Epidemiologic Studies Utilizing Surveys: Accounting for the Sampling Design

Edward L. Korn, Barry I. Graubard
Edward L. Korn, Biometric Research Branch, National Cancer Institute, EPN-739, Bethesda, MD 20892

(1) The Controversy

We start by describing a controversy that recently appeared in the medical literature concerning the appropriate way to analyze an association between a person's body iron stores and the risk of developing cancer. The data source used for this analysis was a follow-up of a large national health survey called the first National Health and Nutrition Examination Survey (NHANES I). The original survey was conducted in 1971-1975 and the follow-up was conducted in 1982-1984.

Using NHANES I and its follow-up, Stevens et al. [1] found a statistically significant difference in two measures of body iron, total iron-binding capacity and transferrin saturation, in men who developed cancer and men who did not. Age and smoking status were controlled for in the analysis. The sampling design was ignored in the analysis. Yip and Williamson wrote a letter to the editor criticizing their lack of use of the sampling design and suggested that if the sampling design had been incorporated, "it is doubtful that [these differences] ... would have remained 'statistically significant'"[2].

Stevens responded in his own letter to the editor, "It is important to take into account the probability sampling method used by NHANES when one is attempting to estimate the level of a variable in the U.S. population at large. To test for differences between case patients and controls within the NHANES cohort, however, the methods we used are appropriate." [3]

Before going on, let us examine these two analyses. We concentrate attention on men and on one of the two iron variables, "total iron-binding capacity". Results for the other variable and for women can be found elsewhere [4].

Table 1:

	Cancer				
Analysis	No	Yes	Diff.	SE	p-value
OLS	62.8	61.3	1.5	0.6	0.013
Design-Based	63.6	62.7	0.9	0.9	0.29

The top line represents an ordinary least squares analysis. The 62.8 is the adjusted mean value among the 3116 men who did not develop cancer, and the 61.3 among the 242 men who did develop cancer. The means have been adjusted by linear regression for age and smoking status. The difference in these values is 1.5 moles per liter with a standard error of .6 for a p-value of .013. This analysis is one of the analyses that Stevens did in his paper [1]. The second row contains the results of a design-based analysis. This analysis uses the survey design in the standard way, i.e., incorporates both the sample clustering and sample weights of the observations. The difference has become smaller, the standard error larger, and the result is no longer statistically significant.

(2) Sample Clustering

For reduction of costly traveling, and for the lack of complete list sampling frames, large scale surveys usually have a clustered design. We use NHANES I as an example throughout; full design details are available elsewhere [5-7]. The mainland United States is divided into approximately 1900 primary sampling units (PSU's) which are grouped into strata. Each PSU consists of a standard metropolitan area, or at most three contiguous counties. One or two PSU's are sampled from each strata, from which a limited number of census enumeration districts are sampled. From these a limited number of segments are sampled, and then a limited number of households. Finally, individuals are sampled from the selected households.

Clustering can lead to a dependence among the observations which will lead to an inflation in variances. This is true not just for population means, but for associations too. Consider the following simple example. Suppose we select a simple random sample of households, and then sample 4 children in each household. Assume we are interested in the association of drinking well water with having had chickenpox. If we assume that everyone in the house has the same water supply, and if one child has had chickenpox then his siblings will have had it too, then the effective sample size for measuring this association is 1000 and not 4000. If we ignore the sampling design and use 4000 as our sample size, we will be underestimating the variance of the observed association by a factor of 4, or the standard error by a factor of 2.

Now while this extreme example shows how clustering can lead to a dependence of the observations, it does not imply that it must. In fact it is easy to write down sufficient conditions for ignoring the clustering in the analysis, for example, if the residuals are independent. It is not so easy to decide from the data whether the clustering has or has not lead to dependence that will affect your analysis.

Fortunately, survey samplers have over the years developed methods which estimate standard errors no matter what the dependence, e.g., Taylor series linearization, balanced half sample repeated replication, and the jackknife [8]. These methods capture the variability of estimators by measuring the variability at the PSU level. All dependence due to clustering at the lower levels of sampling is automatically incorporated. Thus, these replication methods yield approximately unbiased estimates of variances whether or not there is dependence.

Since adjusting for the sample clustering will on average increase the standard errors, its use in the analysis can be thought of as a conservative procedure. What is lost by using the clustering in an analysis when it was unnecessary? One potential problem with these replicated variances is their variability. This becomes more of a problem with limited numbers of sampled PSU's. In particular, a replicated variance is distributed approximately as a multiple of a chi-square distribution with d degrees of freedom. For a stratified design, d is equal to the number of sampled PSU's minus the number of strata. For NHANES I body iron analyses, the design can be approximated by a sample of 67 PSU's from 32 strata which leaves 35 degrees of freedom, a reasonably large number. But consider another HANES survey, the Hispanic HANES (HHANES). The design of that survey can be approximated by a sample of 16 PSU's from 8 strata [9], leaving only 8 degrees of freedom.

A good way to describe this variability is in terms of the inefficiency of using the sample clustering when it was actually unnecessary [10,11]:

For 1 parameter

$$\text{Inefficiency}_0 = 1 - (z^{\alpha/2}/t_d^{\alpha/2})^2$$

For parameters (simultaneously)

$$\text{Inefficiency}_0 = 1 - \chi_p^{2\,\alpha}/(pF_{p,d-p+1}^{\alpha})$$

The subscript 0 is a reminder that this is being calculated under the null hypothesis that there is no dependence. For 1 parameter, if one were doing hypothesis testing or constructing a confidence interval using a replicated variance, one would use a t-distribution with d degrees of freedom. This is represented by the upper $\alpha/2$ tail of the appropriate t-distribution. If one knew a priori that the clustering was irrelevant, then one would not need to use a replicated variance, but could use a standard ordinary least squares variance. Since the surveys are large, one could use the upper tail of a normal distribution.

This inefficiency measures the increase in sample size necessary to compensate for using a replicated variance when unnecessary. If the analysis involves simultaneous inference for p parameters, then the inefficiency will be greater as represented by the above formula that involves a chi-square and F distribution. Table 2 presents some examples from NHANES I and HHANES.

Table 2: Inefficiencies of using clustering when unnecessary for $\alpha = .05$

# parm.	NHANES I	HHANES
1	7%	28%
2	9%	37%
5	12%	65%

Notice that for the case we are interested in the inefficiency is quite small, 7%.

In this case when the inefficiency is small, we would recommend using the replicated variances which accounts for the sample clustering. The idea is that if there is dependence, then one has correctly accounted for it. If there is not, then one has not lost much. "Small" here depends somewhat on the context. Ten percent is probably always small, but 100% could be small too if it means the standard error is going from .1 to .2 on an expected relative risk of 3.

What do you do if this inefficiency is not small? We offer the following general recommendations.

Summary of Recommendations -- sample clustering

(1) If Inefficiency_0 is "small", then uses the clustering in a standard (design-based) way to obtain replicated variances.

(2) If Inefficiency_0 is not small, then
   (a) drop the strata boundaries, and/or
   (b) drop the PSU boundaries.

This last recommendation (2) is tentative and presently under study. It will increase the degrees of freedom at the cost of potentially estimating the variance with bias.

(3) Causality

Before discussing sample weights, it is useful to discuss briefly causality. One possible definition of the magnitude of the causal effect of a risk factor on an outcome is the expected association one would see from a randomized trial of that risk factor [12]. With such a definition, a key notion is whether a variable is "exogenous" or not. An "exogenous variable" is one that would not be affected by the treatment assignment in a hypothesized randomized trial of the risk factor [13]. Under suitable conditions, an analysis of a risk factor/disease association will not be "hurt" by conditioning on exogenous variables. In fact, such conditioning may lead one closer to the causal effect of the risk factor.

(5) Sample Weights

For complex surveys, each sampled individual with data has a "sample weight" associated with his data. The sample weight is the number of individuals in the target population that the sampled individual represents. The sample weight may be derived as the product of three components. The first component comes from the fact that surveys frequently over-sample certain groups in the population (base weight). For example, in NHANES I, people living in poverty census enumeration districts were sampled at either two or eight times the rate as people living elsewhere. Additionally, persons aged 65 years or over were sampled twice as often as women aged 20-44, who were in turn sampled twice as often as other adults. Finally, in a small number of segments there was subsampling of households at rates of 1/2 to 1/4 because of inaccuracies of census block listings ("supplemental block" and "duplication control").

The second component of the sample weight is an adjustment for non-response, including both the inability to locate sampled individuals as well as their refusal to participate. In NHANES I the non-response adjustment was based on five family income groups within each PSU. This adjustment to the sample weight was truncated at 3.

The third component of the sample weight is an adjustment so that the sum of the weights for a given sex, race, and age agree with known population figures (poststratification adjustment). Finally, since not all sampled persons in NHANES I had their body iron stores measured, a special set of weights were derived for use in the analysis of these data.

The reason survey analysts use sample weights is that they lead to approximately unbiased estimates of population quantities. For example, weighted least squares regression, in which the weights are the sample weights, estimates population regression coefficients. Ordinary least squares (OLS) will not in general.

Why not always do a weighted analysis? As with the use of clustering when unnecessary, the use of sample weights when unnecessary leads to an inefficient analysis. Since in this case, OLS is optimal, the inefficiency can be defined as

$$\text{Inefficiency}_0 = 1 - \text{Var}(\hat{\beta}_1^{OLS}) \,/\, \text{Var}(\hat{\beta}_1^{WLS})$$

where β_1 is some regression coefficient of interest. The subscript 0 is again a reminder that this inefficiency is being computed under a null hypothesis (that the OLS coefficient is actually unbiased). We now discuss three ways of estimating this inefficiency from the data. We start with the case of a simple mean, for which the unweighted and weighted estimates are given by

$$\bar{y}^{OLS} = \sum_{i=1}^{n} y_i \,/\, n \quad \text{and} \quad \bar{y}^{WLS} = \sum_{i=1}^{n} w_i y_i \,/\, \sum_{i=1}^{n} w_i \ ,$$

respectively.

(Method 1) Assume y_i are i.i.d.

$$\text{Inefficiency}_0 = 1 - \frac{(\sum w_i)^2}{n \sum w_i^2}$$

(Method 2) Assume y_i are independent

$$\text{Inefficiency}_0 = 1 - \frac{\sum (y_i - \bar{y})^2 \,/\, n^2}{\sum w_i^2 (y_i - \bar{y})^2 \,/\, (\sum w_i)^2}$$

(Method 3) No assumption on the y_i

$$\text{Inefficiency}_0 = 1 - \hat{Var}(\bar{y}^{OLS}) \,/\, \hat{Var}(\bar{y}^{WLS})$$

Method 1 utilizes only the variability of the sample weights. For method 2 the y's do not have to be identically distributed. In particular, it allows for the possibility that the y's are correlated with the weights. Method 3 makes no distributional assumption concerning the y's, and so these variances

could be estimated by one of the replication methods discussed earlier.

We now consider the corresponding formulas for estimating the inefficiency of a single regression coefficient β_1 based on the model y = X β + error.

(Method 1) Assume errors are i.i.d.

$$\text{Ineff}_0 = 1 - \frac{[(X'X)^{-1}]_{11}}{[(X'WX)^{-1}(X'WWX)(X'WX)^{-1}]_{11}}$$

(2) Assume errors are independent

Ineff_0 = 1 -where Σ = diagonal $((\text{residual}_i^2))$

(3) No assumption on errors

$$\text{Ineff}_0 = 1 - \hat{Var}(\hat{\beta}_1^{OLS}) \,/\, \hat{Var}(\hat{\beta}_1^{WLS})$$

Here W is the diagonal matrix with the sample weights on the diagonal, and the subscript (11) refers to the diagonal element of the matrix corresponding to β_1.

The inefficiencies for the analysis of total iron-binding capacity are quite high:

Table 3: Inefficiency_0 of using sample weights for analysis of Total Iron-Binding Capacity for Males

Method	Assumption on Residuals	Estimating: Mean TIBC	β_1
1	i.i.d.	47%	58%
2	indep.	48%	70%
3	none	20%	60%

This is not surprising given the high variability of the sample weights:

Table 4: Percentiles of Sample Weights for Men

Maximum	135,824
95%	31,556
75%	19,504
Median	8,170
25%	3,614
5%	876
Minimum	477

Although the estimation of the inefficiency using method 3 has appeal because it makes the fewest assumptions, notice that it might be very variable because it uses replicated variances. To study this we performed a limited simulation. Using the observed sample weights from the 3358 observations, we generated an independent normal deviate to be the y for each observation. Then the inefficiency was calculated for the mean using methods 2 and 3. (For method 1, the inefficiency is .47 regardless of the y data). We repeated this process 100 times with the following results:

	Mean	SD	MIN	MAX
Method 2	.47	.03	.38	.54
Method 3	.46	.17	.004	.77

Since the y data for this simulation was generated to be i.i.d., the correct answer is .47. Although all methods are unbiased in this simulation, the variability of method 3 is unacceptably large. Therefore, we do not recommend this method.

Since the inefficiency of performing a weighted analysis is unacceptably high, what do we recommend? One approach is to use an unweighted analysis but to control for any exogenous variables utilized in determining the sample weights. In the present context, in addition to age which is already included in the model, we augment the model with income levels, an indicator variable whether or not age<65, an indicator variable whether or not the household was in a poverty census enumeration district, and race. We then perform an OLS analysis to estimate the regression coefficients. The standard errors of the coefficients are estimated using a replication method, since that inefficiency was small.

Table 5: Regression analysis augmented with exogenous design variables (OLS estimates -- replicated variances)

Variable	β	p-value
No Cancer vs. Cancer	1.4	.010
Age (years)	-.048	.004
Smoking (vs. never)		.027
current	0.9	
former	0.7	
unknown	1.6	

Table 5 (continued)

Variable	β	p-value
Income (5 levels)	small	.95
Age<65 (Yes/No)	-1.5	.004
Poverty ED (Yes/No)	0.1	.83
Race (White/non.)	-1.3	.022

Adding these design variables changes the estimated coefficient for the cancer indicator (x_1) very little, from 1.5 to 1.4. Additional inclusion of interactions among the covariates did not change the coefficient for the cancer indicator. We do not recommend including interactions of the design variables with x_1 in the model for the following reason. If these interactions are included, then there will be many estimates of the treatment effect corresponding to different values of the design variables. One is then left with trying to average these estimates for an overall estimate of the treatment effect. This averaging can be done in a weighted or unweighted manner, neither of which works well.

As a final check on the model, we recommend testing whether the unweighted regression coefficient from the augmented model is unbiased. This can be done in a model-based or design-based manner [14,15], with the latter approach being preferred. In the present context, this can be simply done by using the augmented model and estimating the unweighted β_1 (=1.43), the weighted β_1 (=0.72), and replicating the standard error of their difference (SE=0.86). In this case the coefficients are not significantly different.

Here is a summary of the different analyses:

Analysis using Cluster.	Weights	Cancer No	Yes	Diff.	SE	p-value
No	No	62.8	61.3	1.5	0.6	.013
Yes	Yes	63.6	62.7	0.9	0.9	.29
Yes	*	63.5	62.1	1.4	0.5	.010

Our recommended analysis is the last line, which uses the clustering in a standard design-based way, but models the weights by including design variables in the regression model. For this particular example, our recommended analysis turns out to be very similar to an analysis which ignores the sampling design completely.

Summary of Recommendations -- sample weights

(1) If Inefficiency_0 is "small", then use the weights in a standard (design-based) weighted analysis.

(2) If Inefficiency_0 is not small, then

(a) Augment the model with exogenous design variables.

(b) Do not put interactions with the main treatment variable (x_1) in the model.

(c) Test whether the OLS coefficient for x_1 is biased, i.e., the weights matter.

(3) If this test (2c) does not reject, stop.

(4) If this test does reject, then truncate the weights the maximum amount so that the test of the bias of the truncated beta does not reject.

This last recommendation (4) is tentative and presently under study.

(6) Discussion

We end with two suggestions for the designers and producers of large-scale health surveys that will be utilized by other investigators. Following our recommendations, one will sometimes need to utilize variables relating to the design and non-response adjustments of the survey. Our first suggestion is that all these variables be documented and included in public use data files. Our second recommendation concerns the design of the surveys themselves. As we have discussed, a small number of sampled primary sampling units, and/or individuals with extremely high sample weights can make the use of the sample design in the analysis very inefficient. Given the large number of secondary analyses that are performed using these survey data, we suggest that the policy makers provide sufficient resources so that the surveys can be designed for efficient secondary analyses using the design.

References

(1) Stevens RG, Jones DY, Micozzi MS et al. Body iron stores and the risk of cancer. *N Engl J Med* 319:1047-1052, 1988

(2) Yip R, Williamson DF. Body iron stores and risk of cancer. *N Engl J Med* 320 (To the Editor):1012, 1989

(3) Stevens RG, Jones DY, Micozzi MS, et al. Body iron stores and risk of cancer. *N Engl J Med* 320 (To the Editor):1014, 1989

(4) Korn EL, Graubard BI. Epidemiologic studies utilizing surveys: Accounting for the Sampling Design. *Am J Public Health* 81:1166-1173, 1991

(5) Miller HW. Plan and operation of the Health and Nutrition Examination Survey. *Vital Health Stat* 1(10a). Hyattsville, MD: National Center for Health Statistics, 1985

(6) Landis JR, Lepkowski JM, Eklund SA et al. A statistical methodology for analyzing data from a complex survey: the First National Health and Nutrition Examination Survey. *Vital Health Stat* 2(92). Hyattsville, MD: National Center for Health Statistics, 1982

(7) Cohen BB, Barbano HE, Cox CE, et al. Plan and operation of the HNANES I Epidemiologic Follow-up Study: 1982-1984. *Vital Health Stat* 1(22). Hyattsville, MD: National Center for Health Statistics, 1987

(8) Wolter KM. *Introduction to Variance Estimation.* New York, NY: Springer-Verlag, 1985

(9) Russell-Briefel R, Dresser CM, Ezzati TM, et al. Plan and operation of the Hispanic Helath and Nutrition Examination Survey 1982-1984. *Vital Health Stat* 1(19). Hyattsville, MD: National Center for Health Statistics, 1985

(10) Cornfield J. Randomization by groups: a formal analysis. *Am J Epidemiol* 108:100-102, 1978

(11) Korn EL, Graubard BI. Simultaneous testing of regression coefficients with complex survey data: use of Bonferroni t-statistics. *The American Statistician* 44:270-276, 1990

(12) Cochran WG. The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A* 128:234-255, 1965

(13) Korn EL, Graubard. Examining neighborhood confounding in a survey: an example using the National Health and Nutrition Survey II. *Stat Med* 7:1087-1098, 1988

(14) DuMouchel WH, Duncan GJ. Using sample survey weights in multiple regression analysis of stratified samples. *J Am Stat Assoc* 78:535-543, 1983

(15) Fuller WA. Least squares and related analyses for complex survey designs. *Survey Methodology* 10:97-118, 1984