COMPARATIVE EFFECTIVENESS OF ADMINISTRATIVE DATA AND COUNTY-LEVEL AGGREGATES IN NONRESPONSE ADJUSTMENTS FOR SURVEYS OF LOW-INCOME POPULATIONS

Frank Potter, Mourad Touzani, Ronghua Lu, Yuhong Zheng, and Zhanyun Zhao
Mathematica Policy Research Inc, Princeton NJ 08540

In the evaluation of social programs, low-income populations are surveyed to study the program effectiveness. In these surveys, limited data are available from administrative data to use in the nonresponse adjustments (generally, age, gender and sometimes race of the program participant). The data can be supplemented by using aggregated data at the county level or at the zip-code level (e.g., Area Resource File or data from the Census Bureau). These data provide contextual information on where the program participant resides, but are not directly related to the individual's propensity to respond. In addition, using these aggregated data can add to the costs of the nonresponse adjustment activities. Our paper will investigate the benefits of using the additional aggregated data for explaining the variation response by logistic response propensity models when the aggregated data are used, relative to models using only the available administrative data. We will also assess the explanatory ability of the supplemental data in relation to the sampling designs (clustered and unclustered designs) for three study populations of interest in a complex survey of children either enrolled or recently disenrolled from a State Children Health Insurance Program (SCHIP). Our study is based on more than 50 response logistic propensity models developed for a 10-state evaluation of SCHIP.

1. METHODS FOR NONRESPONSE ADJUSTMENT

Unit-level nonresponse exists in all statistical surveys. The basic approach is to adjust the weights for the respondents to compensate for the nonrespondents (Brick and Kalton 1996, Lessler and Kalsbeek 1992). The classical methods use a single global adjustment or adjustments within groups of sample members (the weighting class approach). More recently, logistic regression is used to model the response propensity to identify factors associated to nonresponse (Little 1986, Potter et al. 1998). Response propensity modeling can be viewed as the extension of the standard weighting class procedure; however allowing for the use of more factors (including both continuous and discrete factors) and complex interactions among factors to explain the differential propensity to be located or to respond. The factors for inclusion in the models can be identified using a bivariate cross-tabulations and multivariate procedures such as interaction detection procedures (for example CHAID, Chi-squared Automatic Interaction Detection software (Magidson 1993)). In general, these models are developed using information from the sampling frame or other sources (such as the Census Bureau) as covariates. Through a series of step-wise procedures and analyses, we can reduce the number of covariates and interactions among those covariates to a minimum number for the final model. The model is then evaluated using the R-square and other measures of goodness of model fit and the statistical significance of the coefficients of the covariates in the model. We then use the specific covariate values for each respondent to estimate a propensity to respond. The inverse of these propensity scores is used to adjust the weights from the prior step of weight computations[1].

As a last step in weighting, the nonresponse-adjusted weights are examined to look for excessively large weights produced as a result of the sampling or response adjustments. We determine whether to truncate weights and which ones to truncate based upon an investigation of the effect on the extreme weights on the variation in the weights and potential effect on key survey estimates. When weights are truncated, we apply a smoothing adjustment within weighting classes to recover the lost weight (Potter 1990).

2. SURVEY DESIGN

The survey used in our study was the part of the Congressional-mandated evaluation of the State Children Health Insurance Program (SCHIP) conducted by Mathematica Policy Research, Inc. For this evaluation, we sampled new and established SCHIP enrollees and recent disenrollees in 10 states (Touzani et al. 2002). For all sample members, we interviewed the parent, guardian or caregiver of the sampled child (that is the person most knowledgeable

[1] Propensity scores can also be used to develop weighting classes.

about the health care needs and services received for the sampled child). The sample for the SCHIP survey included new enrollees, established enrollees, and recent disenrollees and was designed so that study findings can be used to make inferences about these SCHIP enrollment domains for each of the 10 states participating in the SCHIP evaluation and to make comparisons across these states.

The state selection process flowed from three criteria specified in the legislation for the evaluation. These criteria specified that the 10 states were to: (1) include a significant portion of uninsured children; (2) use diverse programmatic approaches to providing child health assistance; and (3) represent various geographic areas. Guided by these selection criteria, we chose the following states to participate in the SCHIP evaluation: California, Colorado, Florida, Illinois, Louisiana, Missouri, New Jersey, New York, North Carolina, and Texas.

In considering options for this study, we evaluated telephone-only interviewing and face-to-face interviewing. The high costs and clustered nature of face-to-face interviews led to our adoption of a dual frame sample design. The dual-frame design combines an unclustered sample that is interviewed by telephone only (when a telephone number could be found using centralized locating efforts) with a clustered sample that is interviewed by telephone but has in-person field followup to locate the nontelephone households. The field locator provided the parent, guardian or caregiver with a cell phone for completing the interview by telephone. With this approach, we can achieve the greater precision associated with the unclustered design, while retaining the enhanced response and coverage rates of the face-to-face interview approach.

a. Target Population of Children

The target population for the SCHIP portion of the evaluation was restricted to children who, at the time of frame construction, were newly enrolled or established enrollees of SCHIP or who had recently disenrolled. The target population was further limited to children living in the 10 states at the time of data collection and to SCHIP enrollees age 18 and younger and recent SCHIP disenrollees age 19 and younger. The age limit of 19 years was set for disenrollees so that we could capture recently disenrolled children who aged out of SCHIP when they reached their 19th birthday.

Enrollment status was defined based on the enrollment status data recorded in the SCHIP databases delivered by each state. We developed explicit operational definitions for the three SCHIP enrollment domains based on the enrollment process used in the state (presumptive, retrospective, or other) and within the logistical constraints of the SCHIP enrollee databases from the 10 states. We adopted the following operational definitions of the enrollment domains for the SCHIP samples:

- *New enrollees*: children who were enrolled in the program at least one month, but less than three months at the time of frame construction[2]
- *Intermediate enrollees*: were children who were enrolled in the program for two or more than one month but less than five months at the time of frame construction
- *Established enrollees*: were children who were enrolled for five or more months in the program at the time of frame construction
- *Recent disenrollees*: children who were disenrolled from the program at the time of frame construction but who were enrolled in the preceding two months

The SCHIP surveys were limited to new enrollees, established enrollees, and recent disenrollees.[3] We refined the above definitions of the target population and enrollment domains after we received SCHIP data from each of the 10 states. We investigated the exact definition of "enrollment" for each state and on what date a child is to be considered as enrolled relative to the date when the parent and caregiver would know that the child was enrolled. That is, some states had retroacted enrollment (covering costs of medical services received prior to the application process) and we used information from the state files to determine when would the parent or caregiver know of coverage. In those states, we used date of application or date of authorization (that is the approximate date when the parent was notified) as the point when the parent knew of the coverage and would have started to consciously seek medical care for the child under

[2] The "time of frame construction" is defined as the most recent month for which the state provided SCHIP and Medicaid enrollment data.

[3] Intermediate enrollees were not included in the evaluation, as they would be too far away from their enrollment to recall their pre-enrollment experience but would not have been enrolled for sufficient time to acquire experience with the program.

SCHIP. The purpose of this was to classify the child enrollment status as the basis of the parent's knowledge that healthcare services would be covered, rather than for the period when health care costs were paid for by SCHIP. In these surveys, we only included children for whom the determination process has been completed and eligibility confirmed.

b. Sampling Frame

The sampling units for the surveys were the new and established enrollees and recent disenrollees of SCHIP (all 10 states). For this study, we used state SCHIP eligibility and enrollment files to construct the frames. To avoid burdening the respondent (who must be a person living with the child and most familiar with the child's health and health care), the sampling approach first selected households containing eligible children and then selected one eligible child for interview from each household.

c. Sampling Designs

Our design used exclusively telephone interviewing to avoid mode effects. First, we used central office locating efforts to find the telephone number for the sampled child's household and the household could be reached by telephone ("telephone households"). For households that could not be located by the central office locating, we used in-person locating of households by field locators with cellular telephones to ensure adequate coverage of sampled children living in nontelephone households or households that could not be reached by telephone ("nontelephone households"). For these nontelephone households, we then conducted the interview by telephone using the same computer-assisted telephone interviewing system that was also used for the telephone households.

We adopted the following operational definitions for this study:

- *Telephone households*: households with telephone service for which telephone numbers can be located.
- *Nontelephone households*: (1) those households without telephone service, and (2) those households for which a telephone number cannot be located.[4]

These definitions reflect the practical constraint—only those sampled children residing in households with telephone service for which telephone numbers can be located (that is, those that meet the operational definition of "telephone household") can have their parents interviewed by telephone.

For this evaluation, we considered designs suitable for telephone versus face-to-face data-collection modes in terms of their ability to provide maximum efficiency and coverage for minimum costs. We faced two design challenges. First, the broad geographic representation of the unclustered telephone design was considered desirable, but so was the improved population coverage of the face-to-face data collection using a clustered design. Second, the per-interview costs of face-to-face data collection required that we limit the number of households we attempted to contact in-person. To have the benefits of in-person contact and in-field locating and yet avoiding mixing data collection modes, we used in-person field locating in the clustered sample with all interviews conducted by telephone. That is we addressed these challenges by incorporating both sampling approaches in the final design.

The design was a variation of the classic subsampling-for-nonresponse-followup design. In each state (except New Jersey)[5], two independent samples were selected for the SCHIP survey and for the Medicaid survey—one clustered and one unclustered. All telephone households were interviewed in both samples. All (or a sample of) the nontelephone households were also interviewed by telephone in the clustered sample. That is, across both designs, households were interviewed by telephone only. This restriction is necessary for the integration of the two samples; it also reduces mode effects across samples, because telephone households were always interviewed by telephone, regardless of the design for which they were sampled. The clustered design used field staff to locate the

[4] The latter group includes households with unlisted numbers who do not have their current number recorded in the SCHIP or Medicaid database.

[5] For New Jersey, we used only an unclustered design because the state is sufficiently geographically small that the use of a clustered sample was deemed unnecessary.

household and then provide a cell phone for use by the nontelephone household for the interview.

Each design used multiple stages of selection and a composite size measure was used to maintain nearly equal selection rates within each domain (Folsom, Potter, and Williams 1987). A sample design was replicated for up to three different sample rounds and fielded in each state. Each sample round was composed of sampled children from each SCHIP enrollment domain. For states with smaller enrollment populations, the multiple rounds were needed to ensure that sufficient sample sizes of new enrollees and recent disenrollees were obtained from each program. The sample for the last round for each state included a reserve sample, from which additional sample cases were released for data collection if response or eligibility rates were different from expectations.

Because of the enrollment population sizes for California and Texas, the full sample was selected from the March 2002 enrollment files. For six states (Florida, Illinois, Missouri, New Jersey, New York and North Carolina), two sample rounds were used which were based on the January and March enrollment files. The samples for Colorado and Louisiana, which had the smallest enrollment populations, were selected using three sample rounds (using January, March, and May enrollment files). For each design, we selected three separate samples of enrollees and recent disenrollees. These sample rounds were drawn to avoid sampling multiple children from the same household or sampling households for more than one sample round. Each sample draw was derived from the universe existing at the time of sampling, but taking into account whether a household was in the sampling frame, or the sample, of the prior round(s).

We used the final definitions to select the SCHIP samples of new enrollees, established enrollees, and recent disenrollees for each of the three sample rounds. To define the enrollment domains for each sample round, we began by classifying persons into the three domains (new enrollees, established enrollees, and recent disenrollees) using the databases provided by the state. Note that the populations of established enrollees on the three sampling rounds overlapped extensively, but, by definition, new enrollees and recent disenrollees were unique to a specific sample round. Some changes also occurred in their enrollment status from one sampling round to another (for example established enrollees at one time could become recent disenrollees at the next time). At each sampling occasion, the sample was composed of a clustered sample and an unclustered sample of children in the SCHIP (except for New Jersey) enrollee domains. We used sampling procedures that prevent the selection of the same child or household at subsequent rounds, while preserving the probability structure for the two independent samples at each round.

In summary, the sampling design for SCHIP enrollees included 38 samples across the 10 states (18 clustered samples and 20 unclustered samples). For California and Texas, clustered and unclustered samples were selected using only the March 2002 extract and for Colorado and Louisiana, clustered and unclustered samples were selected using extracts for January, March and May of 2002. The samples for other 6 states used extracts from January and March. In total, 27,770 children were selected across the 10 states and an overall weighted response rate of 74 percent was achieved (see Table 1 for sample sizes and response rates for the individual domains and sample designs).

d. Sampling Weights

The sampling weights were developed for each sample member separately in each state for each round, sampling design, and sample domain using the selection probabilities for each stage of selection. The weights were then combined across round to make weights for sampling design and sample domain in each state. After the weights were combined across rounds, the weights were post-stratified to average monthly enrollment estimates for each domain. In total 57 sets of weights required nonresponse adjustment: 30 sets of weights for clustered samples (10 states and three domains for each state) and 27 sets of weights for unclustered samples (9 states and three domains for each state).

3. METHODS AND DATA FOR NONRESPONSE ADJUSTMENT

In all surveys, nonresponse occurs and the standard procedure is to adjust the sampling weights to compensate for this nonresponse and, thereby, minimize the potential for nonresponse bias. The weights for respondents who are similar to those who do not respond are adjusted to reduce the potential for this bias. We initially conducted an analysis to identify the factors that may be related to nonresponse. Because the extract files from the states contained limited data (age and sometimes race) for identifying similarities among respondents and nonrespondents, we accessed county-level data from the Area Resource File (ARF) to supplement the state-provided data. The ARF contains county-level counts and other data compiled from the Census Bureau, the Bureau of Economic Analysis, the US Department of Agriculture, the National Center for

Health Statistics and other sources. The data obtained from the ARF are given in Table 2. These variables were selected as measures of racial and ethnic composition and as measures related to the extent of poverty in the counties in which the sample member resided. As a rule, we perceive these variables are proxy measures for unobservable factors associated with response and that these variables themselves do not imply any direct relationship with response patterns.

For the response models, we used a series of summarizations and tabulations of these continuous variables to form categories based on the characteristics of each sample. We formed categories to ensure adequate sample counts in each category and that the categories were somewhat logical breaks in the distribution of the continuous variable. We then proceeded with step-wise logistic modeling to identify the categorized variables and the state-provided data on the child's age and race that best explained the response pattern found for each sample. We used both forward and backward stepwise logistic regression with normalized weights (and assuming simple random sampling) to identify main effects and second-order interactions as candidate covariates. We then used logistic regression in SUDAAN with the sampling design specified to assess statistical significance of the covariates and the interactions using the design-based variance estimates. Once a final model was identified, we computed measures of goodness of fit including the percentage of concordance and the Hosmer-Lemshow goodness of fit statistic.

Since the states and the enrollment population differed substantially, no single set of variables were consistently the best variables to explain the response pattern. However, response, in general, was associated with the degree of urbanicity with lower response in some urban areas and higher response in rural areas. Other factors that were found useful in the explaining the response pattern were ethnicity and race and the percentage of persons ages 0-17 in poverty.

To assess the effectiveness of using the additional data from the Area Resource File, we reran the nonresponse adjustment process for the 57 sets of weights using only data available from the sampling frames. For all states, the extract files from the state contained age of the child and county of residence. We used the county of residence to use the rural/urban continuum code for a county (see http://www.ers.usda.gov/Data/RuralUrbanContinuumCodes/ for additional information on these codes). We could classify a county as in a metropolitan statistical area (MSA) with a population of one million or more or a county in an MSA with 250,000 or fewer persons. On the other hand, race information was available for 8 of the states (the extract files from FL and NY did not include any race information). We used essentially identical procedures to develop the response propensity models. That is again we used both forward and backward stepwise logistic regression with normalized weights (and assuming simple random sampling) to identify main effects and second-order interactions as candidate covariates. We used logistic regression in SUDAAN to assess statistical significance of the covariates and the interactions. We then computed measures of goodness of fit including the percentage of concordance and the Hosmer-Lemshow goodness of fit statistic on the final reduced model.

4. ANALYSIS

For our analysis, we used the percentage of concordance (Table 3) and the R-square statistic (Table 4). The percent concordant was higher by 6 to 10 percent of the models using the frame data supplemented by the ARF data. This difference was statistically significant over all models, for models when the sample is based on the clustered sample design and for two of the study domains (new enrollees and established enrollees). While the R-square statistic will be increase purely by increasing the number of covariates in a model, we did find a comparable pattern using this measure (Table 4.). When looking at the differences for individual states (data not shown), only CA and FL had significantly higher percent concordance (note that FL did not provide race information) and CA, FL, LA and NY had significantly better R-square values. There was considerable variation across the study domains within a state. Once again, the models using only the frame data were based primarily on the age of the child, the metropolitan status of the county of residence and, for 8 of the 10 states, the race of the child both as main effects and as interactive variables.

It should be noted that no separate adjustment was made for the inability to locate a person so response is inclusive of both the ability to locate a households or a person and after locating the person, the ability to get a completed interview. For the recent disenrollees, a substantial proportion of the sample was disenrollees because of aging out of the program (that is they were 19 or older) and locating these persons was more difficult. Age was therefore a very significant predictor of the propensity to response. We had expected that the ARF data would have substantially more impact in the unclustered samples since the clustered samples limited to 30

primary sampling units for all but CA (60 PSUs were used in CA).

5. CONCLUSION

In evaluations of social programs, surveys of low-income populations are an integral part of the evaluation and rely on administrative data as sampling frames. These administrative data files generally have limited data to use in sample selection or in the nonresponse adjustments (generally, age, gender and sometimes race of the program participant). For nonresponse adjustments, the data can be supplemented by using aggregated data at the county level or at the zip-code level (e.g., Area Resource File or data from the Census Bureau). These data provide contextual information on where the program participant resides, but are not directly related to the individual's propensity to respond.

In the Congressionally-mandated evaluation of the State Children Health Insurance Program (SCHIP), we used administrative files to select the samples of children. For data collection efficiency and to minimize nonresponse, we used both unclustered and cluster sampling designs with multiple stages of selection. The complexity of the evaluation design and the study populations resulted in 57 sets of weights for nonresponse adjustment. Because of the lack of information from the administrative files, we accessed additional data from ARF and used these data in the response propensity modeling. The use of the ARF data required additional time and resources and we sought to assess the benefits of using the ARF data.

We found that the additional information generally improved the goodness of fit of the response propensity models, but the results were not always consistent nor matching our expectations. We found more improvement in the models for weights that were based on the clustered sample designs (we expected models for weights from the unclustered sample designs to have the greater benefit since these samples were widely distributed across the counties in a state). For the study domains of new enrollees and established enrollees, the additional ARF data provided significant improvement in the models (measured by the percent concordance). For the disenrollees, the ARF data generally did not help; however, as noted previously, the age of these sample members was a highly significant factor in all of the response propensity models.

The conclusion is that the ARF data can provide useful information when limited data are available from the sampling frame for nonresponse adjustment. However, these county-level data cannot be a substitute for additional person-level data. The characteristics of the sample population have a substantial effect on the extent of usefulness of these additional county-level data.

REFERENCES

Brick, J.M. and Kalton, G. "Handling missing data in survey research." *Statistical Methods in Medical Research*, 5, 215–238 (1996)

Folsom, R.E., F. Potter, S.R. Williams. "Notes on a Composite Measure for Self-Weighting Samples in Multiple Domains." 1987 American Statistical Association *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 1987, pp. 792-796.

Lessler, J.T. and Kalsbeek, W.D. (1992). "Nonsampling Error in Surveys." New York: John Wiley.

Little, R.J.A. (1986) "Survey nonresponse adjustments for estimates of means." *International Statistical Review*, **54**, 139–157.

Magidson J. SPSS for Windows: CHAID, release 6.0. Belmont MA: Statistical Innovations, Inc., 1993.

Potter, F.J. "A Study of Procedures to Identify and Trim Extreme Sampling Weights." *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 1990, pp. 225-230

Potter, F.J., Iannacchione, V.G., Mosher, W.D., Mason, R.E., and Kavee, J.D. "Sample design, sampling weights, imputation, and variance estimation in the 1995 National Survey of Family Growth." Vital and Health Statistics. Series 2, Data Evaluation and methods research; no. 124 (1998).

Touzani, M, Potter, F.J. and Zambrowski A. "Sampling Children for a Study of the State Children's Health Insurance Program." 2002 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM]. Alexandria, VA: American Statistical Association (2002).

TABLE 1. DESIGN-SPECIFIC SAMPLES AND WEIGHTED RESPONSE RATES

	All Samples			Unclustered Sample Design			Clustered Sample Design		
Domain	State Samples	Sampled Children	Response Rate	State Samples	Sampled Children	Response Rate	State Samples	Sampled Children	Response Rate
All Domains	57	27,770	74.0%	30	16,100	76.9%	27	11,660	70.8%
New Enrollee	19	8,966	75.8%	10	5,158	78.6%	9	3,808	72.6%
Established Enrollee	19	8,850	76.8%	10	5,086	79.8%	9	3,764	73.5%
Recent Disenrollee	19	9,954	69.5%	10	5,866	72.4%	9	4,088	66.3%

TABLE 2. DATA FROM AREA RESOURCE FILE (ARF) USED IN MODELING

1. Rural/urban continuum code (10 levels)
2. Population percentage for white, black/African American, Asian, and other
3. Percent Hispanic or Latino population
4. Percent of persons 25 or older with less than 9 years of school
5. Percent of persons 25 or older with some high school diploma or more
6. Percent of persons 25 or older with 4 or more years of college
7. Median family income
8. Median household income
9. Percent of families below poverty level
10. Percent of persons below poverty level
11. Percent of families with female head
12. Percent of persons in poverty
13. Percent of persons ages 0-17 in poverty
14. Percent of related children ages 5-17 in poverty.

TABLE 3. LOGISTIC REGRESSION GOODNESS OF FIT: CONCORDANCE

	Samples	Frame Data + ARF Data	Frame Data Only	Difference
Total	57	60.8%	54.7%	6.1%**
Sampling Design				
Clustered	27	60.6%	52.7%	7.9%**
Unclustered	30	61.0%	56.5%	4.5%
Domains				
New Enrollee	19	61.8%	55.9%	6.0%*
Established Enrollee	19	60.9%	50.8%	10.1%**
Recent Disenrollee	19	59.7%	57.4%	2.2%

* Significantly different from zero at the 0.05 level, two-tailed test
** Significantly different from zero at the 0.01 level, two-tailed test.

TABLE 4. LOGISTIC REGRESSION: R-SQUARE

	Samples	Frame Data + ARF Data	Frame Data Only	Relative Difference
Total	57	0.073	0.053	-14.8%**
Sampling Design				
Clustered	27	0.078	0.056	-18.0%**
Unclustered	30	0.068	0.051	-11.9%*
Domain				
New Enrollee	19	0.072	0.047	-29.7%**
Established Enrollee	19	0.073	0.049	-24.9%**
Recent Disenrollee	19	0.073	0.063	10.1%

Relative difference = 100 * [(Frame Data Only) – (Frame + ARF Data)] / (Frame + ARF Data).

* Significantly different from zero at the 0.05 level, two-tailed test
** Significantly different from zero at the 0.01 level, two-tailed test.