LIMITING THE RISK OF DATA DISCLOSURE BY USING SWAPPING TECHNIQUES IN VARIANCE ESTIMATION

Sylvia Dohrmann, Leyla Mohadjer, Jill Montaquila, Westat;
Randy Sitter, Wilson Lu, Simon Fraser University;
Lester R. Curtin, National Center for Health Statistics
Sylvia Dohrmann, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Confidentiality; Replication; Taylor Series Linearization; Area Sample

1. Introduction

The National Health and Nutrition Examination Surveys (NHANES) are one of a 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 has been implemented as a continuous, annual survey. Each single year and any combination of consecutive years comprise a nationally representative sample of the U.S. population.

A four-stage sample is selected for NHANES. The first stage of selection is the primary sampling unit (PSU). Within each of the selected PSUs, an average of 24 segments, or secondary sampling units (SSUs), consisting of census blocks or groups of census blocks are selected; a subsample of the households within these segments are selected and screened. Within the screened households, members of particular race/ethnicity-income-sex-age subdomains are selected with prespecified probabilities.

The practical constraints surrounding the collection of medical data in mobile examination units have limited the NHANES survey to 15 PSUs in each annual sample.[1] The small number of PSUs in the sample has dramatically increased the chance of data disclosure. To improve the precision of the published results and to reduce data disclosure risks, NCHS currently prepares public-use files for two-year samples rather than annual samples. The PSU identifiers are, of course, not included on these files.

The first data release of the continuous NHANES survey included the combined 1999-2000 annual samples. 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 before the release of these data. The alternative approaches for creating variance estimation replicates or pseudo-PSU identifiers that would mask the original PSUs were presented in Dohrmann et al., 2002. Since that paper, we have continued our investigation of this issue towards improving our methodology for the 2001-2002 data release.

Section 2 presents an overview of the PSU-splitting method used for the 1999-2000 NHANES data release as well as the disadvantages of that method. Section 3 provides a description of the methods used to create alternative sets of PSU and stratum identifiers for variance estimation. A comparison of the variance estimates from each of the approaches considered, and a discussion of how the new approach is an improvement over the previously released method for estimating variances in NHANES is presented in Section 4.

2. PSU-Splitting

Because no explicit stratification was used to select the PSUs for NHANES 1999-2000 (Montaquila et al., 1998) and because of the small number of PSUs in the sample, the delete-1 jackknife was used to create replicates for variance estimation for the analysis of the NHANES 1999-2000 data; 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

Various methods for splitting each PSU into two dissimilar pseudo-PSUs creating a total of 52[2] were considered for the purpose of disclosure limitation. The impact on the performance of the resulting jackknife variance estimates and on the disclosure of original PSU indicators was examined (see Dohrmann et al., 2002 for more detail). The final chosen method for the 1999-2000 NHANES release (termed the "clustered-split

[1] There were 12 PSUs in the 1999 annual sample.

[2] There were 27 PSUs in the 1999-2000 sample, including one certainty PSU fielded in both years. All the noncertainty PSUs were split to form a total of 50 pseudo-PSUs. The two certainty PSUs were combined formed two pseudo-PSUs as in the baseline design for a total of 52 pseudo-PSUs.

PSU" alternative in Dohrmann et al.) entailed ordering the SSUs on minority density and then assigning the first half within a PSU to one pseudo-PSU and the second half to another, as depicted in Figure 2. Due to the ordering on minority density, one expects that the resulting pseudo-PSUs formed from this method will not have the same characteristics as the full PSU. In addition, the order of the replicates was then scrambled to further ensure confidentiality.

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 2. Clustered-split PSU replication design

There is little practical difference in terms of confidentiality between supplying the end-user the 52 resulting set of jackknife replicate weights or giving the pseudo-PSU indicators, as one can easily obtain one from the other[3] (Yung, 1997; Shah, 2001; Lu, 2004). The questions are: i) is it now easy to re-match units to original PSUs; and ii) will the resulting jackknife variance estimate still produce variance estimates that are similar to the variance estimates from the baseline design on various characteristics.

Initially, the protection of confidentiality seemed more satisfactory than did the performance of the resulting variance estimator. For the 70 characteristics investigated, the cluster-split jackknife variance estimates on average performed reasonably well for estimates with design effects of less than 2. However, for design effects ranging from 2 to 5, this method resulted in an underestimation pattern that became even more pronounced for design effects larger than 5. This pattern is re-created in Figure 3 which plots the values of the design effects of the baseline estimates on the x-axis and the ratios of the estimated standard errors using the PSU-splitting alternative to the estimated standard errors using the baseline design on the y-axis, by race/ethnicity. Plots by all other subgroups showed similar patterns. There are two aspects to this plot. One is the observation that on average the estimated standard errors of the split-PSU jackknife are under-estimates relative to the baseline jackknife. The other aspect is the overall pattern with design effect.

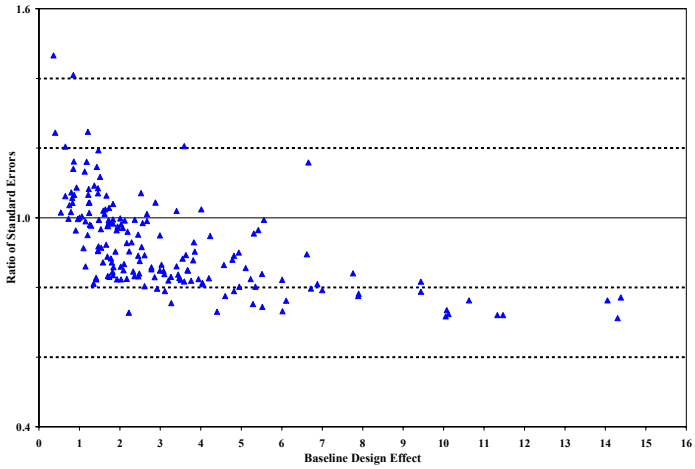

Figure 3. Ratios of standard errors (clustered-split divided by baseline) against baseline design effect by race/ethnicity using PSU-splitting alternative

2.1 Evaluation of PSU-Splitting

Further research revealed the reasons behind the underestimation exhibited in Figure 3. Consider the more simple case of n clusters selected with replacement with inclusion probabilities $\pi_i = np_i$. Let $\hat{Y} = \sum_{i=1}^{n} \hat{Y}_i / (np_i)$ be a linear unbiased estimator of population total Y, where $\hat{Y}_i$ is a linear unbiased estimate of the cluster total Y_i based on sampling at the second and subsequent stages. This estimator of Y may be written as

$$\hat{Y} = \sum_{(ik)\in s} w_{ik} y_{ik}, \qquad (1)$$

where s is the total sample of elements and w_{ik} and y_{ik} respectively denote the sampling weight and the item value attached to the (ik)-th sampled element ($k = 1, \ldots, n_i$; $i = 1, \ldots, n$). The delete-1 jackknife estimator of $\hat{Y}$ is constructed by recalculating the weights w_{ik} each time a sample cluster j is deleted ($j = 1, \ldots, n$). This is done as follows: $w_{ik(j)} = 0$ if $j = i$; $w_{ik(j)} = nw_{ik}/(n-1)$ if $j \neq i$. Replacing w_{ik} with the jackknife weights $w_{ik(j)}$ in (1) we get $\hat{Y}_{(j)}$ and the jackknife estimator is given by

$$v_J = \frac{(n-1)}{n} \sum_{j=1}^{n} (\hat{Y}_{(j)} - \hat{Y})^2 . \qquad (2)$$

If we write $\hat{Y} = \bar{r} = \sum_{i=1}^{n} r_i / n$, where $r_i = \sum_{i=1}^{n} nw_{ik} y_{ik}$, it is not difficult to show that

[3] This assumes that any replicate weight adjustments are similar to the full-sample weight adjustments.

$$v_J = \frac{1}{n(n-1)} \sum_{j=1}^{n} (r_j - \bar{r})^2 \ . \qquad (3)$$

Under the clustered-split design, the 2n PSUs are treated as if they were true PSUs. With each PSU separated into two pseudo-PSUs $G_{1,i}$ and $G_{2,i}$, $r_i = \sum_{k \in G_{1,i}} nw_{ik} y_{ik} + \sum_{k \in G_{2,i}} nw_{ik} y_{ik} = r_{1,i} + r_{2,i}$, and $\hat{Y} = (1/n) \sum_{i=1}^{n} (r_{1,i} + r_{2,i})$. To delete the (g,j)-pseudo-PSU we define $\hat{Y}_{(g,j)} = \sum_i \sum_k w_{ik(g,j)} y_{ik}$, where $w_{ik(g,j)} = 0$ if $j = i$ and $k \in G_g$ and $w_{ik(g,j)} = 2nw_{ik}/(2n-1)$ otherwise. Then $\hat{Y}_{(g,j)} = \hat{Y} - (2r_{g,j} - \bar{r})/(2n-1)$ for $g = 1, 2$ and $j = 1, \ldots, n$. Applying (3) the jackknife estimator v_{J2} becomes

$$v_{J2} = \frac{1}{2n(2n-1)} \sum_{j=1}^{n} \sum_{g=1}^{2} (2r_{g,j} - \bar{r})^2 \qquad (4)$$

which reduces to

$$v_{J2} = \frac{(n-1)}{(2n-1)} v_J + \frac{1}{n(2n-1)} \sum_{j=1}^{n} (r_{1,j} - r_{2,j})^2 \ . \qquad (5)$$

Viewing (5) we see that v_{J2} is approximately equal to $v_J/2$ plus a nonnegative number that is dependent on the between SSU-variability of within the split PSUs.

If it were possible to determine a PSU-splitting scheme that would result in the second term being almost zero, we could simply instruct users to double the resulting variance estimates to arrive at a value of v_{J2} almost identical to v_J. In Dohrmann et al. (2002) we describe a method we called the "scrambled" design in which the pseudo-PSUs were created in a random manner. As a result, the between-SSU variability was quite small and v_{J2} greatly underestimated v_J. With the "clustered-split" method, the between-SSU variability is high between the two pseudo-PSUs $G_{1,i}$ and $G_{2,i}$ since they were created by grouping segments with similar minority levels together; as a result, the estimates of v_{J2} were larger.

2.2 Need for New Approach

After the release of the 1999-2000 NHANES data, with the cluster-split jackknife weights, there was a need to change the basic methodology used for variance estimates. Beginning in 2002, NHANES is a stratified design, with two-PSUs per stratum for the two-year samples. Given this, and the great number of replicates needed for continual two-year releases, NCHS decided that in future data releases only PSU and strata indicators will be released for variance estimation. As a result a new method of variance estimation had to be developed for use with the publicly released data.

After reviewing the research described in Section 2.1, the obvious alternative to the clustered-split design is to recombine the splits $G_{1,i}$ and $G_{2,i}$, with splits from other PSUs to form the pseudo-PSUs. This strategy of PSU-splitting and recombining is merely one method of changing the SSU assignment, or swapping segments between PSUs.

3. SSU Swapping

Many surveys swap data values between cases for disclosure limitation. Rather than swapping individual values or records, we examined alternative approaches of swapping segments. That is, for two similar segments in different PSUs, swapping the PSU and variance strata identifiers for all sampled cases.

After considering several swapping algorithms it became clear that any segment swapping algorithm used for NHANES should (1) ensure an adequate swapping rate per PSU for confidentiality protection such that the original PSUs cannot be readily reconstructed after swapping, and (2) select swapping partners that ideally are identical in the matching characteristics to minimize the bias of variance estimates. However, solving a linear programming problem to satisfy both swapping rate and matching criteria is nontrivial and some compromises are necessary to reduce the computation time.

An alternative is to apply record linkage techniques to identify optimal swapping partners and to control the swapping rate through sampling. In summary, this method involves three basic steps: matching, sampling, and bias evaluation. These steps are repeated to adjust the sampling (swapping) rate and the matching method. The process is stopped when we are satisfied that, on average, swapping has negligible effects on key variance estimates.

The three basic steps of this approach are implemented as follows. Step 1 applies record linkage techniques to conduct complete matching of the segments. Matching uses constraints to prohibit the pairing of segments from the same PSU, and includes a barrier to avoid poor matches (segments with no good matching partners are not swapped). The selection of matched pairs uses a one-to-one matching algorithm to provide an optimal matched set with the maximum overall likelihood of being good matches. The next section provides more detail on the concepts of record linkage. Step 2 involves sampling a fixed percentage of the matched segments within each PSU for swapping. Sampling controls the maximum number of segments for swapping (i.e., the swapping rate) per PSU. Since only a limited number of segments are swapped, there is no obvious information to help reconstruct the original PSUs. Step 3 is to conduct variance analyses to measure the potential bias resulting from swapping.

3.1 Matching with AutoMatch

We used a probability-based record linkage technique to form optimal segment pairs for swapping. The theory for record linkage given by Fellegi and Sunter (1969), and Winkler (1995) discusses the implementation and parameter estimation. We used the software AutoMatch (by MatchWare Technologies, Inc., 1996) for implementation (see Winglee et al., 2000; and Gomatam et al., 2002; for applications with this package). This software requires the user to estimate the conditional probabilities of agreement (m_v) and disagreement (u_v) for each matching variable and calculates the log-odds weights for all possible record pairs. It then determines the optimal set of pairs by taking the set with the greatest sum of weights. An iterative procedure can also be used to refine the values of the conditional matching probabilities m_v and u_v.

Table 1. Parameters used for matching 2001-2002 NHANES segments

Segment level demographic variable	m_v	u_v	d	Agree weight	Disagree weight
Variable 1	0.90	0.015	10	5.86	-3.45
Variable 2	0.70	0.010	10	6.09	-1.68
Variable 3	0.80	0.015	5	5.58	-2.32
Variable 4	0.95	0.015	20	6.06	-4.47
Variable 5	0.40	0.045	10	3.13	-0.67
Variable 6	0.90	0.010	20	6.49	-3.29

For NHANES, we used six variables describing various demographic characteristics (for example, the percentage of a particular race or ethnicity within the segment) to determine segment pairs. We ran AutoMatch several times to refine the m_v and u_v values that should be used for each variable. Table 1 shows the matching parameters used for NHANES 2001-2002. The first four variables ranged from 0 to 100, and were matched numerically. An extra parameter, d, was included that allowed the weight to be prorated if it differed by a certain amount. For example, if the value of variable 1 differed between two segments by 1 percent, the weight for that pair would be slightly less than the full agreement weight, rather than the full disagreement weight. Only if the difference was over 10 percent would the pair be given the full disagreement weight.

Two variables had much smaller ranges, and were matched by comparing the percentage difference between the segment pairs. Again, an extra parameter was used to allow for small disagreements between the pairs.

3.2 Application and Evaluation on NHANES

Three basic strategies were investigated by application to NHANES, referred to as Methods 1, 2, and 3. Each method swapped segments (SSUs) between PSUs. Once segments were swapped, SUDAAN was used to calculate variance estimates via Taylor series using the resulting pseudo-PSUs. In all cases the certainty PSUs were not included in the procedure.

Method 1 used the basic strategy of: Step 1, considering the entire set of segments during the matching step; and then, Step 2, carefully viewing the matched pairs in order to choose which to swap.

In Method 1, a cutoff in the program was used in Step 1 to control the total number of matches through a score. Various values of this score were tried until at least a predetermined number of SSUs were matched from each PSU (recall that each PSU contains approximately 24 SSUs). In Step 2, to determine segments to be swapped after the matches were made, first several pairs were discarded so that each PSU had approximately α% of its segments swapped and replaced by segments from other PSUs. The decision to discard each pair was made case-by-case so that the pairs that did not match as well were eliminated first, but each PSU still had the correct number of segments swapped. After this procedure was finished, two follow-up procedures (Methods 1a and 1b) were performed.

Method 1a was similar but left fewer segments swapped from each PSU by discarding more pairs in Step 2. In Method 1b, only the pairs in the PSUs determined to have the highest disclosure risk were examined. Pairs in each of these PSUs that had the highest match weight and did not match to segments in other high-risk PSUs were selected. By so doing the greatest exposure risks were masked while total swapping was reduced.

Method 2 used a slightly different strategy which in some sense reversed Steps 1 and 2 in Method 1. First a random sample of approximately α% of the SSUs from each PSU was drawn and then the sampled SSUs matched to each other via the same record linkage algorithm. This was repeated a few times and the best result retained, denoted Method 2, i, ii, iii, and iv. Method 2a (i, ii, iii, and iv) was similar to Method 2 but swapped fewer SSUs per PSU.

Method 3 employed the same basic strategy as Method 1, with segment matching performed within the PSU pairs formed for variance estimation.

For each swapping method, descriptive statistics were calculated for the standard error relative to the baseline standard error for a variety of characteristics selected from various components of NHANES (see Tables 2 and 3), which were not used in the matching. Ideally, these values would be close to 1, meaning the standard errors were not greatly affected by the swapping procedures.

4. Results

Table 2. Summary statistics for the ratio of standard errors for swapping methods to the baseline standard errors (NHANES 1999-2000)

Method	Mean	Std Dev	Min	1st Qrtile	Me-dian	3rd Qrtile	99th Pctile	Max
1	0.96	0.22	0.40	0.82	0.94	1.06	1.65	2.91
1a	0.97	0.17	0.44	0.87	0.96	1.05	1.44	2.91
1b	0.99	0.11	0.38	0.95	1.00	1.03	1.28	1.78
2	0.96	0.20	0.28	0.84	0.95	1.06	1.56	2.23
2i	0.98	0.29	0.39	0.84	0.95	1.07	1.76	6.89
2ii	0.97	0.29	0.38	0.82	0.94	1.07	1.65	5.97
2iii	0.98	0.25	0.45	0.84	0.95	1.08	1.74	4.02
2ai	0.98	0.18	0.44	0.87	0.97	1.06	1.53	2.35
2aii	0.98	0.25	0.49	0.88	0.97	1.06	1.52	6.24
2aiii	1.00	0.21	0.28	0.89	0.98	1.09	1.65	3.27
2aiv	0.98	0.14	0.43	0.89	0.97	1.05	1.40	1.50
PSU Splitting	0.95	0.17	0.54	0.85	0.93	1.02	1.49	3.41

Table 2 shows some descriptive statistics for the swapping techniques examined for the NHANES 1999-2000 data. Based on these results, we only examined Methods 1b and 2a for the NHANES 2001-2002 sample. The results of these methods are shown in Table 3.

Table 3. Summary statistics for the ratio of standard errors for swapping methods to baseline standard errors (NHANES 2001-2002)

Method	Mean	Std Dev	Min	1st Qrtile	Median	3rd Qrtile	99th Pctile	Max
1b	0.98	0.10	0.52	0.93	0.99	1.02	1.33	1.60
2ai	0.97	0.20	0.50	0.88	0.96	1.04	1.50	4.19
2aii	0.97	0.16	0.33	0.87	0.97	1.05	1.49	1.84
2aiii	0.97	0.18	0.38	0.87	0.96	1.04	1.55	3.03
2aiv	0.97	0.16	0.48	0.86	0.96	1.05	1.45	1.77
2av	0.98	0.19	0.51	0.89	0.96	1.04	1.50	4.09
2avi	0.97	0.17	0.48	0.87	0.96	1.05	1.49	2.00

In addition to the summary tables, several plots were created. The first of these plotted the baseline design effect on the x-axis and the ratio of estimated standard errors on the y-axis as shown in Figure 4 for Method 2ai by race/ethnicity.

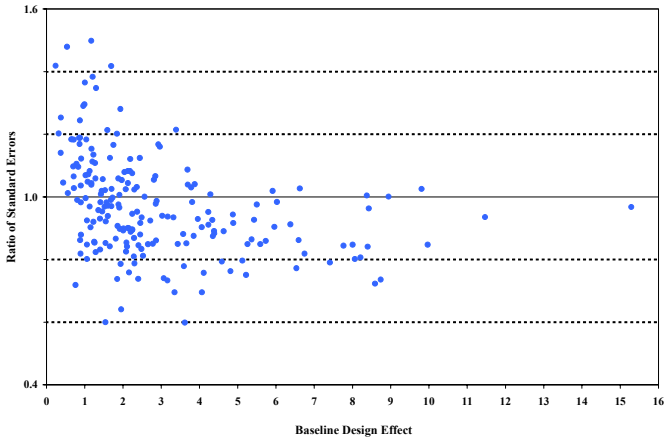

Figure 4. Ratios of standard errors against baseline design effect by race/ethnicity using SWAP 2ai alternative

At first, this plot seemed like a reasonable representation of the impact of swapping. However, the observed pattern, which was similar to that observed using PSU-splitting as depicted in Figure 3 was of concern. It had been hoped that the more sophisticated methodology would not only reduce the under-estimation problem observed in Figure 3, which it did as seen in Tables 2 and 3, but also eliminate the observed curved pattern. This caused us to reconsider the nature of this plot entirely. We concluded that this plot is not necessarily the best way of viewing the swapping performance. The pattern represents the relationship between clustering structure and design effects. To evaluate the masking procedure the chart of the baseline standard error against the swapped standard errors (or the respective CVs to remove the effect of unequal weighting) should be considered. That chart is shown in Figure 5.

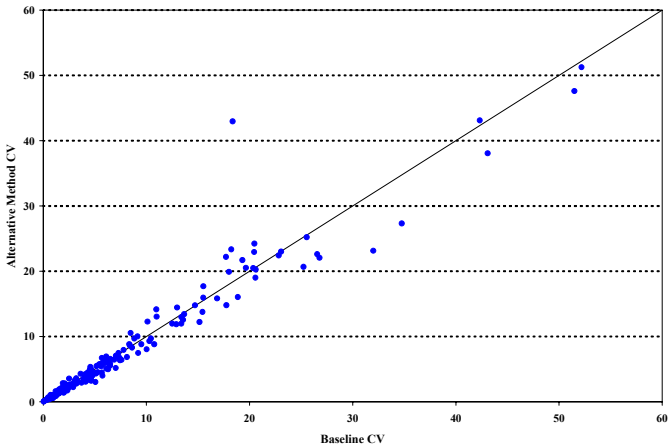

Figure 5. Chart of swapped CV against baseline CV by race/ethnicity for SWAP 2ai alternative

The method ultimately selected for the NHANES was a compromise between Method 1b and Method 2a. The resulting pseudo-PSU and strata indicators were released for the 1999-2000 sample in addition to the 2001-2002 sample. The two sets of indicators together may be used to compute variances for the combined four-year sample.

5. Further Research

We continue to investigate the optimal method of keeping the NHANES PSUs confidential. Future research includes automating the matching procedure further to reduce the amount of manual work, and to allow us to more quickly evaluate different matching variables and swapping parameters. Lu (2004) develops this capability.

6. References

Dohrmann, S., Curtin, L., Mohadjer, L., Montaquila, J., and Lê, T. (2002). National Health and Nutrition Examination Survey: Limiting the Risk of Data Disclosure Using Replication Techniques in Variance Estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 807-812.

Fellegi, I.P., and Sunter, A.B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, **64**, 1183-1210.

Gomatam, S., Carter, R., Ariet, A., and Mitchell, G. (2002). An Empirical Companion of Record Linkage Procedures. *Statistics in Medicine,* **21**, 1485-1496.

Lu, W. (2004). *Confidentiality and Variance Estimation in Complex Surveys*. Doctoral dissertation, Simon Fraser University, Burnaby, Canada.

Matchware Technologies, Inc. (1996). *AutoMatch: Generalized Record Linkage System User's Manual*. Silver Spring, MD: Matchware Technologies, Inc.

Montaquila, J., Mohadjer, L., and Khare, M. (1998). The Enhanced Sample Design of The Future National Health and Nutrition Examination Survey (NHANES). *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 662-667.

Shah, B.V. (2001). *A Method to Create Pseudo Strata and PSUs Based on BRR Weights.* Unpublished memorandum to Ken Kaplan of the Bureau of the Census.

Winglee, M., Valliant, R., Brick, J.M., and Machlin, S. (2000). Probability Matching of Medical Events. *Journal of Economic and Social Measurement*, **26***,* 129-140.

Winkler, W.E. (1995). Matching and record linkage. In B.G. Cox et al. (Eds.), *Business survey methods* (pp. 355-384). New York: J. Wiley.

Yung, W. (1997) Variance Estimation for Public Use Files Under Confidentiality Constraints. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 434-439.