AN EVALUATION OF COUNTING RULES IN A NETWORK SURVEY - ESTIMATING PREVALENCE RATES OF RARE EVENTS

Cecelia B. Snowden, National Center for Health Statistics
Ronald F. Czaja, University of Illinois

INTRODUCTION

A need exists to develop efficient methodologies for conducting surveys of populations with relatively rare characteristics. The need is especially critical when parameter estimates are required for planning and program evaluation. For example, rapidly increasing costs of health care have created a need for public health programs to provide support and relief for those faced with financial devastation following a serious, chronic illness such as cancer. Cost effective programs to deal with this problem must be developed.

Surveys employing traditional sampling frames have not provided the required estimates for three reasons: (1) identifying a large national probability sample of patients is difficult, (2) the costs associated with this effort are enormous, and (3) family members and health care providers often limit access to patients needed to obtain accurate information of direct and indirect costs. Despite these barriers, the National Cancer Institute (NCI) continues to need accurate baseline data to assess the cost benefits of screening procedures for early detection, new treatment interventions, and rehabilitation strategies. The NCI therefore funded a series of survey experiments to test methods to be used in a national survey of cancer care costs.

This paper studies the accuracy of conventional and multiplicity counting rules for the estimation of cancer prevalence using a subset of the data collected in one of these experiments.

MULTIPLICITY ESTIMATION

Since 1970 Monroe Sirken and colleagues at the National Center for Health Statistics have been using network survey methods to estimate prevalence of various kinds of morbidity. In network sampling counting rules are algorithms that link enumeration units (or listing units) to elementary units. Rules that allow one enumeration unit to be linked to one elementary unit are conventional rules; rules that allow elementary units to be linked to more than one enumeration units are multiplicity rules.

The estimates in this study are based on the following rules:

(1) patient or events are linked to their own household (conventional),

(2) patients are linked to their own household and households of their siblings (sibling),

(3) patients are linked to their own household and the households of their children (children),

(4) patients are linked to their own household and the households of their siblings and children (relative).

In the following these will also be referred to as rule 1, rule 2, rule 3 and rule 4, respectively. The estimator of prevalence of disease from a simple random sample of m households without replacement from a universe of M households for rule r is

$$\hat{\Theta}_r = \frac{1}{m}\frac{M}{N_t}\sum_{i=1}^{M}\sum_{j=1}^{N}\beta_{rij}\,a_i/S_{rj}, \quad r=1,2,3,4$$

where: N_t = total number or persons in the target population

N = total number of events (patients) in the target population

$$\beta_{rij} = \begin{cases} 1 \text{ if } j^{th} \text{ event is reported at the } i^{th} \text{ household (HH) by rule r} \\ 0 \text{ otherwise} \end{cases}$$

$$a_i = \begin{cases} 1 \text{ if } i^{th} \text{ HH is selected} \\ 0 \text{ otherwise} \end{cases}$$

S_{rj} = multiplicity of j^{th} event by rule r

Note: For rule 1, $S_{1j} = 1$, $j = 1,2,\ldots,N$ by the conventional rule each event is linked to a de jure residence;

for $r = 2,3,4$ $S_{rj} \geq 1$, which represents the total number of different households in which the patient and respective relatives reside.

Before we give expressions for the mean, variance and mean square error of the proposed estimator, let us first define the following counting rule related parameters:

$\gamma_{1,r} = \frac{1}{N}\sum_{j=1}^{N}(1/S_{rj})$ - average of reciprocal of multiplicity

$\gamma_{2,r} = \frac{1}{N}\sum_{j=1}^{N}(1/S^2_{rj})$ - average of reciprocal of squared multiplicity

P_1 = probability that cancer patient is reported at de jure residence of patient

P_r = probability that cancer patient is reported at residence of relative as specified by rule r, r=2,3,4

$\Theta = \frac{N}{N_t}$ = true prevalence rate of disease in target population

Then,

$$E(\hat{\theta}_1) = \theta P_1$$

$$E(\hat{\theta}_r) = \theta[P_r + (P_1 - P_r)\gamma_{1,r}], \quad r=2,3,4$$

$$Var(\hat{\theta}_1) = \frac{\theta P_1}{m}\left(\frac{M}{N_t} - P_1\theta\right)$$

$$Var(\hat{\theta}_r) = \frac{1}{m}\left\{\frac{\theta M}{N_t} P_r \gamma_{1,r} + \frac{\theta M}{N_t}(P_1 - P_r)\gamma_{2,r} - [E(\hat{\theta}_r)]^2\right\}$$

$$MSE(\hat{\theta}_r) = Var(\hat{\theta}_r) + Bias^2(\hat{\theta}_r)$$

Estimators based on conventional and multiplicity rules are compared in terms of bias (B), mean square error (MSE) and the number of households from which the multiplicity estimator has the smallest MSE.

SURVEY DESIGN

Four experiments were conducted in the overall NCI research project, only the first two experiments are appropriate to network surveys. Experiment 1 was designed to evaluate the degree to which cancer patients will be reported in their own households using a general health interview. An additional purpose was to obtain the name and address of one randomly selected child or sibling of the cancer patient to interview in the second experiment. Experiment 2 was designed to test the reporting level and accuracy of siblings and children of cancer patients. In addition, its purpose was to determine if network sampling was more efficient than traditional sampling, and, if so, did children, or siblings, or both, provide the most efficient sampling frame.

The samples for Experiments 1 and 2 included cancer patients from two regional tumor registries in Illinois, randomly selected relatives of the patients and a sample of households from the general population who resided in close proximity to the patient and relative households. The two hospitals which provided the patient sample diagnosed approximately 1,500 new cancer cases annually. Each registry was requested to provide patients who were living, noninstitutionalized, who resided in Illinois, and who were diagnosed at the institution. Because of the need to protect the privacy of the patient the sampling of cancer patients was managed by the Illinois Cancer Council (ICC).

A sample of 325 cancer patients and 275 decoy cases comprised Experiment 1. From the information provided in the patient interviews, a sample of 205 relatives living in Illinois was combined with 167 decoy households for experiment 2.

The registry sample was organized into nine cancer groups, three geographic regions, and three one-year diagnostic periods from August 1977 through July 1980.

Thirty of the 325 patient households were classified as out-of-scope (See Figure 1). The response rate for the remainder of the patient sample was 89 percent; 10 percent refused to be interviewed and one percent was not interviewed for other reasons.

For the relative sample, 12 households were defined as out-of-scope. The majority of these cases were households where the relative was not enumerated. Eighty-four percent of the relative households cooperated; 14 percent refused to be interviewed; and 2 percent were not interviewed for other reasons.

RESULTS

The proportion of cancer patients who were reported in their own households and in households of relatives are shown in Table 1. The reporting rates are presented by patient age, sex, and for white males and white females. Data are not presented separately for nonwhites because of their small sample size (n = 30).

Eighty-nine percent of all patient households reported cancer patients. This was higher than the reporting rate in the relative households which was 80 percent (sibling-75%, children-85%). This result was surprising because prior to data collection we hypothesized that reporting would be lower in cancer patient households due to the reluctance of patients and of other household members to discuss the disease.

In patient households, age was not an important determinant in whether the patient was reported. There was a small difference by sex, and when race was controlled, the difference was further increased between white males and white females, 88% and 97% respectively.

Reporting rates by type of relative household show that children report better than siblings. This relationship is consistent across all domains and characteristics presented in the table. Children reporting rates range from 80% for all fathers to 93% for white mothers. The lowest and highest reporting rates for siblings occur for male patients and white female patients (63% and 90%) respectively.

The bias component of the MSE is presented as a ratio of the three counting rules to the conventional rule in Table 2. The conventional rule has the smallest bias with the children rule having the smallest bias among the multiplicity rules.

To determine the performance of the conventional rule and each of the multiplicity rules, the ratio of each multiplicity rule to the conventional rule is given in Table 3. A range of theoretical prevalence rates, and samples sizes in terms of number of households is also presented. The MSE intersection column gives the sample sizes in households for each prevalence rate where the MSE of the multiplicity rule is equal to the MSE of the conventional rule. Above the indicated sample size, the conventional rule is more efficient.

The multiplicity rule for children has uniformly the largest number of households where this rule gives a more accurate estimate for prevalence rates. A second consistent pattern

is the relationship between the prevalence rates and the sample sizes of households where the MSEs intersect. The MSE intersection is inversely proportional to the prevalence rates for all three counting rules. Thus, as the prevalence rates increase by one unit the samples of households where the MSE ratios are equal to one decrease by 50 percent. Data in Table 3 also indicate that only the prevalence of one tenth of 1% for the children's rule would the multiplicity estimator be more accurate for a sample size of households comparable to the approximate number of households in the Health Interview Survey (HIS).

SUMMARY AND CONCLUSIONS

This paper provides evidence that a network survey can improve the accuracy of estimates of rare events without increasing sample size. Particularly for researchers with limited resources these findings offer a plausible alternative to the conventional survey. Criteria for evaluation herein are in terms of bias and mean square error considered separately. Before definitive conclusions can be reached, cost models including optimum yield, minimum bias and variances considered jointly are examples of additional criterion that must be evaluated.

ACKNOWLEDGEMENTS

Thanks are due to Dr. Robert J. Casady and Dr. Monroe G. Sirken for statistical advice.

Janice Melvin for the typing of this paper.

REFERENCES

Nathan, Gad, (1976) "An Empirical Study of Responses and Sampling Errors for Multiplicity Estimates with Different Counting Rules:, Journal of the American Statistical Society, Vol. 71, No. 356, pp. 808-815.

Sirken, Monroe G., (1970) "Household Surveys with Multiplicity", Journal of the American Statistical Society, Vol. 65, pp. 257-266.

_____, (1973) "Design of Household Sample Surveys to Test Death Registration Completeness", Demography, Vol. 10, No. 3, pp. 469-478.

Survey Research Laboratory, University of Illinois: Household Network Surveys of Cancer Care Cost: A Research Pilot. Final Report; Vol. 1, September, 1982.

Figure 1. Sample Disposition and the Results of the Patient Reporting

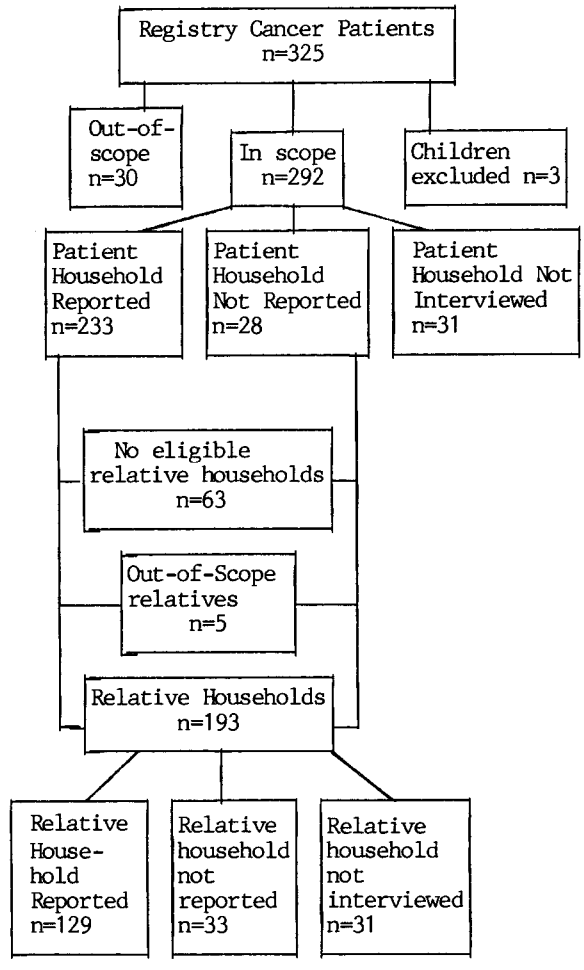

TABLE 1

Reporting Rates for Patient and Relative Households by Selected Cancer Patient Characteristics

Patient Characteristics	HOUSEHOLDS PATIENT $P_1(N)$	SIBLING $P_2(N)$	CHILDREN $P_3(N)$	RELATIVE $P_4(N)$
Total Sample	.89 (260)	.75 (84)	.85 (78)	.80 (162)
Age				
18-64	.90 (132)	.80 (49)	.85 (34)	.82 (83)
65 and over	.88 (128)	.69 (35)	.84 (44)	.77 (79)
Sex				
MALE	.86 (131)	.63 (38)	.80 (45)	.72 (83)
FEMALE	.92 (129)	.85 (46)	.91 (33)	.87 (79)
WHITE MALE	.88 (120)	.68 (34)	.84 (43)	.77 (77)
WHITE FEMALE	.97 (110)	.90 (40)	.93 (30)	.91 (70)

TABLE 2

Ratio of the Bias of Multiplicity to Conventional Counting Rule

Sibling/Conventional	1.50
Children/Conventional	1.16
Relative/Conventional	1.51

TABLE 3

Ratio of Mean Square Error Estimates of Multiplicity to Conventional Counting Rules by Prevalence Rates and Samples of Households:

Prevalence Rates	HOUSEHOLDS 500	1,000	5,000	15,000	30,000	40,000	MSE INTER-SECTION
				SIBLING RULE			
.001	.63	.65	.79	1.07	1.33	1.44	12,129
.002	.65	.69	.95	1.33	1.60	1.71	6,061
.005	.70	.79	1.26	1.69	1.91	1.98	2,421
.01	.79	.95	1.54	1.91	2.06	2.10	1,207
.02	.95	1.17	1.80	2.06	2.15	2.17	600
.03	1.07	1.34	1.92	2.12	2.18	2.19	398
				CHILDREN RULE			
.001	.63	.64	.70	.83	.94	.99	41,865
.002	.64	.66	.77	.94	1.06	1.11	20,929
.005	.66	.70	.91	1.10	1.20	1.22	8,367
.01	.70	.77	1.03	1.20	1.26	1.28	4,180
.02	.77	.87	1.15	1.26	1.30	1.31	2,086
.03	.82	.94	1.20	1.29	1.31	1.32	1,388
				SIBLING & CHILDREN RULE			
.001	.44	.46	.63	.94	1.23	1.36	17,537
.002	.46	.51	.80	1.23	1.55	1.67	8,766
.005	.53	.63	1.15	1.65	1.89	1.97	3,502
.01	.63	.80	1.47	1.90	2.06	2.11	1,748
.02	.80	1.06	1.76	2.06	2.16	2.18	870
.03	.94	1.24	1.90	2.13	2.19	2.21	578