NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY LIMITING THE RISK OF DATA DISCLOSURE USING REPLICATION TECHNIQUES IN VARIANCE ESTIMATION

Sylvia Dohrmann, Westat; Lexter R. Curtin, National Center for Health Statistics;
Leyla Mohadjer, Jill Montaquila and Thanh Lê, Westat
Sylvia Dohrmann, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Disclosure Limitation, Replication Techniques, and Jackknife Replicates

1. Introduction

The National Health and Nutrition Examination Surveys (NHANES) are one of the series of health-related programs conducted by the National Center for Health Statistics (NCHS). A unique feature of these surveys is the collection of health data by means of medical examinations carried out for a nationally representative sample of the U.S. population. Beginning in 1999, NHANES is being implemented as a continuous, annual survey.

The counties in the primary sampling units (PSUs) from two panels of the 1995 NHIS were used as the sampling frame for NHANES 1999-2001. Each single year and any combination of consecutive years comprise a nationally representative sample of the U.S. population. This design will facilitate potential linkage to other health and nutrition surveys that provide yearly estimates and will allow aggregate-level national estimates from NHANES each year.

A four-stage sample is selected for NHANES. Within each of the selected PSUs, an average of 24 segments are selected. A subsample of the households within those segments are selected and screened. Within the screened households, members of particular race/ethnicity-sex-age subdomains are identified as potential sampled persons; all other members of the household are excluded.

NHANES 1999-2000 was fielded in 27 stands comprising 26 PSUs (one certainty PSU was visited twice), which were primarily single counties. Because no explicit stratification was used to select the PSUs from the two panels of NHIS and because of the small number of PSUs in the sample, the jackknife method (JK1) was used to create replicates for variance estimation for the analysis of the NHANES 1999-2000 data.[1] With the JK1 method, one unit (PSU) is dropped at a time in forming each replicate. Thus, groups of sampled units selected from the same PSU can be readily identified.

The risk of PSU identification coupled with the fact that the data files contain some geographic data and other characteristics of the area led to concerns about disclosure risks in the release of the NHANES 1999-2000 data file. As a result, NCHS initiated research to examine the disclosure risks of NHANES. NCHS reviewed both univariate and multivariate distributions of variables to identify problems (e.g., outliers, unique combinations of variables that could lead to disclosure, etc.). At the same time, NCHS requested research on alternative approaches for creating variance estimation replicates to mask the PSUs. This document describes the methods and results for the latter.

The research consisted of developing alternative replication approaches (Section 2), assessing their disclosure limitations (Section 3), and evaluating their performance (Section 4). Section 5 contains the recommended variance estimation approach for 1999-2000 NHANES release.

2. Creation of Replicates

As noted above, the original set of replicates (hereafter referred to as the "baseline") for NHANES 1999-2000 was created using the standard JK1 method; for noncertainty PSUs, the PSU is the variance unit, and for the certainty PSU, two variance units were formed by alternating segments. Figure 1 depicts the creation of replicates in the baseline design. The shaded area denotes that in creating the given replicate, the particular PSU or the particular segment was dropped.

Certainty status	PSU	Replicate				
		1	2	…	26	27
Noncertainty PSUs	A					
	B					
	…					
Certainty PSUs	Z1, Z2 1st seg					
	Z1, Z2 2nd seg					
	Z1, Z2 3rd seg					
	Z1, Z2 4th seg					
	…					

Figure 1. Baseline replication design

The first alternative (hereafter referred to as the "split PSU" alternative) creates replicates by alternating segments within each noncertainty PSU, as was done for the certainty PSU in the baseline replication design. The order of the replicates is then scrambled to further ensure confidentiality. This approach attempts to preserve as much of the design as possible while taking one step toward preserving confidentiality. That is, since the sampled units in a given PSU are split between two variance units, there is no single, easily identifiable grouping of units in a given PSU. Figure 2 depicts the split PSU replication design.

[1] With so few PSUs and the design used to select the NHANES PSUs, it was concluded that the effect of subsampling from the NHIS PSUs is minor when compared to the magnitude of variances of the NHANES statistics.

Certainty status	PSU	Replicate				
		1	2	…	51	52
Noncertainty PSUs	A 1st seg					
	A 2nd seg					
	A 3rd seg					
	A 4th seg					
	…					
Certainty PSUs	Z1, Z2 1st seg					
	Z1, Z2 2nd seg					
	Z1, Z2 3rd seg					
	Z1, Z2 4th seg					
	…					

Figure 2. Split PSU replication design

The second alternative (hereafter referred to as the "clustered-split PSU" alternative) for splitting noncertainty PSUs groups the first half of the segments in the original order of selection (for example, segments 1 to 12) into one replicate and the second half (for example, segments 13 to 24) into another replicate, rather than alternating the segments. One might expect that this approach would result in clustering that would resemble too closely the PSUs, and thus not serve our main objective of maintaining confidentiality. However, since the segments are selected in increasing order of minority density, the resulting replicates formed from this method do not have the same characteristics as the full PSU. In addition, the order of the replicates is then scrambled to further ensure confidentiality. Since the segments are assigned alternatively to replicates, the replicates are quite likely to have the same characteristics of the full PSU. The same may not be said for the aforementioned "split PSU" design. Figure 3 depicts the clustered-split PSU replication design.

Certainty status	PSU	Replicate				
		1	2	…	51	52
Noncertainty PSUs	A 1st seg					
	A 2nd seg					
	…					
	A 12th seg					
	A 13th seg					
	A 14th seg					
	…					
	A 24th seg					
	…					
Certainty PSUs	Z1, Z2 1st seg					
	Z1, Z2 2nd seg					
	Z1, Z2 3rd seg					
	Z1, Z2 4th seg					
	…					

Figure 3. Clustered-split PSU replication design

The third alternative (hereafter referred to as the "scrambled PSU" alternative) assigns segments to replicates with little regard to PSUs. The noncertainty PSUs are randomly sorted, then the segments are assigned sequentially to 48 replicates. This approach attempts to ensure confidentiality at the possible expense of properly reflecting the effects of PSU-level clustering on the variances of survey estimates. Figure 4 depicts the scrambled PSU replication design.

Certainty status	PSU	Replicate						
		1	2	3	4	…	49	50
Noncertainty PSUs	1st stand 1st seg							
	1st stand 2nd seg							
	1st stand 3rd seg							
	1st stand 4th seg							
	…							
Certainty PSUs	Z1, Z2 1st seg							
	Z1, Z2 2nd seg							
	Z1, Z2 3rd seg							
	Z1, Z2 4th seg							
	…							

Figure 4. Scrambled PSU replication design

3. Disclosure Limitation

As mentioned in the introduction, the JK1 method of variance estimation drops one unit at a time from the sample to form replicates. The weights for the dropped units, PSUs in the case of NHANES, are set to zero making it quite simple to identify the records included in the unit. This poses a problem in terms of confidentiality. There is concern that the demographic characteristics of the people in these units would resemble those of the PSU so closely that the location of the PSU would be evident. While the possibility of identifying dropped units cannot be entirely eliminated, the alternative methods proposed in Section 2 attempt to mask the linkage of these records to the geography from which they were selected. This section contains an evaluation of the disclosure limitations of the alternative designs.

In order to determine the disclosure limitations of the baseline and alternative variance estimation designs, the demographics of the variance units (i.e., the records dropped to form each replicate) from each of the designs were compared to those of the respective PSUs. Recall the methods used to form variance units from the baseline and alternative designs described in Section 2:

- Baseline—Each noncertainty PSU is a variance unit; the certainty PSU forms two variance units by alternating segments;
- Split—Each PSU forms two variance units by alternating segments;
- Clustered-split—Each PSU forms two variance units by placing lower minority segments (roughly half the segments) into one variance unit, and the higher minority segments into the other variance unit; and
- Scrambled—Variance units are formed with no regard to PSUs.

Since the scrambled design forms variance units containing records from many PSUs, the characteristics

of the variance units' demographics are not expected to be associated with those of any single PSU. There is little chance that records could be traced back to their geography by means of examining replicates formed in this design. For the other designs, it is necessary to evaluate the possible disclosure limitations.

In order to evaluate the disclosure limitations of the baseline, split, and clustered-split designs, the demographics of the sampled persons falling into each variance unit were compared with PSU-level demographics (using data from the 2000 Census). Demographics were selected that could be calculated from the sample data, would vary among the PSUs, and could possibly aid in identifying the PSUs. The demographics, listed below, were all expressed as percentages so that comparisons could be made directly:

- Race/ethnicity including the percentage of the population that is black, Hispanic, Mexican American, and/or White nonHispanic;[2]
- Population greater than or equal to 65 years of age;
- Population less than 18 years of age;
- High school graduates;
- College graduates;
- Households containing people less than 18 years of age; and
- Households containing people greater than or equal to 65 years of age.

The results of the comparisons are shown in Figures 5. On the chart, there is a line drawn indicating the points at which the PSU and variance unit percentages would be the same. The circles indicate the demographics of each variance unit formed in the baseline design with those of the respective PSU. The "fan" pattern in this chart shows that while there is little difference in the rare characteristics (i.e., those with less than 10% prevalence), the larger the percentage the more the variance unit and the PSU differ. This difference is greater between the split design variance units and the PSUs diamond markers and most pronounced between the clustered-split design variance units and the PSUs triangle markers.

Note that there are points on the figure with very large percentages in the PSU and the variance units. (These points are in the top right-hand corners of the chart.) These are percentages of White nonHispanics in very low minority areas of the Midwest and Northeast. While the variance unit estimates are close to the PSU estimates, there are other variance unit estimates close to these same values which would prevent any positive identification of the PSU.

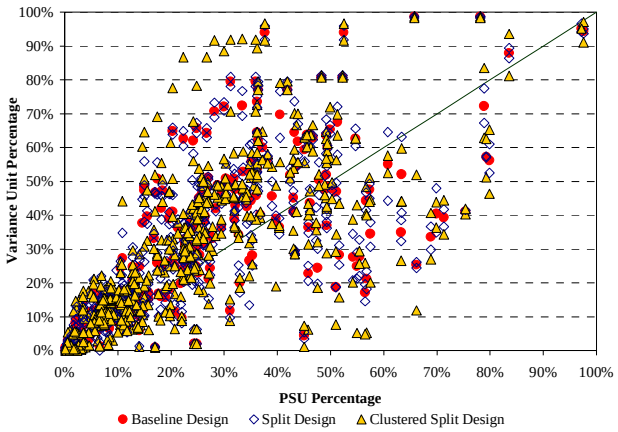

Figure 5. Demographic characteristics of PSUs and sample variance units

These results indicate that the demographics of the variance units formed in the alternative designs, especially the clustered-split design, are quite different from the PSUs. Thus, it would be difficult to match the sampled persons in a variance unit back to the PSU from which they were selected.

4. Data Analyses

The analyses compared each of the three alternative methods, the split design, the clustered-split design, and the scrambled design, with the baseline design. Descriptive statistics for 70 survey items were computed and 3 logistic regression models were fit. Standard errors were computed, with WesVar,[3] along with the survey estimates and the regression parameter estimates, using the replicate weights from each design and the JK1 method of variance estimation. The analyses did not include estimates of differences between subgroups (e.g., the difference in the number of alcoholic drinks consumed by men and women). There were two reasons for this. The appropriate comparisons, from a virtually limitless list of possibilities, would have been very difficult to identify. Also, the time needed for such analyses would have impeded a timely 1999-2000 data release.

To compare the alternative designs, we computed the ratios of the alternative method standard errors to the baseline standard error, for each survey estimate and each parameter estimate from logistic regression models. Section 4.1 contains the details of the descriptive statistics analysis. Section 4.2 contains the results of the logistic regression analysis.

The number of degrees of freedom used for all analyses across all four designs was 27, the number

[2] In combination with one or more other race/ethnicity listed.

[3] WesVar is developed by Westat (www.Westat.com).

associated with the baseline design. The reason for this is regardless of which design is used the fact that there are only 27 PSUs in the sample remains unchanged.

4.1 Descriptive Statistics

For each survey item described in Section 4, one of the following estimates using the full-sample weight were computed:

- Proportion (for example, proportion of people who are now taking prescribed medicine);
- Mean (for example, average level of total cholesterol, average diastolic blood pressure, average systolic blood pressure);
- Geometric mean (for example, mean level of cadmium, mean level of lead, and mean level of mercury as measured from blood, mean level of uranium as measured in urine); and
- Quantile such as median and the 95th percentile (for example, median height, 95th percentile for height, 95th percentile for weight).

Together with the survey estimate, the standard error of the estimate, unweighted sample size, and the design effect were computed using the JK1 method and the replicate weights. The estimates and their descriptive statistics were computed overall, and by gender, by race/ethnicity, and by collapsed race/ethnicity-sex-age domain. These analyses were carried out for each of the baseline, split PSU, clustered-split PSU, and scrambled PSU designs.

Scatter plots were also created to present the comparisons of the designs. Figures 6 and 7 plot the values of the design effects of the estimates on the x-axis and the ratios of the standard errors using the alternative designs to the standard errors using the baseline design on the y-axis, for the overall level and by race/ethnicity. In each plot, a horizontal line is drawn at 1 to indicate the level at which the two standard errors would be equal. Figure 6 shows the plot of estimates overall. Figure 7 shows the plots of estimates by race/ethnicity. Plots by all other subgroups showed patterns similar to those seen in Figures 6 and 7.

In general, the larger the design effect, the more the alternative designs underestimate the variance. This is most evident for estimates with design effects larger than two. This effect was also seen in plots for more detailed subgroups, not included in this report.

As can be seen in Figures 6 and 7, all the alternative designs under estimate the variance. This is most extreme with the scrambled design. The variance estimates under this design are also quite variable. The split and clustered-split designs produce standard errors that are approximately 20 percent smaller than those from the original design. (The ratios of standard errors tend to lie around 0.8.) However, the estimates under the clustered-split design tend to be larger, and thus more conservative, and less variable than those for the split design.

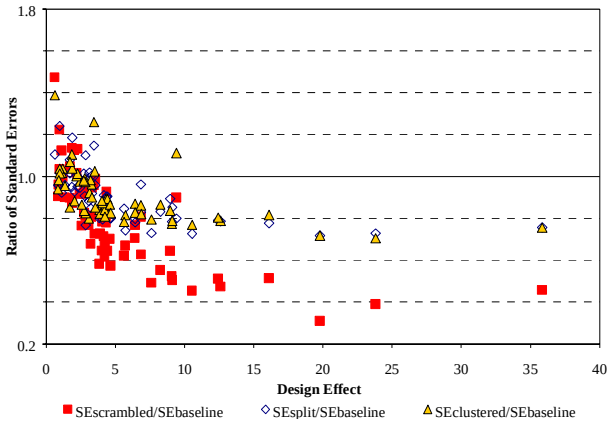

Figure 6. Ratios of standard errors against baseline design effects

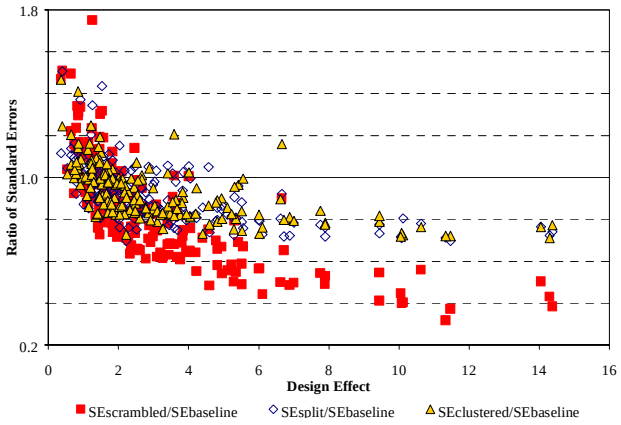

Figure 7. Ratios of standard errors against baseline design effects by race/ethnicity

4.2 Logistic Regression Models

Three dichotomous logistic regression models specified by NCHS were also performed. These models were:

- Prevalence of hypertension, predicted by demographic characteristics (age, sex and race/ethnicity), socio-economic characteristics (level of education), social behavior (currently smoking, drinking alcohol), and health condition (severity of obesity, ever told had high blood pressure, taking medicine for high blood pressure);

- Prevalence of coronary heart disease, predicted by demographic characteristics (age, sex and race/ethnicity), socio-economic characteristics (level of education), social behavior (currently smoking, drinking alcohol), and health condition (severity of obesity, high cholesterol ever told had high blood pressure); and
- Prevalence of hepatitis-C virus infection, predicted by demographic and socio-economic characteristics (age, sex, race/ethnicity, education) and social behavior (illegal drug use, high risk sexual behavior).

Each logistic regression model was fit for the baseline design, and each of the alternative designs. Changing the replicate designs only changed the standard errors of the parameter estimates, and thus the results of all hypothesis tests and the confidence intervals. The parameter estimates themselves are not effected by the replicate design. Tables 1 through 3 contain, for each of the models described above, the best models from the baseline and alternative designs. The significance levels of the predictors, as well as the overall measures of fit are also included. (The significant predictors in the final models are in boldface type.)

Tables 1 through 3 show that the significant predictors may differ from one replicate design to another. For example, using the baseline design, the prevalence of hypertension could be predicted by age, race/ethnicity, alcohol consumption, BMI (body-mass index) level and smoking status. Using the split, scrambled, or clustered-split design, sex becomes an additional significant predictor. There are no changes in the model on prevalence of coronary heart disease when the replicate design is changed. For the model on prevalence of hepatitis-C, the split and scrambled designs result in race/ethnicity and drug use as significant predictors, while drug use is the only significant predictor if the baseline or clustered-split design is used. The reason for this is that the alternative designs underestimate the variance, and thus some variables will appear significant in these models that would not be significant in the baseline design.

Before undertaking these analyses there was some concern that logistic models would need an extreme number of iterations in order to converge due to increasing the number of replicates. This was not the case. The alternative designs required the same number of iterations as the baseline design to reach convergence.

5. Summary and Conclusion

This research activity evaluated three alternative replication approaches to minimize data disclosure risks:

1. Split PSU Design—Each PSU forms two variance units by alternating segments;
2. Clustered-split PSU Design—Each PSU forms two variance units by placing lower minority segments (roughly half the segments) into one variance unit, and the higher minority segments into the other variance unit; and
3. Scrambled PSU Design—Variance units are formed with no regard to PSUs.

Each of the alternative designs was compared to the baseline design, which used the standard JK1 method. The analyses compared the demographic distribution of the variance units to the respective PSU to determine the risk of PSU identification. The performance of the different methods was evaluated by comparing the alternative variance estimates of many NHANES statistics and logistic regression predictors to those of the baseline design.

Based on the results of these analyses, it is recommended that the clustered-split design be used to form replicates for any data that are publicly released. The replicates formed in the clustered-split design have the dual advantage of preserving to the extent possible, the clustering of segments from the original design, while creating replicates that do not too closely resemble the PSUs in terms of demographic characteristics. This design results in variance estimates that are more stable than the scrambled design and slightly more stable than the split design. The clustered-split design also creates standard errors closer to the baseline estimates. The logistic regression analysis also suggests that using the clustered-split model will result in models more consistent with the baseline design when compared to the other alternatives.

However, since the variance estimates from the NHANES 1999-2000 clustered-split design are approximately 20 percent lower than the baseline design (or approximately equal to 80 percent of the baseline estimates), it is recommended that analysts multiply standard errors by the factor, *f*, to approximate what would have resulted from the baseline design:

$$f = 1/.80 .$$

Since this research was conducted on only 70 survey estimates, it may be that a different factor is more appropriate, or possibly a factor specific to the types of estimates being made. NCHS continues to investigate this issue, and we anticipate that they will release the results of their research in analytic guidelines accompanying the 1999-2000 data release.

Table 1. Results from dichotomous logistic regression of prevalence of hypertension model

All parameters investigated	Prob>F for best model			
	Baseline design	Split design	Clustered-split design	Scrambled design
Age in single years	0.0007	0.0003	0.0001	0.0000
Gender (Male, Female)		0.0257	0.0473	0.0315
Race/ethnicity (Mexican, Black, Other)				
Race/ethnicity recoded (Black, Other)	0.0394	0.0312	0.0302	0.0052
Education level (<=HS, >HS)				
More than 3 drinks/day (Yes, No)	0.0007	0.0247	0.0049	
BMI (High, Medium, Low)	0.0002			
BMI (High, Medium/Low)		0.0008	0.0021	0.0002
Ever told had high blood pressure (Yes, No)				
Smoked in last 5 days (Yes, No)	0.0326	0.0156	0.0220	
Measures of fit				
Overall F-test, Prob>F:	0.0001	0.0001	0.0002	0.0000
Negative log-likelihood:	0.0695	0.0755	0.0755	0.0782
Overall Score Test, Prob>S:	0.0009	0.0021	0.0021	0.0000

Table 2. Results from dichotomous logistic regression of prevalence of coronary heart disease

All parameters investigated	Prob>F for best model			
	Baseline design	Split design	Clustered-split design	Scrambled design
Age in single years	0.0000	0.0000	0.0000	0.0000
Gender (Male, Female)	0.0009	0.0050	0.0021	0.0018
Race/ethnicity (Mexican, Black, Other)				
Race/ethnicity recoded (Black, Other)	0.0000	0.0065	0.0006	0.0293
Education level (<=HS, >HS)				
More than 3 drinks/day (Yes, No)				
BMI (High, Medium, Low)				
BMI (High, Medium/Low)				
Cholesterol level (High, Average)	0.0017	0.0006	0.0012	0.0015
Had high blood pressure (Yes, No)				
Smoked in last 5 days (Yes, No)				
Measures of fit				
Overall F-test, Prob>F:	0.0000	0.0000	0.0000	0.0000
Negative log-likelihood:	0.1818	0.1818	0.1818	0.1818
Overall Score Test, Prob>S:	0.0000	0.0000	0.0000	0.0000

Table 3. Results from dichotomous logistic regression of prevalence of hepatitis-C virus infection

All parameters investigated	Prob>F for best model			
	Baseline design	Split design	Clustered-split design	Scrambled design
Age in single years				
Gender (Male, Female)				
Race/ethnicity (Mexican, Black, Other)				
Race/ethnicity recoded (Black, Other)		0.0383		0.0263
Education level (<=HS, >HS)				
High risk sex behavior (Yes, No)				
Used cocaine/other drug (Yes, No)	0.0000	0.0000	0.0000	0.0000
Measures of fit				
Overall F-test, Prob>F:	0.0000	0.0000	0.0000	0.0000
Negative log-likelihood:	0.0827	0.0884	0.0827	0.0884
Overall Score Test, Prob>S:	0.0000	0.0000	0.0001	0.0001