ESTABLISHING SAMPLING RATES FOR THE FAMILY HISTORY OF CANCER VALIDATION STUDY

Lou Rizzo, Westat; Barry Graubard, NCI; Ralph DiGaetano, Westat; Louise Wideroff, NCI
Lou Rizzo, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Sensitivity, Specificity, Stratification, and Optimal

1. Introduction

The primary purpose of the Family History of Cancer Validation Study (FHCVS) is to examine the quality of cancer reports given by adults for their blood relatives. These family history reports of cancer are used clinically to screen for patients who may carry an inherited genetic mutation that increases cancer risk in affected family members, and are also used in epidemiological studies of cancer risk. It has been found in numerous studies that there is some error in these family histories when compared against cancer registry and medical records for the family member reported. Because of the considerable importance of family history reports, it is important to measure the associated error rates.

In the recent studies regarding this issue, the two key parameters, which measure this accuracy, are the sensitivity and the specificity. The sensitivity *SN* in this application is defined to be the percentage of relatives having a specified cancer who are accurately reported as such. The sensitivity *SN* in this application is defined to be the percentage of relatives not having a specified cancer who are accurately reported as such.

There have been a number of recent papers describing studies which have explored the quality of family history reports and have provided estimates of sensitivity and specificity. These papers include the case-control Utah Diet, Activity, and Reproduction in Colon Cancer Study (Kerber and Slattery 1997), a second case-control study in Australia as reported by Aitken, et al. (1995), and two studies of cancer cases only (i.e., only persons with cancer are providing reports for their relatives): Mussio, et al. (1998), and Anton-Culver, et al. (1996). The sensitivity estimates (for different populations) range from 70 percent to 92 percent, and the specificity estimates (given in Aitken, et al., Mussio, et al., and Anton-Culver, et al.) range from 97 percent to 99 percent.

The reported estimates of sensitivity differ considerably across these studies. The wide variation in the sensitivity estimates may be attributable to high sampling error for all of the studies and considerable differences in the study populations and methodologies. The FHCVS study described in this paper is designed to be a fully population-based study of a single state (Connecticut), with large enough sample sizes to provide high precision in the estimators.

2. Sample Design for the FHCVS

The FHCVS survey begins with a random digit dialing (RDD) sample of telephone households in Connecticut. The RDD sample design is list-assisted (Tucker, et al. 1993). A household will be eligible if it has one adult age 25 to 64. Within eligible households one adult in this age range will be randomly sampled from the set of all such adults. We are planning a sample size of 6,000 telephone numbers, expecting to yield 1,800 adult respondents to complete the first questionnaire. As will be seen in the following sections, we believe this is sufficient to meet the precision requirements of the survey.

The sampled adult will be asked to roster his/her first-degree biological relatives (parents, full siblings, children) and second-degree biological relatives (half siblings, grandparents, siblings of parents, and nephews and nieces). The roster will collect information about the age of the relatives, their current status as living or deceased, and their cancer status (whether or not they have had any of the following five cancers: prostate, breast, lung, colorectal, and ovarian). In this paper we will only discuss first-degree male relatives and only discuss the validation of prostate cancer status, in the interests of brevity. The analysis was very similar for prostate validation in second-degree male relatives and for the other four cancers. Details are available on the overall sample design from rizzol1@westat.com on request.

For first-degree male relatives and prostate validation, a stratified random sample of relatives will be taken from this roster, from each of four strata. The strata reflect age and report status:

- High-incidence age group cancer reports: first-degree male relatives age 55 and older with a report of prostate cancer;
- High-incidence age group non-cancer reports: first-degree male relatives age 55 and older with a respondent report of no prostate cancer;
- Low-incidence age group cancer reports: first-degree male relatives age 45 to 54 with a report of prostate cancer; and

- Low-incidence age group non-cancer reports: first-degree male relatives age 45 to 54 with a respondent report of no prostate cancer.

Male relatives younger than 45 were excluded from consideration for prostate cancer validation, as their incidence of prostate cancer is negligibly low. In order to minimize burden on each respondent and family, we subsampled relatives from the full roster to have their cancer reports validated. The expected sample size of first-degree male relative reports under our sample design is 1.35 per household, or about 2,400 relative reports. Of these we expect a yield of roughly 1,200 55+ first-degree male relative reports, and roughly 300 45-54 year old first-degree male relative reports. This is our pool from which to draw reports of prostate cancer and reports of no prostate cancer to validate.

3. Estimation of Sensitivity and Specificity

We define p_h h=1,2 as the probability of a true case for a particular cancer in age stratum h (i.e., the probability of sampling a person who had that cancer at least once in their lifetime), and R_h as the number of relative reports in the stratum. We define the sensitivity and specificity within each stratum as SH_h and SP_h respectively, and define the population 2 by 2 cell probabilities for each stratum as follows:

	True cases	True non-cases	Total
Reports of cancer	$p_{1h}=p_h*SN_h$	$p_{4h}=(1-p_h)*(1-SP_h)$	$p_{rh}=p_{1h}+p_{4h}$
Reports of non-cancer	$p_{2h}=p_h*(1-SN_h)$	$p_{3h}=(1-p_h)*SP_h$	$p_{nh}=p_{2h}+p_{3h}$
Total	$p_h=p_{1h}+p_{2h}$	$1-p_h=p_{3h}+p_{4h}$	1

The overall sensitivity and specificity (over all the strata) can be defined as follows:

$$SN=\frac{\sum_{h=1}^{H} R_h p_{1h}}{\sum_{h=1}^{H} R_h(p_{1h}+p_{2h})} \qquad SP=\frac{\sum_{h=1}^{H} R_h p_{3h}}{\sum_{h=1}^{H} R_h(p_{3h}+p_{4h})}.$$

At the time of relative rostering we only know the report status for each relative: the true cancer status is only known at the end of the tracking process. The stratified sample design is defined essentially by choosing sampling rates for cancer reports and non-cancer reports within each age stratum. The sampling rates are designated as f_{rh} for the cancer report sampling rate in stratum h, and f_{nh} for the non-cancer report sampling rate in stratum h, with n_{rh}, n_{nh}, and n_h designating the realized sample sizes among the cancer reports, non-cancer reports, and all sampled relative reports respectively in stratum h. The expected values of these sample sizes are as follows:

- $E(n_{rh})=R_h*f_{rh}*(p_{1h}+p_{4h})$;
- $E(n_{nh})=R_h*f_{nh}*(p_{2h}+p_{3h})$; and
- $E(n_{nh})=R_h*((f_{rh}*(p_{1h}+p_{4h}))+(f_{nh}*(p_{2h}+p_{3h})))$.

We can view the realized sample within each stratum h as a sample of size n_h from the population distribution defined by four sample distribution probabilities as follows:

- $q_{1h}=f_{rh}*p_{1h}/(f_{rh}(p_{1h}+p_{4h})+f_{nh}(p_{2h}+p_{3h}))$;
- $q_{2h}=f_{nh}*p_{2h}/(f_{rh}(p_{1h}+p_{4h})+f_{nh}(p_{2h}+p_{3h}))$;
- $q_{3h}=f_{nh}*p_{3h}/(f_{rh}(p_{1h}+p_{4h})+f_{nh}(p_{2h}+p_{3h}))$; and
- $q_{4h}=f_{rh}*p_{4h}/(f_{rh}(p_{1h}+p_{4h})+f_{nh}(p_{2h}+p_{3h}))$.

Write n_{1h} through n_{4h} as the number of sampled relatives falling into each cell, with $n_{1h}+n_{2h}+n_{3h}+n_{4h}=n_h$, and $E(n_{ih})=n_h*q_{ih}$. Consistent estimators of SN and SP from this sample are

$$\hat{SN}=\frac{\sum_{h=1}^{H}\frac{n_{1h}}{f_{rh}}}{\sum_{h=1}^{H}\frac{n_{1h}}{f_{rh}}+\sum_{h=1}^{H}\frac{n_{2h}}{f_{nh}}} \qquad \hat{SP}=\frac{\sum_{h=1}^{H}\frac{n_{3h}}{f_{nh}}}{\sum_{h=1}^{H}\frac{n_{3h}}{f_{nh}}+\sum_{h=1}^{H}\frac{n_{4h}}{f_{rh}}}.$$

Note that these are combined ratio estimators (cite e.g., Cochran 1977, Section 6.11). If we condition on the realized $n_h's$, then $(n_{1h},n_{2h},n_{3h},n_{4h})$ are independent multinomial random variable with parameters $(n_h,q_{1h},q_{2h},q_{3h},q_{4h})$. A straightforward application of the delta method proves the consistency of $\hat{SN}$ and $\hat{SP}$, and gives an approximation for the variances:

$$Var(\hat{SN}) \approx \frac{1}{\left\{\sum_{h=1}^{2} R_h(p_{1h}+p_{2h})\right\}^2}\left\{(1-SN)^2\sum_{h=1}^{N}\frac{R_h}{f_{rh}}p_{1h}(1-q_{1h}) + SN^2\sum_{h=1}^{N}\frac{R_h}{f_{nh}}p_{2h}(1-q_{2h}) + 2SN(1-SN)\sum_{h=1}^{N}R_h p_{1h}p_{2h}\right\}$$

$$Var(\hat{SP}) \approx \frac{1}{\left\{\sum_{h=1}^{2} R_h(p_{3h}+p_{4h})\right\}^2}\left\{(1-SP)^2\sum_{h=1}^{N}\frac{R_h}{f_{nh}}p_{3h}(1-q_{3h}) + SP^2\sum_{h=1}^{N}\frac{R_h}{f_{rh}}p_{4h}(1-q_{4h}) + 2SP(1-SP)\sum_{h=1}^{N}R_h p_{3h}p_{4h}\right\}$$

4. Optimal Sampling Rates for Prostate Validation

The prostate cancer incidence for the 55+ group is estimated based on Connecticut tumor registry data (SEER) to be 11 percent. The prostate cancer incidence for the 45-54 group is estimated to be 0.4 percent. For these calculations we will assume the following true values for sensitivity and specificity:

- 55+ First-degree male relatives: 80 percent sensitivity, 90 percent specificity; and
- 45-54 First-degree male relatives: 90 percent sensitivity, 95 percent specificity.

The higher rates for 45-54 group relatives is based on our anticipation that accuracy will be greater for this age group as few of these relatives will be deceased and most will have had their cancers recently, if they have had the cancer, with both effects increasing accuracy of relative reports.

Determining the sampling rates for the four sample strata finalize the sample design for first-degree male relative prostate cancer validation. As a function of sampling rates, the standard errors will be lowest when all of the sampling rates are 1. We wish to examine the extent to which we can reduce any of the sampling rates and thus survey costs without increasing the standard errors beyond their desired limits.

The approximate variance of the sensitivity and specificity estimators are polynomial functions of the inverses of the sampling rates, and as such are easily analyzed. Figures 1 and 2 show the standard errors of sensitivity and specificity as functions of the 55+ reported cancer sampling rate and the 55+ reported non-cancer sampling rate, while keeping the sampling rates for the 45-54 strata at their maximum value of 1.

Figure 1. Standard errors as a function of the 55+ reported cancer stratum sampling rate, with the other sampling rates set at 1

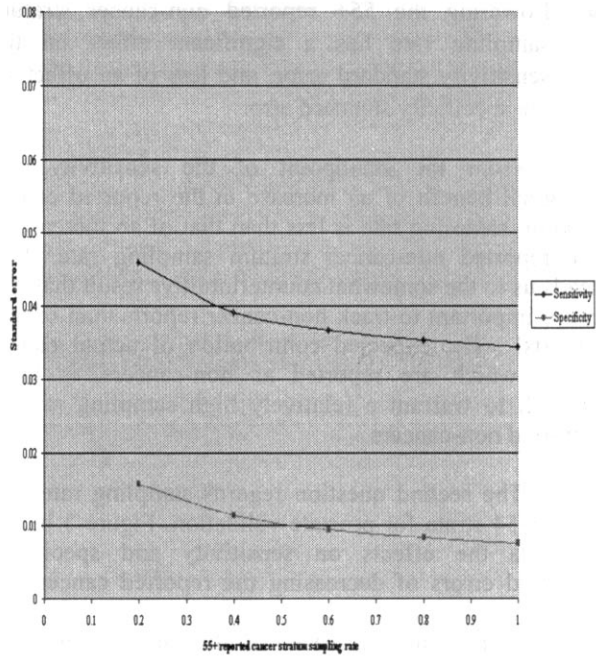
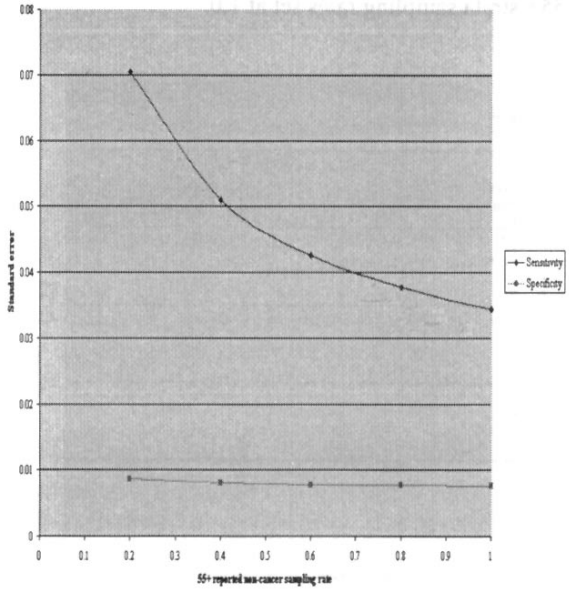


Figure 2. Standard errors as a function of the 55+ reported non-cancer stratum sampling rate, with the other sampling rates set at 1.0



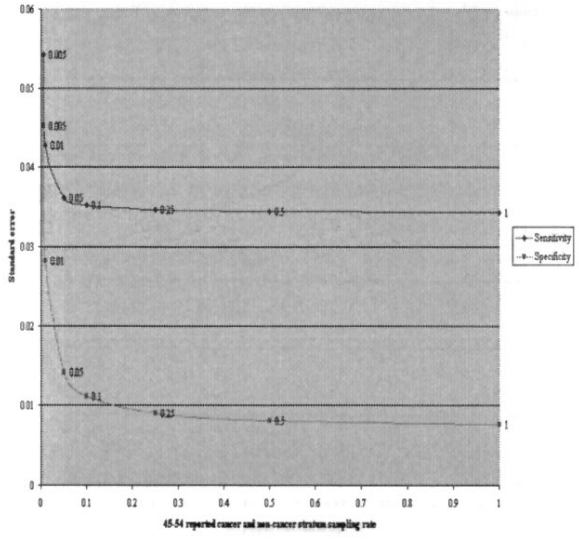
The following patterns are readily observed:

- Lowering the 55+ reported cancer sampling rate has a significant effect on the specificity standard error, and less of an effect on the sensitivity standard error; and
- Lowering the 55+ reported non-cancer stratum sampling rate has a significant effect on the sensitivity standard error, and less of an effect on the specificity standard error.

From the standpoint of the sensitivity, the marginal benefit of an increase in the reported cancer stratum sampling rate is less than that of an increase in the reported non-cancer stratum sampling rate. This leads us to the somewhat counterintuitive result that it is more important to track non-cancer reports than cancer reports! The expected contribution of actual cancer cases, which are reported as non-cancers, is large enough to warrant a relatively high sampling rate of reported non-cancers.

The second question regards sampling rates for the 45-54 strata for prostate validation. Figure 3 below presents the effects on sensitivity and specificity standard errors of decreasing the reported cancer and reported non-cancer stratum sampling rates (the two sampling rates are decreased by the same amount: the first case is with both rates one, the second case with both rates 0.5, the third case with both rates 0.25, etc.).

Figure 3. Standard errors as a function of the 45-54 reported cancer and non-cancer stratum report and report sampling rate (with the two rates equal), with the 55+ strata sampling rates set at 1.0



5. Discussion

An examination of the formulas for the approximate variances of $\hat{SN}$ and $\hat{SP}$ can shed light on the results presented in Section 4.

In the expression for the variance of the sensitivity $Var(\hat{SN})$ the first term has $(1-SN)^2$ as a factor and has f_{rh} in the denominator, and the second term has SN^2 as a factor and has f_{nh} in the denominator. As SN is generally taken to be closer to 1 than to 0 the SN^2 factor tends to make the second variance term larger than the first variance term, which has $(1-SN)^2$ as the factor. The second (larger) variance term is also the term which has f_{nh} in the denominator. Thus $Var(\hat{SN})$ tends to be more sensitive to the non-cancer report sampling rate than the cancer report sampling rate, making a larger non-cancer report sampling rate optimal for sensitivity estimation unless the cost of validating a non-cancer report is much higher.

For the $Var(\hat{SP})$ the situation is reversed: the first term has $(1-SP)^2$ as a factor with f_{nh} in the denominator, and the second term has SP^2 as a factor with f_{rh} in the denominator. SP tends to be very close to 1, so again the second term will tend to make the larger contribution to the variance for a fixed set of sampling rates. The second term is the one with f_{rh} in the denominator, so changes in the reported cancer sampling rate can have a dramatic effect on the precision of the specificity estimator.

The low values of the optimizing sampling rates for sensitivity estimation for both cancer reports and non-cancer reports in the low incidence stratum can be explained by the small values of the q_{1h} and q_{2h} factors in the low-incidence stratum: q_{1h} and q_{2h} correspond to 'true cases', whose numbers are quite small in the low-incidence stratum. This is not true for specificity: the q_{3h} and q_{4h} factors are much larger, as there are many true non-cases in this stratum. For sensitivity estimation the very low sampling rates will be optimal for the low incidence stratum, whereas for specificity estimation higher rates (but still lower than the high incidence stratum rates) are optimal for the low incidence stratum.

6. References

Aitken, J., Bain, C., Ward, M., Siskind, V., and MacLennon, R. (1995). How accurate is self-reported family history of colorectal cancer. *American Journal of Epidemiology*, **141**, p. 863-871.

Anton-Culver, H., Kurosaki, T., Taylor, T.H., Gildea, M., Brunner, D., and Bringman, D. (1996). Validation of family history of breast cancer and identification of the BRCA1 and other syndromes using a population-based cancer registry, *Genetic Epidemiology*, **13**, p. 193-205.

Cochran, W.G. (1977). *Sampling Techniques,* (3rd edition). New York: John Wiley & Sons.

Kerber, R.A., Slattery, M.L. (1997). Comparison of self-reported and database-linked family history of cancer data in a case-control study. *American Journal of Epidemiology,* **146**, p. 244-248.

Mussio, P., Weber, W., Brunetti, D., Stemmermann, G., and Torhorst, J. (1998). Taking a family history in cancer patients with a simple questionnaire, *Anticancer Research*, **18**, p. 2811-2814.

Tucker, C., Casady, R., and Lepkowski, J. (1993). A hierarchy of list-assisted stratified telephone sample design options. Paper presented at the Annual Conference of the American Association for Public Opinion Research.

Acknowledgments

We thank Graham Kalton for his observation that led to our discussion section.