Epidemiologic Studies Utilizing Surveys: Accounting for the Sampling Design

Edward L. Korn, PhD, and Barry I. Graubard, MA

ABSTRACT

Background. Since large-scale health surveys usually have complicated sampling schemes, there is often a question as to whether the sampling design must be considered in the analysis of the data. A recent disagreement concerning the analysis of a body iron stores-cancer association found in the first National Health and Nutrition Examination Survey and its follow-up is used to highlight the issues.

Methods. We explain and illustrate the importance of two aspects of the sampling design: clustering and weighting of observations. The body iron stores-cancer data are reanalyzed by utilizing or ignoring various aspects of the sampling design. Simple formulas are given to describe how using the sampling design of a survey in the analysis will affect the conclusions of that analysis.

Results. The different analyses of the body iron stores-cancer data lead to very different conclusions. Application of the simple formulas suggests that utilization of the sample clustering in the analysis is appropriate, but that a standard utilization of the sample weights leads to an uninformative analysis. The recommended analysis incorporates the sampling weights in a nonstandard way and the sample clustering in the standard way.

Conclusions. Which particular aspects of the sampling design to use in the analysis of complex survey data and how to use them depend on certain features of the design. We give some guidelines for when to use the sample clustering and sample weights in the analysis. (*Am J Public Health.* 1991;81:1166–1173)

Introduction

Many biomedical questions involving human populations cannot be answered by experimentation. We cannot ethically randomize patients to certain risk factors, e.g., exposure to asbestos. For factors we can manipulate, we may be unwilling to wait the required number of years or to follow the large number of individuals necessary for a meaningful analysis. Epidemiologic studies offer an alternative to randomized experiments that can avoid the major ethical problems and sometimes allow reduced sample sizes and earlier completion times.

Depending on the risk factors and outcomes, epidemiologic studies may still require large sample sizes. Large multipurpose health surveys can be a readily available source of data. An advantage of this type of data is that many risk factors can be examined. For example, consider the National and Hispanic Health and Nutrition Examination Surveys—NHANES I, II, III and HHANES—conducted in 1971–1975, 1976–1980, 1988–1994, and 1982–1984, respectively. Table 1 gives some examples of cross-sectional associations that have been studied using these surveys. Also, because these surveys were done at different times, one can examine cross-sectionally changes over time in various measured quantities (Table 1). In 1982–1984, a follow-up survey was conducted of NHANES I sampled individuals aged 25 to 74 years. This follow-up survey allows true longitudinal studies to be conducted (see Table 1 for examples). With diminishing financial resources, we believe that these types of surveys will be used increasingly to investigate biomedical and public health hypotheses.

A question that usually arises in the analysis of such survey data is whether the complex sampling design, which involves the clustering and weighting of observations, needs to be accounted for, or if it can be ignored by treating the sampled data as the population of interest. In some[5,7–12,15–20,22,23,25–36] of the above mentioned studies the sampling design was used in the analysis, while in others[1–4,6,13,14,21,24,37–53] it was not. The recent analysis by Stevens et al.[41] that ignored the sampling design in the analysis crystallizes the issue. Among other findings, Stevens et al.[41] found statistically significant differences in the mean total iron-binding capacity and mean transferrin saturation between men who developed cancer and men who did not, controlling for age and smoking status. Yip and Williamson[54] criticized this study because the sampling design was not used, and suggested that if it had been incorporated in the analysis, "it is doubtful that [these differences] would have remained 'statistically significant'." Stevens and colleagues replied, "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 US population at large. To test for differences between case patients and controls within the NHANES cohort, however, the methods we used are appropriate."[55] What is the correct approach to the analyses of these types of studies?

Edward L. Korn is with the Biometric Research Branch and Barry I. Graubard is with the Biometry Branch, National Cancer Institute.

Requests for reprints should be sent to Edward L. Korn, Biometric Research Branch, National Cancer Institute, Executive Plaza North, Room 739, Bethesda, MD 20892.

This paper was submitted to the journal August 23, 1990, and accepted with revisions April 2, 1991.

TABLE 1—Examples of Studies Using Data from the National Health and Nutrition Examination Surveys

Outcome	References	Risk Factors/Subgroups
Cross-sectional associations		
Blood pressure	1–8	Nutrient intake; alcohol intake; blood lead levels; anthropometric measurements; race, sex, and age
Body weight	9–10	Smoking; alcohol
Anxiety	11	Aspects of diet
Periodontal disease	12	
Tooth decay	13	
General health	14	
Serum cholesterol	15	
Dietary supplements	16	
Blood iron levels	17	
Dietary pattern	18	
Constipation	19	
Glucose tolerance	20	Oral contraceptives; job characteristics
Cardiovascular disease	21–22	
Blood iron	23–24	Race, sex, and age
Blood lead	25	
Serum vitamin A	26	
Serum retinol	27	
Serum α-tocopherol	28	
Dietary intake	29	
Low birth weight	30	
Changes over time measured cross-sectionally		
Blood pressure	31	
Blood lead levels	32	
Anthropometric measurements	33	
Oral contraceptive use	34	
Incidence of childhood asthma	35 36	
Longitudinal studies		
Incidence of cancer	37–44	Alcohol intake; serum cholesterol; dietary fat; anthropometric measurements; blood iron; bowel function; physical activity; depression
Cardiovascular disease mortality	45	Cardiac size
Mortality from injuries	46	Alcohol consumption
Mortality (all causes)	47–50	Diabetes; race; antihypertensive drugs; self-rated health status
Depression	51	Physical activity
Hip fracture	52	Anthropometric measurements
Psychological well-being	53	Age

This article offers a brief tutorial about the sampling design and analysis of large household surveys. We examine whether the sampling design should be used when analyzing cross-sectional or longitudinal data derived from such surveys. We do not consider case-control studies, for which it is generally agreed that the sampling design must be considered in the analysis.[56–58] We consider conditions under which the sampling design can be safely ignored and the inefficiency resulting from utilizing the sampling design when it was unnecessary. The NHANES I and its follow-up are used as illustrations throughout, with special emphasis on the body iron stores-cancer association found by Stevens et al.[41]

Design of Complex Surveys: Cluster Sampling and Sample Weights

In this section the NHANES I design is used to describe the complex sampling design of household surveys. We stress the points of the design relevant to the issues discussed in this paper; other sources[59–61] provide further details. Although simple random samples are the easiest to analyze, they are impractical for large-scale national surveys for two reasons. First, no national registry of names to be sampled may exist. Second, even if a simple random sample of names could be obtained, it would be impractical to travel to, say, 20 000 locations scattered throughout the United States to interview the sampled individuals. Multistage survey designs, on the other hand, enable sampling frames to be developed for a relatively small number of areas and limit the number of locations to which interviewers must travel. For example, in the first stage of NHANES I sampling, the mainland United States was divided into approximately 1900 geographic areas or primary sampling units (PSUs), each one consisting of a standard metropolitan statistical area or, at most, three contiguous counties. These PSUs were grouped into 40 strata from each of which one or two PSUs were sampled; some very large PSUs were sampled with certainty. In each of these sampled PSUs, successive stages of sampling involved census enumeration districts (approximately 300 households), segments (approximately 8 contiguous housing units), and sampled individuals.

At first glance accounting for this type of multistage clustering in the analysis of survey data seems a formidable task. Fortunately, approximate analytic methods have been developed that require only the knowledge of the stratum and PSU of each individual.[62] In practice, computer data files from large surveys contain this information so that it can be readily accessed by statistical computer packages. For NHANES I, the design can be approximated[60] for the statistical analysis as a selection of 70 PSUs, two from each of 35 strata.

For complex surveys, each sampled individual with data has a sample weight associated with his or her 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 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 as much as eight times more often than people living elsewhere. Additionally, persons aged 65 years or older were sampled twice as often as women ages 20–44, who in turn were sampled twice as often as other adults.

TABLE 2—Adjusted Mean Values[a] for Total Iron-Binding Capacity for Subjects in Whom Cancer Developed and Those Who Remained Cancer-free

	Clustering Used?	Sample Weights Used?	Cancer[b]	No Cancer[c]	Difference Mean±SE	Difference 95% CI	Difference P
				Men			
A[d]	No	No	61.30	62.78	−1.48±0.60	−2.65 to −0.31	.013
B	No	Yes	62.65	63.57	−0.92±1.06	−2.99 to 1.15	.38
C[e]	Yes	No	61.30	62.78	−1.48±0.54	−2.58 to −0.38	.010
D[e]	Yes	Yes	62.65	63.57	−0.92±0.86	−2.67 to 0.83	.29
E	Yes	Partly[f]	62.07	63.51	−1.44±0.53	−2.51 to −0.37	.010
				Women			
A[d]	No	No	66.32	66.40	−0.08±0.79	−1.62 to 1.46	.92
B	No	Yes	65.60	66.32	−0.73±1.18	−3.04 to 1.58	.54
C[e]	Yes	No	66.32	66.40	−0.08±0.69	−1.48 to 1.33	.91
D[e]	Yes	Yes	65.60	66.32	−0.73±1.29	−3.34 to 1.88	.58
E	Yes	Partly[f]	66.20	66.35	−0.16±0.60	−1.38 to 1.06	.82

Note. SE = standard error; CI = confidence interval.

[a]Means are adjusted for age and smoking status. For the analysis labeled E, means are also adjusted for race (White vs non-White), family income (<$3000, $3000–$6999, $7000–$9999, $10 000–$14 999, ≥$15 000), poverty vs nonpoverty census enumeration district, and senior status (age ≥65 or <65). For the 3.8% of the sample whose family income was unavailable, income was imputed from education of the head of household in the same manner used in the construction of the sample weights.

[b]For men, N = 242; for women, N = 203.

[c]For men, N = 3116; for women, N = 5165.

[d]Analyses labeled A are from Stevens et al.[41] Adjusted means for cancer and no cancer groups are slightly different here because their analysis implicitly assumed that all four smoking categories were equally likely. Analyses here use observed proportions in each smoking category (analyses A and B) or estimated population proportions in each category (analyses C, D, and E).

[e]Mean estimates for analysis C and D are the same as estimates for analysis A and B, respectively.

[f]Analyses labeled E are unweighted regressions with means adjusted for many of the variables used in defining sample weights (see footnote a).

Finally, in a small number of segments there were more households encountered by the interviewers than expected. These households were subsampled at rates of one half to one quarter.

The second component of the sample weight is an adjustment for nonresponse, including both the inability to locate sampled individuals and their refusal to participate. Although there is no ideal approach to adjust for nonresponse, adjustment using weights is as follows. The population is divided into groups so that the probability of nonresponse is thought to be roughly similar for all individuals within a group. The data from respondents is then weighted upwards depending on the nonresponse rate of their corresponding group. For example, in NHANES I the nonresponse groups were based on five family income groups within each PSU.

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, because not all sampled persons in NHANES I had their body iron stores measured, a special set of weights was derived for use in the analysis of these data.

Reanalysis of the Body Iron Stores-Cancer Association

In this section we follow the analyses given in Stevens et al.[41] with the exception of using the sample clustering and/or the sample weights. We demonstrate how the conclusions of the analysis change when these two aspects of the sampling design are incorporated. We restrict attention to the associations described by Stevens et al.[41] of total iron-binding capacity (mol/L) and transferrin saturation (%) with the development of cancer. Briefly, the data to be analyzed are from 3358 men and 5368 women who (1) had their total iron-binding capacity measured in NHANES I in 1971–1975, (2) had their cancer status determined in the follow-up survey in 1981–1984, and (3) were alive and cancer-free for at least 4 years after their biochemical measurements. These sample sizes are slightly different from those of Stevens et al.[41] because the follow-up public use data files have been revised since their analyses.

Although there are many ways to analyze this type of follow-up data, we follow the simplest analysis given by Stevens et al.,[41] which compares the mean biochemical values in those respondents who developed cancer and those who did not, adjusted by linear regression for age (years) and smoking status (current smoker, former smoker, never a smoker, and unknown). Tables 2 and 3 contain the results of five different analyses of these same associations, labeled A through E.

The A analyses repeat the Stevens et al.[41] analyses. We conclude from the A analyses that for both variables there is a small but statistically significant association for men and no association for women.

The B analyses use the sample weights of the survey but not the clustering. A dramatic effect of using the weights is the increase in the standard errors (SEs), the most extreme case being the change from 0.79 to 2.53 for the SE of the difference in adjusted means for the women's transferrin saturation. The correct interpretation for this particular weighted analysis is not that there is no association between transferrin saturation and the development of cancer in women. Instead, it is that because the SE and confidence interval (CI) are so large, the data on the association are uninformative.

The C and D analyses use the sample clustering, without and with the sample weights, respectively. Only the SEs, not the means, are possibly affected by use of the clustering. For these particular analyses, it turns out that the standard errors are not much affected. The results and conclusions, therefore, are very similar to the unclustered A and B analyses, respectively.

We recommend the E analyses for several reasons. They are unweighted analyses that use the sample clustering, but for which the adjusted means have been further adjusted by linear regression for many of the variables used in defining the sample weights (see footnote a of Table 2). The results and conclusions of the E analyses turn out to be very similar to the unweighted unclustered A analyses. This happens to be true even though these analyses do not ignore the sampling weights and clustering.

The computer programs REG[63] and SURREGR[64] were used to perform the analyses A and B–E, respectively. Because 3 of the PSUs have no total iron-binding capacity data available, for the C, D, and E analyses the strata containing these PSUs are pooled with neighboring strata to yield a design approximated by 67 PSUs selected from 32 strata. The analyses A–E give an idea of the different conclusions that are possible when using various aspects of the sample design in the analysis. In the sections below, we give

some rules of thumb for choosing the most appropriate analyses based on the sample survey design alone.

Should Sample Clustering Be Used in Analysis?

Sample clustering can affect the variability (SEs) of estimates of associations. First, a simple but extreme example will show this. Consider a simple random sample of 1000 households, where disease and risk factor status are obtained from each member of the household. Suppose that if one member of a household has the risk factor, then all members of the household have it, and if one member of the household has the disease, then all have it. The appropriate sample size for the analysis of this sampled data is 1000, not the number of individuals sampled. Thus, for example, if 4000 individuals had been sampled in the 1000 households, and the sample (household) clustering was ignored in the analysis, then SEs would be too small by a factor of 2 ($= \sqrt{4000/1000}$), because SEs are inversely proportional to sample sizes. Admittedly, this is an extreme example to make the point; in general the effect of clustering depends on both the cluster size and the intraclass correlation.[62]

In many realistic situations, the "disease" may be a continuous outcome (e.g., blood pressure) and the risk factor may also be a continuous variable (e.g., dietary calcium intake). Additionally, the risk factor-disease association of interest may be the one in which certain covariates are controlled for, e.g., sex and age. A general sufficient condition for ignoring the clustering in the analysis is the distribution of the outcome for given levels of the risk factor and covariates does not depend on which cluster the individual is in. Given the complicated multistage sampling in a survey such as NHANES I, it may not be obvious whether this condition is satisfied or not.

Since adjusting for the sample clustering will on average increase the SEs, its use in the analysis can be thought of as a conservative procedure. What is lost by using the clustering in the analysis when it was unnecessary? Unnecessary clustering leads to an inefficient analysis, that is, an analysis with CIs wider than necessary and with less statistical power for rejecting the null hypothesis. An easy way to express this inefficiency is in terms of relative sample sizes. For example, an analysis that is 20% inefficient with respect to another analysis means that a sample size of 1000 in the inefficient analysis is equivalent to a sample size of 800 in the more efficient analysis. A key number for estimating the inefficiency of using clustering when unnecessary is the degrees of freedom, d, available for the estimation of SEs. The d is usually taken to be the number of sampled PSUs minus the number of strata. For example, in the NHANES I analysis of body iron stores and cancer, d = 35 = 67 − 32.

The approximate inefficiency in doing an analysis that uses the sample clustering when unnecessary is:[65]

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

where $z^{1-\alpha/2}$ and $t_d^{1-\alpha/2}$ are the $1 - \alpha/2$ percentiles of a normal distribution and a t-distribution with d degrees of freedom, respectively, and α is the significance level for testing or one minus the confidence level for CIs. For example, for the analyses of NHANES I given in the previous section in which we used 95% CIs for the mean differences, suppose we are also interested in hypothesis testing at the α = .05 level. Because d = 35, any standard set of statistical tables gives $z^{0.975} = 1.96$ and $t_{35}^{.975} = 2.03$. The inefficiency equals 6.8%, an inefficiency so small that one should use the clustering in the analyses. If the clustering is important, it must be used so that the variability of means is properly estimated; if the clustering is unimportant, little is lost by using it unnecessarily. The similar magnitude of the SEs for analyses A and C in Tables 2 and 3 suggest that, in fact, clustering was not important for these data.

In general, the inefficiency of using the sample clustering when unnecessary becomes greater with decreasing degrees of freedom. For example, the design of HHANES can be approximated[66] by sampling 16 PSUs from 8 strata. For this design, the inefficiency is 28% when the clustering is unnecessary (using d = 8 in the above formula), making the decision of whether or not to use the clustering in the analysis more difficult than with NHANES I. These inefficiency calculations are meaningful whenever a single disease-risk factor association is being examined. For simultaneous analysis of many disease-risk factor associations or other multivariate analyses, the inefficiency of using the clustering when unnecessary can increase.[67]

TABLE 3—Adjusted Mean Values[a] for Transferrin Saturation (%) for Subjects in Whom Cancer Developed and Those Who Remained Cancer-free

	Clustering Used?	Sample Weights Used?	Cancer[b]	No Cancer[c]	Difference Mean±SE	Difference 95% CI	Difference P
				Men			
A[d]	No	No	33.15	30.73	2.42±0.78	0.89–3.94	.002
B	No	Yes	32.55	30.43	2.12±1.48	−0.78–5.03	.15
C[e]	Yes	No	33.15	30.73	2.42±0.86	0.68–4.16	.008
D[e]	Yes	Yes	32.55	30.43	2.12±1.22	−0.35–4.59	.09
E	Yes	Partly[f]	32.92	30.48	2.44±0.85	0.71–4.17	.007
				Women			
A[d]	No	No	27.98	27.17	0.81±0.79	−0.73–2.36	.30
B	No	Yes	31.37	27.74	3.63±2.53	−1.32–8.58	.15
C[e]	Yes	No	27.98	27.17	0.81±0.76	−0.72–2.35	.29
D[e]	Yes	Yes	31.37	27.74	3.63±2.42	−1.28–8.53	.14
E	Yes	Partly[f]	28.36	27.57	0.78±0.74	−0.72–2.28	.30

Note: SE = standard error; CI = confidence interval.

[a]Means are adjusted for age and smoking status. For the analysis labeled E, means are also adjusted for race (White vs non-White), family income (under $3000, $3000–$6999, $7000–$9999, $10 000–$14 999, ≥$15 000), poverty vs nonpoverty census enumeration district, and senior status (age ≥65 or <65). For the 3.8% of the sample whose family income was unavailable, income was imputed from education of the head of household in the same manner used in the construction of the sample weights.

[b]For men, N = 232; for women, N = 197.

[c]For men, N = 3058; for women, N = 5073.

[d]Analyses labeled A are from Stevens et al.[41] Adjusted means for cancer and no cancer groups are slightly different here because their analysis implicitly assumed that all four smoking categories were equally likely. Analyses here use observed proportions in each smoking category (analyses A and B) or estimated population proportions in each category (analyses C, D, and E).

[e]Mean estimates for analysis C and D are the same as the estimates for analysis A and B, respectively.

[f]Analyses labeled E are unweighted regressions with means adjusted for many of the variables used in defining sample weights (see footnote a).

Should Sampling Weights Be Used in Analysis?

It is well known that the unweighted mean of a sample that is not a simple random sample can lead to a biased estimate

TABLE 4—Percentiles of Sample Weights for Individuals Analyzed

Sex	Minimum	5%	25%	Median	75%	95%	Maximum
Men (n = 3358)	477	876	3614	8170	19 504	31 556	135 824
Women (n = 5368)	611	1141	3162	6664	11 119	24 090	186 062

of the population mean, whereas a weighted mean provides an approximately unbiased estimate. It is not as obvious that an unweighted analysis of a disease-risk factor association can also lead to the wrong conclusions. A simple numerical example given in the Figure shows this to be true. The expected numbers of sampled individuals are given in the Figure for two sampling schemes for the whole population was well as within two subgroups defined by area of residence. On the left, a study using a simple random sample of 10 000 individuals would show on average the correct result that an association exists between the disease and the risk factor: The probability of having the disease is 5.9% without the risk factor and 12.2% with the risk factor, for a relative risk of approximately 2. On the right, a study of the same population that samples nonurban area residents at nine times the rate of urban area residents would on average sample the same number of individuals from each type of area. Unweighted analysis of data from this probability sample suggests incorrectly that there is no association between the disease and the risk factor. A weighted analysis of the data from this probability sample would on average yield the correct association:

$$2.06 = \frac{(9\times245 + 655)/[(9\times245 + 655) + (9\times4655 + 3445)]}{(9\times15 + 85)/[(9\times15 + 85) + (9\times85 + 815)]}$$

A weighted analysis will provide an approximately unbiased estimate of the population association, whereas an analysis ignoring the weights, in general, will not. A general sufficient condition for when the sampling weights can be ignored, however, is as follows: The distribution of the outcome (disease) for given levels of the risk factor and covariates does not depend on any variables used in the sampling design or used to adjust for nonresponse.

Because it may be difficult to verify this sufficient condition, what is the harm in always using the sample weights in the analysis? As with the use of clustering when unnecessary, the use of sampling weights when unnecessary leads to an inefficient analysis. This inefficiency can be defined as

$$1 - (SE_{unwtd}/SE_{wtd})^2$$

where SE_{unwtd} and SE_{wtd} are the unweighted and weighted SEs of the mean difference. For the linear regression models used in the analyses here, the calculation of the efficiency depends on the distribution of the sample weights and independent variables (see DuMouchel and Duncan[68] for the exact formulas). An approximation is given by the following formula:

$$\text{Inefficiency} = 1 - \frac{(w_1 + w_2 + \ldots + w_N)^2/}{[N(w_1^2 + w_2^2 + \ldots + w_N^2)]}$$

where $w_1, w_2, \ldots, w_N$ are the sample weights for the N sampled individuals. The inefficiency increases when the sample weights are more variable, and is zero if all the sample weights are identical.

Table 4 presents selected percentiles of the distribution of sample weights for the individuals analyzed here. The range of the weights is very large, with the largest weight 285 (300) times greater than the smallest weight for the men (women). The inefficiencies of using the sample weights when unnecessary are calculated to be 47% and 48% for men and women, respectively. Considering that the unweighted SEs are not insubstantial compared to clinically meaningful treatment differences, these inefficiencies are too large for us to assert that little is lost by using the weights when unnecessary.

There is an additional consideration when using the sample weights in the analysis. If the sample weights do matter, they can make the analysis even less efficient than described above. For example, consider the analysis of the transferrin saturation in women given in Table 3. Comparing analysis A to analysis B, the SE has increased by a factor of 3.2 = 2.53/0.79, suggesting an inefficiency of 90%. A closer look at the data for this analysis is instructive. Out of the 197 women who developed cancer, the woman with the highest transferrin saturation (68.4%) also had the third largest sample weight (103 042) out of all 5368 women. If this woman's sample weight had had a median value of 6684, then the mean difference in the weighted analysis B would have been 1.35 ± 1.16 instead of 3.63 ± 2.53.

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 the variables utilized in determining the sample weights. In the present context, the E analyses in Tables 2 and 3 compute the means adjusted for many of these variables. One could argue[69] that because these additional variables (family income, race, etc.) are highly unlikely to be on the causal pathway between body iron stores and the development of cancer, it would be useful to control for these variables even if we had a simple random sample. We view this not as the perfect solution, but as a compromise that avoids the inefficiency of a weighted analysis. As it happens, the results of the E analyses are remarkably similar to the unweighted, unclustered A analyses, but, we can now have more confidence in these results. However, for other data sets and other associations this may not be the case.

Discussion

The impact of sampling design on the analysis of complex survey data should always be considered. Which particular aspects of the design to use in the analysis, however, is a subtler question. Classical survey analysts recommend using both the clustering and the sample weights in the analysis.[70,71] Although this is a philosophically defensible position, we have seen above that it can lead to inferences that make the whole study worthless. On the other hand, the position of Stevens et al.[55] that the sampling design can be ignored when addressing these types of questions is tenuous unless one is willing to make strong assumptions. We have attempted in this article to give simple ways of weighing the costs and benefits of accounting for the sampling design in the analysis. For specific recommendations concerning sample clustering and sample weights, we offer the following rules of thumb.

(1) If the number of sampled PSUs minus the number of strata is greater than or equal to 20 (corresponding to an inef-

SIMPLE RANDOM SAMPLING

(N= 10000 sampled from entire population)

Totals

		Disease -	Disease +	Proportion with Disease
Risk	-	9068	572	5.9%
Factor	+	316	44	12.2%
		Total 10000		

Subgroup 1 (Urban Area)

		Disease -	Disease +	
Risk	-	8379	441	5.0%
Factor	+	153	27	15.0%
		Total 9000		

Subgroup 2 (Non-Urban Area)

		Disease -	Disease +	
Risk	-	689	131	16.0%
Factor	+	163	17	9.4%
		Total 1000		

PROBABILITY SAMPLING

(N= 5000 sampled from each subgroup)

Totals

		Disease -	Disease +	Proportion with Disease
Risk	-	8100	900	10.0%
Factor	+	900	100	10.0%
		Total 10000		

Subgroup 1 (Urban Area)

		Disease -	Disease +	
Risk	-	4655	245	5.0%
Factor	+	85	15	15.0%
		Total 5000		

Subgroup 2 (Non-Urban Area)

		Disease -	Disease +	
Risk	-	3445	655	16.0%
Factor	+	815	85	9.4%
		Total 5000		

FIGURE—unweighted analysis of data from a probability sample can lead to incorrect conclusions.

ficiency of roughly 10%), then use the clustering in the analysis. Otherwise, more advanced statistical methods[72] to account for the clustering can be applied to roughly approximate the SEs of the parameters of interest.

(2) Calculate the approximate inefficiency of including the sample weights using the simple formula given in the previous section. If this inefficiency is less than 10%, then use the sample weights in the analysis. If this inefficiency is greater than 10%, then consider the effect this inefficiency will have on the SEs compared to an expected clinically meaningful treatment difference. If the effect is not unacceptably large, use the sample weights in the analysis. Otherwise, use an "unweighted" analysis that controls for the variables relating to the design and nonresponse adjustments in the analysis.

We end with two suggestions for the designers and producers of large-scale health surveys that will be used by other investigators. If our recommendations are followed, one will sometimes need to use variables relating to the design and nonresponse 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 PSUs 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 policymakers provide sufficient resources so that the surveys can be designed for efficient secondary analyses using the design. □

Acknowledgments

We are indebted to L.S. Freedman, S.B. Green, D. Hitchcock, M.G. Kovar, J.L. Mills, B.H. Patterson, and R.G. Stevens for their helpful comments.

References

1. Blair D, Habicht JP, Sims EAH, Sylwester D, Abraham S. Evidence for an increased risk for hypertension with centrally located body fat and the effect of race and sex on this risk. *Am J Epidemiol.* 1984;119:526–540.
2. Frisancho AR, Leonard WR, Bollettino LA. Blood pressure in blacks and whites and its relationship to dietary sodium and potassium intake. *J Chron Dis.* 1984;37:515–519.
3. Stanton JL, Braitman LE, Riley AM, Khoo CS, Smith JL. Demographic, dietary, life style, and anthropometric correlates of blood pressure. *Hypertension.* 1982;4(suppl III):135–142.
4. Gruchow HW, Sobocinski KA, Barboriak JJ. Alcohol, nutrient intake, and hypertension in US adults. *JAMA.* 1985;253:1567–1570.
5. Harlan WR, Hull AL, Schmouder RL, Landis JR, Thompson FE, Larkin FA. Blood pressure and nutrition in adults. *Am J Epidemiol.* 1984;120:17–28.
6. McCarron DA, Morris CD, Henry HJ, Stanton JL. Blood pressure and nutrient intake in the United States. *Science.* 1984;224:1392–1398.
7. Sempos C, Cooper R, Kovar MG, Johnson C, Drizd T, Yetley E: Dietary calcium and blood pressure in NHANES I and II. *Hypertension.* 1986;8:1067–1074.
8. Harlan WR, Landis JR, Schmouder RL, Goldstein NG, Harlan LC. Blood lead and blood pressure. *JAMA.* 1985;253:530–534.
9. Albanes D, Jones Y, Micozzi MS, Mattson ME. Associations between smoking and body weight in the US population: analysis of NHANES II. *Am J Public Health.* 1987;77:439–444.
10. Williamson DF, Forman MR, Binkin NJ, Gentry EM, Remington PL, Trowbridge FL. Alcohol and body weight in United States adults. *Am J Public Health.* 1987;77:1324–1330.
11. Eaton WW, McLeod J. Consumption of coffee or tea and symptoms of anxiety. *Am J Public Health.* 1984;74:66–68.
12. Ismail AI, Burt BA, Eklund SA. Relation between ascorbic acid intake and periodontal disease in the United States. *J Am Dent Assoc.* 1983;107:927–931.
13. Ismail AI, Burt BA, Eklund SA. The cariogenicity of soft drinks in the United States. *J Am Dent Assoc.* 1984;109:241–245.
14. Schwerin HS, Stanton JL, Riley AM, et al. Food eating patterns and health: a reexamination of the Ten-State and HANES I surveys. *Am J Clin Nutr.* 1981;34:568–580.
15. Kovar MG, Fulwood R, Feinleib M. Coffee studies. *N Engl J Med.* 1983;309 (To the Editor):1249.
16. Koplan JP, Annest JL, Layde PM, Rubin GL. Nutrient intake and supplementation in the United States (NHANES II). *Am J Public Health.* 1986;76:287–289.
17. Looker AC, Sempos CT, Johnson CL, Yetley EA. Comparison of dietary intakes and iron status of vitamin-mineral supplement users and nonusers, aged 1–19 years. *Am J Clin Nutr.* 1987;46:665–672.
18. Schectman G, McKinney WP, Pluess J, Hoffman RG. Dietary intake of Americans

reporting adherence to a low cholesterol diet (NHANES II). *Am J Public Health.* 1990;80:698–703.
19. Sandler RS, Jordan MC, Shelton BJ. Demographic and dietary determinants of constipation in the US population. *Am J Public Health.* 1990;80:185–189.
20. Russell-Briefel R, Ezzati TM, Perlman JA, Murphy RS. Impaired glucose tolerance in women using oral contraceptives: United States, 1976–80. *J Chronic Dis.* 1987;40:3–11.
21. Russell-Briefel R, Ezzati TM, Fulwood R, Perlman JA, Murphy RS. Cardiovascular risk status and oral contraceptive use: United States, 1976–80. *Prev Med.* 1986;15:352–362.
22. Karasek RA, Theorell T, Schwartz JE, Schnall PL, Pieper CF, Michela JL. Job characteristics in relation to the prevalence of myocardial infarction in the US Health Examination Survey (HES) and the Health and Nutrition Examination Survey (HANES). *Am J Public Health.* 1988;78:910–918.
23. Dallman PR, Yip R, Johnson C. Prevalence and causes of anemia in the United States, 1976 to 1980. *Am J Clin Nutr.* 1984;39:437–445.
24. Meyers LD, Habicht JP, Johnson CL. Components of the difference in hemoglobin concentrations in blood between black and white women in the United States. *Am J Epidemiol.* 1979;109:539–549.
25. Mahaffey KR, Annest JL, Roberts J, Murphy RS. National estimates of blood lead levels: United States, 1976–80. *N Engl J Med* 1982;307:573–579.
26. Looker AC, Johnson CL, Woteki CE, Yetley EA, Underwood BA. Ethnic and racial differences in serum vitamin A levels of children aged 4–11 years. *Am J Clin Nutr.* 1988;47:247–252.
27. Looker AC, Johnson CL, Underwood BA. Serum retinol levels of persons aged 4–74 years from three Hispanic groups. *Am J Clin Nutr.* 1988;48:1490–1496.
28. Looker A, Underwood B, Wiley J, Fulwood R, Sempos C. Serum α-tocopherol levels of Mexican Americans, Cubans, and Puerto Ricans aged 4–74 y. *Am J Clin Nutr.* 1989;50:491–496.
29. Block G, Rosenberger WF, Patterson BH. Calories, fat and cholesterol: intake patterns in the US population by race, sex, and age. *Am J Public Health.* 1988;78:1150–1155.
30. Scribner R, Dwyer JH. Acculturation and low birthweight among Latinos in the Hispanic HANES. *Am J Public Health.* 1989;79:1263–1267.
31. Dannenberg AL, Drizd T, Horan MJ, Haynes SG, Leaverton PE. Progress in the battle against hypertension: blood pressure trends in the United States from 1960 to 1980. *Hypertension.* 1987;10:226–233.
32. Annest JL, Pirkle JL, Makuc D, Neese JW, Bayse DD, Kovar MG. Chronological trend in blood lead levels between 1976 and 1980. *N Engl J Med.* 1983;308:1373–1377.
33. Flegal KM, Harlan WR, Landis JR. Secular trends in body mass index and skinfold thickness with socioeconomic factors in young adult women. *Am J Clin Nutr.* 1988;48:535–543.
34. Harlan WR, Landis JR, Flegal KM, Davis CS, Miller ME. Secular trends in body mass in the United States, 1960–80. *Am J Epidemiol.* 1988;128:1065–1074.
35. Russell-Briefel R, Ezzati T, Perlman J. Prevalence and trends in oral contraceptive use in premenopausal females ages 12–54 years, United States, 1971–80. *Am J Public Health.* 1985;75:1173–1176.
36. Gergen RJ, Mullally DI, Evans R. National survey of prevalence of asthma among children in the United States, 1976–80. *Pediatrics.* 1988;81:1–7.
37. Schatzkin A, Jones DY, Hoover RN, et al. Alcohol consumption and breast cancer in the epidemiologic follow-up study of the first National Health and Nutrition Examination Survey. *N Engl J Med.* 1987;316:1169–1173.
38. Schatzkin A, Taylor PR, Carter CL, et al. Serum cholesterol and cancer in the NHANES I Epidemiologic Follow-up Study. *Lancet II.* 1987;298–301.
39. Jones DY, Schatzkin A, Green SB, et al. Dietary fat and breast cancer in the National Health and Nutrition Examination Survey I Epidemiologic Follow-up Study. *JNCI.* 1987;79:465–471.
40. Swanson CA, Jones DY, Schatzkin A, Brinton LA, Ziegler RG. Breast cancer risk assessed by anthropometry in the NHANES I Epidemiological Follow-up Study. *Cancer Res.* 1988;48:5363–5367.
41. Stevens RG, Jones DY, Micozzi MS, Taylor PR. Body iron stores and the risk of cancer. *N Engl J Med.* 1988;319:1047–1052.
42. Micozzi MS, Carter CL, Albanes D, Taylor PR, Licitra LM. Bowel function and breast cancer in US women. *Am J Public Health.* 1989;79:73–75.
43. Albanes D, Blair A, Taylor PR. Physical activity and risk of cancer in the NHANES I population. *Am J Public Health.* 1989;79:744–750.
44. Zonderman AB, Costa PT, McCrae RR. Depression as a risk for cancer morbidity and mortality in nationally representative sample. *JAMA.* 1989;262:1191–1195.
45. Rautaharju PM, LaCroix AZ, Savage DD, et al. Electrocardiographic estimate of left ventricular mass versus radiographic cardiac size and the risk of cardiovascular disease mortality in the Epidemiologic Follow-up Study of the first National Health and Nutrition Examination Survey. *Am J Cardiol.* 1988;62:59–66.
46. Anda RF, Williamson DF, Remington PL. Alcohol and fatal injuries among US adults. *JAMA.* 1988;260:2529–2532.
47. Kleinman JC, Donahue RP, Harris MI, Finucane FF, Madans JH, Brock DW. Mortality among diabetics in a national sample. *Am J Epidemiol.* 1988;128:389–401.
48. Otten MW, Teutsch SM, Williamson DF, Marks JS. The effect of known risk factors on the excess mortality of black adults in the United States. *JAMA.* 1990;263:845–850.
49. Havlik RJ, La Croix AZ, Kleinman JC, Ingram DD, Harris T, Cornoni-Huntley J. Antihypertensive drug therapy and survival by treatment status in a national survey. *Hypertension.* 1989;13:I28–I32.
50. Idler EL, Angel RJ. Self-rated health and mortality in the NHANES-I Epidemiologic Follow-up Study. *Am J Public Health.* 1990;80:446–452.
51. Farmer ME, Locke BZ, Moscicki EK, Dannenberg AL, Larson DB, Radloff LS. Physical activity and depressive symptoms: the NHANES I Epidemiologic Follow-up Study. *Am J Epidemiol.* 1988;128:1340–1351.
52. Farmer ME, Harris T, Madans JH, Wallace RB, Cornoni-Huntley J, White LR. Anthropometric indicators and hip fracture: the NHANES I Epidemiologic Follow-up Study. *J Am Geriatr Soc.* 1989;37:9–16.
53. Costa PT, Zonderman AB, McCrae RR, Cornoni-Huntley J, Locke BZ, Barbano HE. Longitudinal analyses of psychological well-being in a national sample: stability of mean levels. *J Gerontol.* 1987;42:50–55.
54. Yip R, Williamson DF. Body iron stores and risk of cancer. *N Engl J Med.* 1989;320 (To the Editor):1012.
55. Stevens RG, Jones DY, Micozzi MS, Taylor PR. Body iron stores and risk of cancer. *N Engl J Med.* 1989;320 (To the Editor):1014.
56. Scott AJ, Wild CJ. Fitting logistic models under case-control or choice based sampling. *J R Stat Soc B.* 1986;48:170–182.
57. Cain KC, Breslow NE. Logistic regression analysis and efficient design for two stage studies. *Am J Epidemiol.* 1988;126:1198–1206.
58. Graubard BI, Fears TR, Gail MH. Effects of cluster sampling on epidemiologic analysis in population-based case-control studies. *Biometrics.* 1989;45:1053–1071.
59. Miller HW. Plan and operation of the Health Nutrition Examination Survey. *Vital Health Stat.* 1985;1(10a). Hyattsville, Md: National Center for Health Statistics.
60. Landis JR, Lepkowski JM, Eklund SA, Stehouwer SA. A statistical methodology for analyzing data from a complex survey: the First National Health and Nutrition Examination Survey. *Vital Health Stat.* 1982;2(92). Hyattsville, Md: National Center for Health Statistics.
61. Cohen BB, Barbano HE, Cox CS, et al. Plan and operation of the NHANES I Epidemiologic Follow-up Study: 1982–84. *Vital Health Stat.* 1987;1(22). Hyattsville, Md: National Center for Health Statistics.
62. Wolter KM. *Introduction to Variance Estimation.* New York, NY: Springer-Verlag; 1985.
63. SAS Institute Inc. *SAS User's Guide: Statistics, Version 5 edition.* Cary, NC: SAS Institute; 1985.
64. Shah BV. SURREGR: *Standard Errors of Regression Coefficients from Sample Survey Data.* Research Triangle Park, NC: Research Triangle Institute; 1982.
65. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol.* 1978;108:100–102.
66. Russell-Briefel R, Dresser CM, Ezzati TM, et al. Plan and operation of the Hispanic Health and Nutrition Examination Survey 1982–1984. *Vital Health Stat.* 1985;1(19). Hyattsville, Md: National Center for Health Statistics.
67. Korn EL, Graubard BI. Simultaneous testing of regression coefficients with complex

survey data: use of Bonferroni t-statistics. *The American Statistician.* 1990;44:270–276.
68. DuMouchel WH, Duncan GJ. Using sample survey weights in multiple regression analysis of stratified samples. *J Am Stat Assoc.* 1983;78:535–543.
69. Korn EL, Graubard BI. Examining neighborhood confounding in a survey: an example using the National Health and Nutrition Examination Survey II. *Stat Med.* 1988;7:1087–1098.
70. Kish L, Frankel MR. Inference from complex samples. *J R Stat Soc* Ser. B. 1974;36:1–37.
71. Hansen MH, Madow WG, Tepping BJ. An evaluation of model-dependent and probability-sampling inferences in sample surveys. *J Am Stat Assoc.* 1983;78:776–793.
72. Pfeffermann D, LaVange L. Regression models for stratified multi-stage cluster samples. In: Skinner CJ, Holt D, Smith TMF, eds. *Analysis of Complex Surveys.* New York, NY: John Wiley & Sons; 1989.

Drug with Potential for Treating the Neuropsychological Effects of HIV Infection to Be Tested

The National Institute of Mental Health (NIMH) today announced the beginning of a clinical study of the effectiveness of Peptide T, a drug of potential value in treating the neuropsychological effects of HIV infection and AIDS.

"This is the first study designed to answer questions about the effectiveness of Peptide T by comparing the compound with a placebo in a controlled study design," said Alan I. Leshner, PhD, NIMH acting director.

The first patient started treatment with Peptide T in mid-March. The study, which is jointly sponsored by NIMH and the National Institute of Allergy and Infectious Diseases, is being conducted at the University of Southern California in Los Angeles.

The human immunodeficiency virus is believed to cause problems of concentration and memory in some persons with AIDS and AIDS-related complex by directly affecting the brain.

Peter N.R. Heseltine, MD, the principal investigator, said "in earlier studies, people with these problems have shown improvement in tests of their mental function when they were treated with Peptide T. The purpose of this new study is to discover if Peptide T is safely able to reduce these HIV-caused problems of concentration and memory and if it is safe and effective when used with the antiretroviral AZT (zidovudine)."

Ellen Stover, PhD, director of the NIMH Office of AIDS Programs, said, "this is the first time that a trial of an AIDS therapeutic will be using improvement in mental functioning as the key outcome. We are working toward making progress in treating HIV infection and are looking forward to the collaboration with USC investigators and the results of this study."

Peptide T was developed by NIMH scientists working at the institute's research facilities on the National Institutes of Health campus in Bethesda, Md. The peptide is an artificially produced protein that blocks the attachment of the AIDS virus to human cells by occupying receptors on the cell surface. Its structure is similar to part of the protein coat of HIV.

In the Phase II study, a minimum of 150 men and women who have been infected with HIV will be enrolled in the double-blind, placebo-controlled trial. (Phase II studies are designed primarily to test the efficacy of new therapies.)

The investigators are making special efforts to enroll women and minorities, who have been underrepresented in some earlier AIDS studies.

Initially, participants will receive either Peptide T or placebo for 6 months, and then during the second 6-month period, all study participants will receive the peptide. During the 1-year study, participants may take any medications approved by the Food and Drug Administration for preventing AIDS-related opportunistic infections, and they may continue treatment with antiviral therapies if they are taking them when they enter the study.

Individuals interested in participating in the study should call Charles Hovis, LAC-USC Medical Center, 213/226-4643, or the PHS-sponsored AIDS Clinical Trials Information Service, 1-800-TRIALS-A.