HOUSEHOLD SAMPLE SURVEYS OF DIABETES: DESIGN EFFECTS OF COUNTING RULES

Sirken, M.G., Inderfurth, G.P., Burnham, C.E., and Danchik, K.M.
National Center for Health Statistics

1. INTRODUCTION

In this paper, we report the findings of an experiment that tested two methods of estimating the prevalence rates of diabetes based on interviews conducted in household sample surveys. This investigation is one phase of a statistical research program in the National Center for Health Statistics to evaluate the design effects of counting rules and other design factors in health surveys [4], [5].

The alternative estimation methods that will be compared in this paper differ with respect to the items of information that are collected in the household interview. The conventional estimator of diabetes in household surveys is based on a counting rule which makes persons with diabetes eligible to be enumerated at a household only if they live there. The alternative estimation method, which will be referred to as the network estimator, requires the collection of ancillary information in the household interview. The network estimator is illustrated in this paper by a counting rule which makes persons with diabetes eligible to be enumerated at a household if either he and/or his siblings live there.

The opportunity to investigate these alternative estimators in a diabetes survey presented itself in the initial pretest of the questionnaire for the 1976 Health Interview Survey. The Health Interview Survey (HIS), one of the major components [1] [2] of the National Health Survey Program [3], assesses the health of the population on the basis of comprehensive interviews that are conducted weekly in a national household sample survey. Although the basic components of the HIS questionnaire are more or less invariant, annual revisions are made in the supplements to the basic questionnaire. After consultation with the National Commission on Diabetes as well as numerous other interested agencies and individuals, several sets of questions were incorporated into the 1976 HIS questionnaire, in an attempt to meet many of the needs for national statistics on diabetes.

In particular, the 1976 HIS questionnaire includes a series of questions that are intended to produce health statistics on the familial aggregation of diabetes in the United States. A subset of these questions is listed in the Appendix to this paper. In addition to their substantive value, questions of this type are potentially useful in improving the quality of the survey estimates of the number of persons with diabetes in the population. The network estimator, illustrated in this paper, is based on information collected by Q. 4, Q. 8a and Q. 8b. The conventional estimator is based entirely on information collected by Q. 4.

The conventional and network estimators of diabetes are described and illustrated in the next section. The pretest of the 1976 HIS questionnaire, with particular emphasis on the questions pertaining to family history of diabetes, is described in Section 3. The diabetes prevalence rates based on the conventional and the network estimators and their respective sampling errors are compared in the section on the pretest findings. A summary and the conclusions are presented in the final section.

2. CONVENTIONAL AND NETWORK ESTIMATORS

Estimation of the number of persons with diabetes in a population by means of a household sample survey may be viewed as a three step process:

(1) selecting a random sample of households,

(2) enumerating persons with diabetes at the sample households,

(3) weighting each person with diabetes enumerated in the survey by the inverse of his chance of being enumerated in the sample.

Conventional and network estimators differ with respect to (a) the counting rules adopted in step 2 for counting diabetic persons in sample households, and (b) the counting rule weights assigned in step 3 to the persons with diabetes enumerated in sample households.

Counting Rules

Every household survey adopts a counting rule that specifies conditions that make persons eligible to be enumerated at households. The counting rule adopted in a diabetes survey specifies conditions that link persons with diabetes to the households where they are eligible to be enumerated. The conventional and network estimators are based on conventional and network rules respectively.

The conventional rule in diabetic surveys seeks to specify conditions that makes every person with diabetes in the population eligible to be enumerated once and only once. For example, the de jure residence rule is the conventional rule that we are evaluating in this report. It specifies that the diabetic person is eligible to be enumerated only once at a household if and only if the household is his de jure place of residence.

On the other hand, the network counting rule in diabetes household surveys permits the same diabetic person to be enumerated at more than one household and permits him to be enumerated more than once at the same household. The multiplicity of a person is defined as the total number of times he is eligible to be enumerated. In this report, for example, we are evaluating a network rule that makes the diabetic

person eligible to be enumerated at the households of his siblings as well as his own household. The number of times that the diabetic person is eligible to be enumerated at his de jure household is equal to one more than the number of his siblings living with him. The number of times the diabetic person is eligible to be enumerated at a household which is not his de jure residence is equal to the number of his siblings living there. According to this rule, the multiplicity of a diabetic person would be equal to the number of his living siblings plus one.

Counting Rule Weights

Every diabetic person enumerated in the survey is assigned a weight which is the product of the sampling weight and the counting rule weight.

The sampling rule weight is the inverse of the probability of selecting the sample household in which the diabetic person was enumerated. This weight is determined entirely by the sample design of the survey, and it is independent of the counting rule weight.

The counting rule weight assigned to the person with diabetes enumerated in a sample household is the ratio of the number of times the person is eligible to be enumerated in the household divided by the multiplicity of the person. The counting rule weight assigned to a person depends, of course, on the particular counting rule adopted in the survey. For instance, the conventional counting rule weight is always equal to one since the conventional counting rule permits every diabetic person to be enumerated once and only once. Hence, these weights are known a priori. Since the network counting rule weights are usually unknown, they are determined on the basis of ancillary information collected from the sample household where the diabetic persons are enumerated in the survey.

An Example.

We are now ready to compare the information that would be collected in a conventional and a network diabetes survey. In the conventional survey based on the de jure residence rule, which makes the person himself the only eligible reporter, every person living in a sample household is asked the following question:

Q.A "Do you have diabetes?"

In the network survey based on the counting rule that makes siblings of the diabetic person as well as the diabetic himself eligible reporters, every person living in the sample household would be asked Q.1 and also Q.2 and possibly question Q.3.

Q.B "How many living brothers and sisters do you have?"

If the answer to Q.2 is "one or more," the network survey asks:

Q.C "How many of these siblings have diabetes?"

Assume the following set of responses to Q.A-Q.C in a single person household:

Q.A No.

Q.B Three.

Q.C One.

Since the reply to Q.A is "No", no persons with diabetes would be enumerated in this household according to the conventional rule. Since the reply to Q.A is "No" and to Q.C is "One", one diabetic person would be enumerated in this household in compliance with the network rule. This person would be assigned a counting rule weight of one-fourth, because the reply to Q.B indicates that there are three living siblings in addition to the person himself or a total of four persons who are eligible to report the diabetic person.

3. THE PRETEST

The information which we will use to compare the alternative survey estimators was obtained as a by-product of the first pretest of the questionnaire that was being developed for the 1976 Health Interview Survey. The set of questions listed in the Appendix was completed for each member of the household. Persons eligible to be enumerated by the conventional and network counting rules are identified by Q.4 and by Q.4 and Q.8(b) respectively. The counting rule weights required by the network estimator are determined by Q.8(a). The other listed questions in the Appendix are not germane to the particular estimators being investigated in this report. However, the information collected by Q.8(f) and Q.8(h) was used to label the diabetic persons identified by Q.4 and Q.8(b) as either children of diabetics or of non-diabetic parents. Consequently, in the subsequent analysis, we are able to compare the diabetes rates of children of diabetic and non-diabetic parents based on the conventional and the network estimators.

The standard interviewing procedures of the Health Interview Survey including the standard respondent rules were applied in the pretest. The respondent rules in household surveys define the preferred and the eligible household respondents. According to the standard HIS respondent rules, the preferred household respondents are: (1) the parents or guardians of the child under seventeen, (2) the person himself, if he is older than eighteen, and (3) either the parent or the person himself, if the person is seventeen or eighteen. In the event that the preferred respondent is not-at-home when the interviewer visits, the standard HIS respondent rule specifies that any other related adult in the household is eligible to serve as the proxy respondent for the absent person.

In the pretest, conducted in York, Pennsylvania, interviews were completed with 187

households in which 570 persons were enumerated. Approximately one-third represented lower income and one-third represented black families. Three household questionnaires representing 20 persons were discarded due to insufficient information in the diabetes section, leaving a total of 184 households and 550 persons. For about two percent of these persons, items of information were missing in the diabetes section due primarily to interviewer omissions. These blanks were edited on the basis of other information available on the questionnaire. There remained, however, some persons for which items of information about siblings and/or parents were not ascertained from the respondent. These persons are distributed in Table 1 by the item of nonresponse.

Selected findings from Table 1 are summarized below. Nonresponse rates vary by item of information. The rate is higher to Q.8h "Does (did) your father have diabetes?" (5.3%) than to Q.8f "Does (did) your mother have diabetes?" (2.0%) or to Q.8d "How many of your living siblings have diabetes?" (1.1%). The item nonresponse rates also vary by the respondent status. For both males and females 20 years and over, the item nonresponse rates are consistently higher for proxy than for self respondents. For proxy respondents, the item nonresponse rates are consistently larger for females than for males. Nevertheless, the item nonresponse rates are about the same for males and for females because the proportion of self respondents was substantially greater for females (90 percent) than for males (40 percent).

4. PRETEST FINDINGS

The pretest produced two sets of diabetes rates - one based on the network estimator, the other based on the conventional estimator. The numerators of the conventional and network rates are based on their respective estimators. However, the denominators of both sets of rates are based on the conventional estimator.

Rates of diabetes were estimated from the pretest for the total population and for the seven population subdomains that are listed in the stub of Table 2. An upper and lower bound is given in Table 2 for each rate based on the network counting rule. The bounds define an interval of uncertainty resulting from item nonresponse to Q.8d, the number of living siblings with diabetes. The lower and upper bound estimates, respectively, assume that none and all of the siblings for which this question was not answered have diabetes. The true estimate based on the network counting rule corrected for item nonresponse lies somewhere in the interval but probably much closer to the lower than to the upper bound. Hence, in the subsequent discussion we will refer to the lower bound of the network estimate in making a comparison with the conventional estimate.

The estimated diabetes prevalence rate is 27.3 per 1,000 persons based on the conventional estimator, or about 10 percent larger than that estimated rate (24.6 per 1,000 persons) based on the network estimator. However, the estimated rates based on the conventional estimator are not consistently larger than the estimated rates based on the network estimator. For example, the conventional estimate is larger than the network estimate for persons 25 years and over, but the reverse applies for persons under 25 years. (Since the ages of siblings with diabetes were collected in the pretest only if they resided in

Table 1. Item Nonresponse Rates by Age, Sex, and Respondent Status.

Age, Sex and Respondent Status	Sample Size	Questions		
		Q.8d Number of living siblings with diabetes?	Q.8f Does (did) mother have diabetes?	Q.8h Does (did) father have diabetes?
Total	550	1.1%	2.0%	5.3%
Male	254	0.8	2.8	5.9
Female	296	1.4	1.4	4.7
Under 20 years	211	0.9	0	4.7
Male	103	1.0	0	5.8
Female	108	0.9	0	3.7
Twenty years and over	339	1.2	3.2	5.6
Male	151	0.7	4.6	6.0
Self respondent	74	0.0	2.7	4.1
Nonself respondent	77	1.3	6.5	7.8
Female	188	1.6	2.1	5.3
Self respondent	168	1.2	1.2	4.8
Nonself respondent	20	5.0	10.0	10.0

the sample household, the ages of the siblings who lived elsewhere were inferred from the ages of their siblings who lived in the sample household.) The diabetes rates based on the conventional estimator are larger for persons who did not have a diabetic parent and for persons who were self respondents in the survey. On the other hand, the diabetes rates based on the network estimator are larger for persons who have a diabetic parent and for persons who did not respond for themselves. The ratio of the diabetes rates of persons whose parents do and do not have diabetes is substantially higher based on the network estimator (2.2) than on the conventional estimator (1.4). However, the sample sizes are small, and none of the differences between conventional and network estimates of diabetes rates noted above are statistically significant.

The relative standard errors of the network estimates are consistently smaller than those of the conventional estimates. The two sets of standard errors are also compared in Table 2. The set of figures in the last column of this table shows the design effect of using the conventional estimator instead of the network estimator. Each figure, representing the squared ratio of the relative standard error of the conventional to the network estimate, indicates how much larger a sample of households would be required by the conventional estimate to obtain the precision of the network estimate. For example, a design effect of 1.75 implies that a 75 percent larger household sample would be required by the conventional estimator to achieve the precision of the network estimator.

5. SUMMARY AND CONCLUSIONS

The major objectives of the experiment reported in this paper were to field test and compare the conventional and network estimators of diabetes. The network estimator illustrated in this report requires the collection of two items of ancillary information for each person enumerated in the household interviews: (a) the number of his siblings who are living elsewhere and (b) the number of these siblings that have diabetes. The ancillary questions required by the network estimator were included in a preliminary version of the questionnaire of the 1976 Health Interview Survey. The sampling and nonsampling error effects of the alternative estimators presented in this report are based on the initial pretest of this questionnaire.

Based on the pretest findings, the estimate of diabetes prevalence in the total population is about 10 percent greater for the conventional estimator than for the network estimator. However, the estimates based on the network estimator were greater than those based on the conventional estimator for four of the seven population subdomains for which separate estimates were derived from the pretest. In the absence of an independent criterion of the diabetes status of persons enumerated in the pretest, and in view of the small sample size and other limitations of the pretest, it is difficult to interpret the differences between the estimators and to categorically state that one of the estimators is subject to smaller bias than the other.

On the other hand, the pretest disclosed that the network estimator was subject to bias errors due to nonresponse to an item of information that is not needed by the conventional estimator. Thus, the number of living siblings with diabetes was not ascertained for about one percent of the persons enumerated in the pretest. Ordinarily, an item nonresponse rate this small would be inconsequential, but it could be of consequence in estimating a relatively uncommon

Table 2. Diabetes Prevalence Rates (Per 1000 Population) and Their Relative Standard Errors by Type of Estimator and by Age, Whether or Not the Parents Had Diabetes and by Respondent Status.

Population Characteristics	Sample Size	Rate per 1000 Population		Relative Standard Errors (in percent)		Design Effect
		Conventional Estimator	Network Estimator	Conventional Estimator	Network Estimator	
All persons	550	27.3	24.6-33.9	26.5%	20.2%	1.72
Under 25 years	246	-	0.8-7.6	-	100.0	-
Twenty-five years and older	304	49.3	43.8-55.2	26.5	20.1	1.74
Self respondents	221	49.8	41.3-49.0	32.0	23.9	1.79
Nonself respondents	83	48.2	50.5-71.8	49.7	38.4	1.68
Neither parent has diabetes	438	27.4	21.2-25.0	30.8	25.1	1.51
One or both parents have diabetes	80	37.5	46.9-68.9	58.2	35.2	2.73
N.A. whether parents have diabetes	32	-	15.6-70.7	-	104.0	-

condition like diabetes which affects only a small fraction of the population if a substantial portion of the NA's were persons with diabetes.

The sampling errors of the estimates of diabetes based on the network estimator were consistently smaller than those based on the conventional estimator. For example, the conventional estimator would require a household sample size about 75 percent larger than the network estimator in order to match the precision of the network estimate. Furthermore, it is believed that the sample size advantage of the network estimator would be even larger for estimates of diabetes prevalence for relatively small demographic subdomains of the population.

Largely on the basis of findings presented in the paper we have concluded that the pretest demonstrated the feasibility of collecting the ancillary information required by the network estimator. The HIS questionnaire and the interviewer instructions used in the initial pretest were revised somewhat with the view to reducing the impact of item nonresponse. After passing a second field test, the revised set of questions and of interviewer instructions were incorporated into the 1976 questionnaire of the Health Interview Survey.

References

[1] National Center for Health Statistics, "Health Survey Procedures (Concepts, Questionnaire Development, and Definition in the Health Interview Survey)", Vital and Health Statistics, Series 1, No. 2 (1964), pp. 1-66.

[2] _____, "Health Interview Survey Procedures: 1957-1974", Vital and Health Statistics, Series 1, No. 11, (1975), pp. 1-153.

[3] _____, "Origin, Program, and Operation of the U.S. National Health Survey", Vital and Health Statistics, Series 1, No. 2 (1963), pp. 1-41.

[4] Sirken, Monroe G., "Survey Strategies for Estimating Rare Health Attributes," Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (1970), pp. 134-144.

[5] _____, "The Counting Rule Strategy in Sample Surveys" Proceedings of the Social Statistics Section, American Statistical Association (1974), pp. 119-123.

Appendix

Selected Items of Information on Diabetes Completed for Each Household Member: First Pretest of the 1976 HIS Questionnaire

4a. Does anyone in the family (you, your ... etc.) have diabetes?

b. Who is this?

c. Does anyone else have diabetes?

8a. How many living brothers and sisters does -- have?

b. How many of these brothers and sisters have diabetes?

c. How many of --'s brothers and sisters are not living?

d. How many of these brothers and sisters had diabetes?

e. Is --'s mother still living?

f. Does (did) she have diabetes?

g. Is --'s father still living?

h. Does (did) he have diabetes?