimates across the states doubles when the states are grouped. For example when data from all states were combined, the GSE with 32 subgroups had a range of 0.41 (from 3.67 (AK) to 4.08 (DC)). When only data from within a state's grouping is used this range is increased to 0.76 for nonblack groupings (3.42 (AK) to 4.18 (WV)) and 0.97 for black groupings (3.48 (MS) to 4.45 (DC)). This is important since one of the concerns about the synthetic estimator is that many analysts believe it "over shrinks" the estimates towards the national average, underrepresenting the true variability from small area to small area.

Similar results are observed for self-perceived poor health status. With states grouped by nonblack populations in central cities (Table 5) the three versions of the synthetic estimator each found Alaska to have the lowest incidence rate and the District of Columbia the highest. It is again intriguing to note that these two states had direct estimates that were very similar. When grouped by black populations in central cities Alaska still has the lowest rate, but Mississipi now has the highest rate.

Table 6 examines the robustness of the synthetic estimator by comparing the variation among the different model-based estimators for each state. Small variation (as was found when national subgroup estimates were used) would indicate that the synthetic estimator is robust to its assumptions that were loosened in the various forms of the GSE that were examined. For doctor visits the variation among the model-based estimators was similar when states were grouped by nonblack population as when all states were combined. When grouped by percentage of the black population in central cities there was significant variation among the model-based estimates for the eleven states with more rural black populations.

Table 4. Variation in state estimates - average number of doctor visits

	Design unbiased	Synthetic	Composite	Heteroscedastic variances	GSE with 16 groupings	GSE with 32 groupings
All states	2.69-6.04	3.63-4.00	2.95-4.60	3.66-4.04	3.65-4.03	3.67-4.08
Percent nonblack	2.69-6.04	3.46-4.13	3.09-5.24	3.50-4.16	3.42-4.18	3.42-4.18
Percent black	2.69-6.04	3.14-4.23	2.72-5.37	3.34-4.28	3.48-4.44	3.48-4.45

Table 5. Percent with self-perceived poor health status - States grouped by percent of nonblack population in central cities

State	Design-unbiased inflation estimator	Synthetic estimator	Composite estimator	GSE with (16) balanced variances	GSE with (32) balanced variances
WV	7.17	3.34	6.02	3.44	3.43
MS	6.28	3.37	5.67	3.56	3.56
AL	5.42	3.10	4.71	3.11	3.13
ME	5.19	3.06	4.52	3.13	3.13
AR	5.10	3.00	4.23	3.02	3.04
KY	5.06	3.10	4.81	3.19	3.19
NC	4.68	2.95	4.36	2.97	3.00
TN	4.46	2.84	4.09	2.86	2.85
OK	3.81	2.65	3.48	2.67	2.67
LA	3.52	2.95	3.40	2.97	2.98
GA	3.44	3.04	3.42	3.18	3.17
SC	3.42	3.25	3.40	3.40	3.41
IN	3.40	2.59	3.27	2.62	2.61
AZ	3.29	2.72	2.81	2.47	2.45
VA	3.29	2.75	3.21	2.77	2.78
TX	3.02	2.59	2.78	2.34	2.34
FL	2.96	3.19	2.98	3.23	3.24
DC	2.95	4.07	3.99	3.82	3.81
AK	2.87	1.88	1.91	1.68	1.68
ID	2.80	2.83	2.81	2.90	2.90
OH	2.79	2.72	2.78	2.75	2.74
MI	2.78	3.09	2.79	3.20	3.20
MO	2.64	2.79	2.66	2.82	2.82
DE	2.61	3.18	2.78	3.30	3.29
OR	2.50	2.53	2.51	2.56	2.56
WY	2.41	2.20	2.28	2.22	2.22
IL	2.40	2.75	2.42	2.78	2.77
CA	2.31	2.36	2.31	2.39	2.39
NY	2.31	3.03	2.63	2.77	2.78
KS	2.21	2.59	2.29	2.61	2.61
MA	2.21	2.60	2.25	2.62	2.62
MT	2.10	2.46	2.27	2.49	2.49
VT	2.09	2.89	2.32	2.94	2.93
NJ	2.03	3.31	2.07	3.43	3.43
SD	2.00	2.51	2.25	2.53	2.54
RI	1.97	2.67	2.24	2.69	2.69
MD	1.84	3.16	1.89	3.30	3.30
UT	1.75	1.89	1.78	1.91	1.91
NV	1.74	2.46	1.97	2.49	2.49
PA	1.69	2.92	1.74	2.95	2.94
NM	1.67	2.25	1.82	2.28	2.28
IA	1.65	2.65	1.81	2.67	2.67
WA	1.44	2.38	1.52	2.41	2.41
WI	1.38	2.52	1.48	2.55	2.55
CT	1.27	2.72	1.43	2.75	2.74
CO	1.20	2.26	1.31	2.28	2.28
MN	1.16	2.39	1.26	2.41	2.41
HI	0.87	2.32	1.16	2.36	2.36
NH	0.68	2.30	0.95	2.32	2.32
NE	NA	2.56	2.56	2.58	2.58
ND	NA	2.47	2.47	2.50	2.49

Grouping the states also increased the variation among model-based estimates for percent with poor health in each state. This variation was particularly pronounced for states will large percentages of its nonblack population, or small percentages of its black population, in central cities.

Thus using subnational groupings of states resulted in model-based estimates that vary more from state-to-state and are, in general, still consistent across the form of estimator that is used. There are some states, however, for which the form of the estimator can have a significant impact on the estimate. We now examine the impact on mean square errors of using the subnational groupings.

Table 6. Robustness of the synthetic estimator

Variable	State grouping	Range among estimators (2), (4), (5), and (6)	
Doctor visits	None	6%	DC
		0-2%	All others
	Non-black	6%	DC
		3%	AZ
		0-2%	All others
	Black	10-11%	All 11 states with <36%
		4%	DC
		0-2%	All others
Poor health	None	0-2%	All states
	Non-black	7-12%	All 5 states with >37%
		3-6%	9 of 12 states with <16%
		0-2%	All others
	Black	10-12%	All 11 states with <36%
		3-4%	8 of 9 states with >77%
		0-2%	All others

3.3.2 Mean Square Errors Using Subnational State Groupings

The six-step process used in section 3.2 for computing mean square errors was recalculated averaging only across those states in the same grouping. The subgroup means were calculated separately for each grouping. The following tables are demonstrative of results for the two different groupings of states examined for each of the two variables.

Table 7 demonstrates that grouping states with similar expected biases can result in state-specific mean square errors where the variance is a major component of the MSE. For states with a high percentage of their nonblack population residing in central cities the variance of the synthetic estimator is on average 21 percent of the mean square error. When all states were taken together the RMSE varied only from .412 to .416. When states are grouped by the distribution of nonblack population we have RMSEs that vary from .190 in Utah to .717 in West Virginia, reflecting the fact that the model-based estimators are able to produce estimates for some states with much greater accuracy than for others.

Table 8 shows more promising results. Using the synthetic estimator to estimate percent with perceived poor health without subgrouping the states, the variance of the model-based estimator is only 0.4% of its MSE and the RMSE varies only from 1.16 to 1.17 in any state. However, when the states are grouped according to the percent of the nonblack population in central cities, the state-specific MSEs are quite different. For states with a high percentage in central cities, the average variance is 78 percent of the MSE! Among these states the RMSEs vary from 0.179 in Alaska to 0.311 in the District of Columbia; thus a confidence interval on the estimate for DC would be almost twice as wide as one for Alaska. The state-specific RMSE for West Virginia is 1.57, almost nine times the RMSE for Alaska.

Table 7. State-specific mean square errors for number of doctor visits using the synthetic estimator, with states grouped by percent nonblack population in central cities

		All states	Nonblack Population in Central Cities		
			Low	Moderate	High
(1)	$E(\bar{y}_{1i.} - \hat{\bar{y}}_{2i.})^2$	.3777	.7536	.2203	.3238
(2)	$aveVar(\bar{y}_1)$	.2074	.2417	.1842	.2736
	$aveMSE(\bar{y}_2)$	.1703	.5119	.0361	.0502
(4)	$aveVar(\bar{y}_2)$	.0020	.0119	.0029	.0106
	$\frac{aveVar}{aveMSE}(\bar{y}_2)$	1.2%	2.3%	8.0%	21.1%
(5)	$aveBias^2(\bar{y}_2)$	.1684	.5001	.0333	.0396
(6)	$MSE(\bar{y}_2)$				
	smallest	(GA) .1701		(UT) .0360	(TX) .0474
	largest	(DC) .1732	(WV) .5141		(DC) .0586
	$RMSE(\bar{y}_2)$	.412		.190	.218
		.416	.717		.242

Table 8. State-specific mean square errors for percent with perceived poor health using the synthetic estimator, with states grouped by percent of nonblack population in central cities

		All states	Nonblack Population In Central Cities		
			Low	Moderate	High
(1)	$E(\bar{y}_{1i.} - \hat{\bar{y}}_{2i.})^2$	1.724	3.011	1.080	.0651
(2)	$aveVar(\bar{y}_1)$	0.369	0.552	0.265	0.600
	$aveMSE(\bar{y}_2)$	1.355	2.459	0.815	0.051
(4)	$aveVar(\bar{y}_2)$	.0005	.0055	.0007	.0040
	$\frac{aveVar}{aveMSE}(\bar{y}_2)$	0.4%	2.2%	0.9%	78.4%
(5)	$aveBias^2(\bar{y}_2)$	1.350	2.404	0.809	0.011
(6)	$MSE(\bar{y}_2)$				
	smallest	(AK) 1.353			(AK) 0.032
	largest	(DC) 1.374	(WV) 2.475		(DC) 0.097
	$RMSE(\bar{y}_2)$	1.16			0.179
		1.17	1.57		0.311

all values are times .0001, except RMSE which is times .01

4. Conclusions and Recommendations

This paper has examined a series of issues involved in developing small-area estimates for the National Health Interview Survey. When designing the NHIS, or any other national survey from which it is hoped to develop small-area statistics, it is very important to stratify the sample in accordance with the small-area boundaries that are to be analyzed. This will improve the ability to produce design-unbiased estimates for those areas with large enough sample sizes. It will therefore improve composite estimators. Finally, it will improve the estimates of MSE for model-based estimators that are based on these data.

A Generalized Synthetic Estimator has been introduced that allows prior information to be included as well as the use of heterogeneous variances across small areas. It also allows one to examine the robustness of the classical synthetic estimator to several assumptions that determine when it is an appropriate estimator. In the case of the NHIS, the classical synthetic estimator appears to be quite robust to the assumption of homogeneity and the use of data from a prior year's survey.

Finally, a procedure is demonstrated for developing state-specific mean square errors for the model-based estimators. Variances for the model-based estimators are computed using replicated variances. An average bias is then calculated across all states. If the sample size is sufficiently large, improvements on the procedure can be made by grouping states together that have common expected biases. Applying these state-specific MSEs to the NHIS found instances where the root mean square error for one state was nine times larger than for another state, indicating the greater level of accuracy in predicting the estimate for the first state.

5. Acknowledgments

The authors would like to acknowledge the guidance of Graham Kalton of Westat, and Bill Ericson and Dick Cornell of the University of Michigan. Assistance was also provided by John Edmonds and David Judkins of Westat, and Don Malec at the National Center for Health Statistics. Part of this work was completed under contract #200-89-7021 with the National Center for Health Statistics. The comments in this paper are the authors' and do not necessarily reflect the opinions of Westat or NCHS.

6. References

Ericson, W. "Subjective Bayesian Models in Sampling Finite Populations." Journal of the Royal Statistical Society B, 31 (1969) 195-233.

_____. "Bayesian Sampling Lecture Notes." Unpublished, University of Michigan 1981.

_____. "A Bayesian Approach to Regression Estimation in Finite Populations." University of Michigan Technical Report #120, July, 1983.

Gonzalez, M., and Waksberg, J. "Estimation of the Error of Synthetic Estimates." Paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria, August 18-25, 1973.

Marker, D. unpublished dissertation research.

National Center for Health Statistics. Synthetic State Estimates of Disability. P.H.S. Publication No. 1759. Washington, D.C.: Government Printing Office, 1968.

Schaible, W.L. "A Composite Estimator for Small Area Statistics." National Institute on Drug Abuse Research Monograph 24, 1978, 36-62.