American Journal of Epidemiology
Copyright © 1996 by The Johns Hopkins University School of Hygiene and Public Health
All rights reserved

Vol. 143, No. 1
Printed in U.S.A.

Optimal Sampling Strategies for Two-Stage Studies

Marie Reilly

The optimal allocation of available resources is the concern of every investigator in choosing a study design. The recent development of statistical methods for the analysis of two-stage data makes these study designs attractive for their economy and efficiency. However, little work has been done on deriving two-stage designs that are optimal under the kinds of constraints encountered in practice. The methods presented in this paper provide a means of deriving designs that will maximize precision for a fixed total budget or minimize the study cost necessary to achieve a desired precision. These optimal designs depend on the relative information content and the relative cost of gathering the first- and second-stage data. In place of the usual sample size calculations, the investigator can use pilot data to estimate the study size and second-stage sampling fractions. The gains in efficiency that can result from such carefully designed studies are illustrated here by deriving and implementing optimal designs using data from the Coronary Artery Surgery Study (*Circulation* 1980;62:254–61). *Am J Epidemiol* 1996;143:92–100.

cost savings; data collection; epidemiologic methods; likelihood functions; models, statistical; research design

Traditionally, epidemiologists would choose a simple cross-sectional, cohort, or case-control design for the acquisition of observational data, thereby enabling straightforward sample size calculations and data analysis. It has been recognized for some time that the power of a test of a simple relative risk or odds ratio can be improved by sampling unequally from exposed and nonexposed (or from case and control) subjects, if the cost of ascertaining data is different for the two groups (1). More recently, Nam and Fears (2) have derived some optimal sampling plans for such tests in matched case-control designs.

Typically, data on a large number of covariates are gathered in epidemiologic studies, and a number of these are of interest in the final analysis, as either risk factors or confounders. Some covariates such as sex and age can be ascertained easily and cheaply, while others such as laboratory test results or radiographic findings might involve considerable expense. Two-stage designs (3) provide a means of reducing study costs, by ascertaining data on some covariates only for a subsample of study subjects. The first stage of a typical two-stage study consists of the response variable and some covariates for all study subjects, while the second-stage data consist of any other covariates of interest for a subsample of subjects. The second-stage sample is selected using stratified random sampling within the strata defined by the variables available at the first stage. For example, in a two-stage case-control study, the second-stage sample consists of a random sample of cases and a random sample of controls. In contrast, a case-cohort study is a two-stage design where the second-stage sample consists of all cases and a random sample of controls. This paper considers the general two-stage design, where data on the outcome variable and some easily obtained covariates are available for all study subjects, while the more expensive covariates are ascertained only for the subsample of second-stage subjects.

With the advent of methods that enable all of the data (first-stage *and* second-stage) to be accommodated in the analysis (3–5), two-stage designs offer the possibility of a more economical study and more efficient estimates. Although Cain and Breslow (4) have illustrated that balanced designs are more efficient than case-control designs for a number of logistic models, little work has been done on determining sampling strategies that are optimal with respect to cost and/or efficiency. The optimal "double sampling" strategies of Buonaccorsi (6) are derived only for random subsampling and require assumptions of normality. Tosteson and Ware (7) provide a means of choosing between sampling plans when the "first-stage" data consist of surrogates for outcome and exposure and a logistic model is supposed.

Received for publication September 21, 1993, and in final form October 2, 1995.

From the Department of Statistics, University College Dublin, Belfield, Dublin 4, Ireland.

In recent work, Reilly and Pepe (8) developed a new "mean-score" method of analysis of two-stage data that permits the derivation of explicit expressions for sampling strategies to 1) maximize precision for a given cost and 2) minimize cost for a fixed precision. The mean-score estimates can be obtained by standard regression analysis with suitable weights (see below), although the variance of the estimates requires some additional calculations. A description of the method and an illustration of optimal sampling are presented in this paper. An outline of the technical details is provided in Appendix 1, and interested readers are referred to the original article (8) for statistical details.

NOTATION AND METHODS

The likelihood-based mean-score estimator, like the estimators of Cain and Breslow (4) and Flanders and Greenland (5), makes use of all of the data (first-stage *and* second-stage) for the analysis. Let us designate the outcome variable Y and the regression model $P_\beta(Y|X)$, where X denotes the covariates of interest. The first-stage covariates, Z, may include some components of X that we will label X_C because these components are complete for all subjects. The second-stage covariates will be denoted X_I, since these components are incomplete for all subjects who were not sampled at the second stage. The covariates in the regression model, $X = (X_C,X_I)$, will be simply X_I in the case where no components of X are ascertained at the first stage. This would happen, for example, when Z consists only of surrogates for some components of X. Interest is focused on the estimation of the parameter β, which for a logistic regression model is the vector of adjusted log odds ratios.

The properties of the mean-score method have been derived for categorical Y and Z. It can be seen from equation 1 in Appendix 1 that the mean-score estimate of β can be found by using the weights $[n(Z_i,Y_i)]/[n_2(Z_i, Y_i)]$ in a statistical package that accommodates the regression model $P_\beta(Y|X)$. Although the estimating equation in expression 1 is the same as that used by Flanders and Greenland, the variance expression 2 differs from the one that they derived using a pseudolikelihood approach. The simple form of our variance expression enables the derivation of optimal designs, as discussed below. The first term of the variance of the mean-score estimator in expression 2 represents the variance of the maximum likelihood estimator if second-stage covariates were available for all subjects, so the second term represents a penalty due to incomplete sampling at the second stage. The I term in expression 2 is the average information over all subjects in the population, while the $\mathcal{V}$ term involves, for each (Z,Y) stratum, the variance of the score function over all subjects in that stratum. The estimate of the variance in equation 4 replaces these quantities by the corresponding sample mean (suitably weighted to account for the second-stage sampling mechanism) and sample variances. Hence, to obtain the variance estimate, the user needs to calculate the estimates (at $\beta = \hat{\beta}$) of the score and information components in the (Z,Y) cells. For the logistic regression model, these components can be expressed very simply in terms of the covariate values and the predicted values from the model. For example, for the dichotomous event Y and the simple logistic model

$$\pi = P(Y = 1) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2}}{1 + e^{\beta_0+\beta_1X_1+\beta_2X_2}},$$

where $\beta = (\beta_0, \beta_1, \beta_2)$, the estimated score vector and information matrix are

$$\frac{\partial}{\partial\beta}\log P_\beta(Y|X)|_{\beta=\hat{\beta}} = \begin{pmatrix} Y - \hat{\pi} \\ X_1(Y - \hat{\pi}) \\ X_2(Y - \hat{\pi}) \end{pmatrix}$$

and

$$-\frac{\partial^2}{\partial\beta^2}\log P_\beta(Y|X)|_{\beta=\hat{\beta}} = \hat{\pi}(1-\hat{\pi})\begin{bmatrix} 1 & X_1 & X_2 \\ X_1 & X_1{}^2 & X_1X_2 \\ X_2 & X_2X_1 & X_2{}^2 \end{bmatrix}.$$

The predicted values $\hat{\pi}$ are provided by most statistical packages, and the symmetry of the covariates in the score and information estimates makes them easily accommodated in calculations.

OPTIMAL SAMPLING

Expression 2 for the variance of $\hat{\beta}$ depends on the sampling fraction of second-stage subjects in each (Z,Y) cell, and on the variance of the score terms in these cells. It is not surprising, then, that optimal studies can be designed by an astute choice of the second-stage sampling fraction in each of these cells. What is considered "optimal" in any setting will depend on the purpose of the study and on various constraints such as budgetary considerations. The various optimal sampling strategies derived in the paper by Reilly and Pepe (8) will be briefly discussed here, while an outline of the technical details can be found in Appendix 2.

The simplest expression arises in the situation where the sizes of the first- and second-stage samples are fixed and it is necessary to maximize the precision (i.e., minimize the variance) of the estimator for some component $\hat{\beta}_k$ of $\hat{\beta}$ which is of primary interest. In this case, the optimal sampling fraction, ρ'_{ZY}, for the (Z,Y) cell is given by equation 5. The optimal sampling fraction for each (Z,Y) cell depends on the prevalence of that cell in the population, ρ_{ZY}, and on the "weighting factor," W_{ZY}, for the cell. This weighting factor is computed from the information matrix and the variance of the score within the cell, and hence reflects in some sense the variability of subjects within the cell. In the simple case of scalar β, the optimal strategy is to sample cell (Z,Y) with intensity proportional to the standard deviation of the score in that cell. This is similar to the traditional optimum allocation for stratified sampling (9) with the variance of the score replacing the simple variance.

In many practical settings, investigators choose a two-stage design because of budget limitations. Let us suppose that the total budget available for the study is B. Denoting the cost of each first-stage observation C_1 and the additional cost of ascertaining second-stage data for a subject C_2, clearly $B = nC_1 + n_2C_2$, where n is the total number of subjects and n_2 is the number of second-stage subjects. With the budget fixed at B, the optimal design is now the study size n and the second-stage sampling fractions ρ'_{ZY} which minimize the variance of $\hat{\beta}_k$. These optimal values are given by equation 6. Note that the expression for ρ'_{ZY} is the same as before, except that $[B - nC_1]/C_2$ replaces n_2. Since C_1 is the unit cost of first-stage data (gathered for all subjects), $B - nC_1$ represents the funds available for second-stage data, so $[B - nC_1]/C_2$ is the "affordable" second-stage sample size. Hence, the optimal ρ'_{ZY} for a fixed budget has this affordable sample size replacing the fixed n_2 in the previous example.

In the planning stages of a study, the investigator may wish to estimate the budget required in order for the study to yield a meaningful result. For simple observational studies, minimal sample size requirements are calculated based on some desired power or precision. In practice, these precision requirements may be relaxed to keep the cost of the study within reasonable limits. The analogue of this for a two-stage study is the determination of the overall sample size and second-stage sampling fractions required to achieve a specified precision at minimal cost. The equations in expression 7 in Appendix 2 present the expressions for such an optimal design, where the study cost is minimized subject to the standard deviation of $\hat{\beta}_k$ being fixed at δ.

The solutions to equations 5–7 may yield some second-stage sampling fractions $\rho'_{ZY} > 1$. In this case, the largest ρ'_{ZY}, say $\rho'_{Z_1Y_1}$, is set equal to 1 (i.e., all subjects in cell (Z_1,Y_1) are included in the second-stage sample) and the remaining second-stage subjects are sampled optimally from the other cells. This is accomplished for equation 5 by replacing n_2 with $n_2 - n(Z_1,Y_1)$ and using summation over $\{(Z,Y) \neq (Z_1,Y_1)\}$ in the denominator. If the solution still yields a sampling fraction greater than 1, the process is repeated. For equations 6 and 7, summation is again taken over $\{(Z,Y) \neq (Z_1,Y_1)\}$ and C_1 is replaced by $C_1 + C_2\rho_{Z_1Y_1}$ so that the cost of those second-stage subjects who will definitely be sampled is being viewed as part of the first-stage costs. This makes intuitive sense, since these costs are independent of the second-stage sampling intensities. Note that for equation 6, $B - nC_1$ will be replaced by $B - nC_1 - n(Z_1, Y_1)C_2$, which is the remaining funds which are to be optimally allocated to the cells $(Z,Y) \neq (Z_1,Y_1)$, so again the second-stage subjects who are to be sampled with probability 1 are treated as part of the first-stage costs.

In addition to the parameter β and the prevalences ρ_{ZY} of the cells, the optimal designs discussed above depend on the information matrix I and on the variance of the score in the various cells. These quantities will be unknown at the study design stage. Hence, deriving an optimal design in a practical situation requires that estimates of a number of quantities be available from previous work or from pilot data. This is similar to the usual problem of calculating sample size, where estimates of prevalence or variance are required. In the examples presented in the following section, pilot samples are taken in order to estimate the components that determine the optimal second-stage sampling fractions. The resulting designs, which are approximations of the true optimal designs, result in significant improvements in precision or reductions in cost.

DATA EXAMPLES

To illustrate how the sampling strategies above might be used in practice, we will use data from the Coronary Artery Surgery Study (CASS) registry (4, 10, 11) in the following examples. Briefly, this registry collected detailed information regarding treatment and outcome on approximately 25,000 patients who underwent coronary arteriography at 15 cooperating sites in the United States and Canada. Enrollment began in 1974 and continued until May 1979. Follow-up data were collected through May 1983. We will limit our attention to a subset of 8,096 patients who underwent bypass surgery and consider operative mortality the outcome of interest. The number and proportion of these 8,096 subjects in each of the strata defined by sex and mortality is given in table 1.

Example 1

Let us suppose that only the information presented in table 1 is available. It is now proposed that we ascertain the age for a second-stage subsample of subjects so that age and sex can be assessed as independent risk factors for mortality. Let us suppose that sampling of these second-stage data involves nontrivial effort (e.g., chart review) and that resources are available for sampling only 1,000 subjects. If the second-stage sample is a simple random sample, it can be analyzed using standard methods to obtain asymptotically unbiased estimates. If the second-stage sample is a stratified random sample from the strata defined by sex and mortality, a mean-score analysis could be used to obtain valid estimates. The question arises as to how this second-stage sample might be chosen in order to maximize the precision of the estimates of interest.

The proportion of the 8,096 subjects in each (Z,Y) cell provides a good estimate of the prevalence, ρ_{ZY}. To estimate the other components of the optimal sampling fractions, it is reasonable to use a small portion of the available resources to take a pilot sample from the four strata. Taking a random sample of 25 subjects from each cell, a mean-score analysis using logistic regression to model operative mortality as a function of age and sex gives coefficients β_0(intercept) = -5.06, $\beta_{\text{age}} = 0.022$, and $\beta_{\text{sex}} = 0.674$. (This analysis and the calculations described below were carried out using programs written in the GAUSS (Aptech Systems Inc., Kent, Washington) programming language.) The information matrix, I, can be estimated using the predicted values, $\hat{\pi}$, from the logistic model and the covariate information on the 100 subjects in the pilot sample. The (i, j) component of the information contribution of an individual is $\hat{\pi}(1 - \hat{\pi})X_iX_j$, where X_i and X_j denote the values of the ith and jth covariate for the individual, with $X_1 = 1$ (the intercept in the model), X_2 = age, and X_3 = sex. The information matrix I is simply a weighted average of the individual information contributions. The score vectors for each of the 100 subjects are calculated as shown above in "Notation and Methods"; the ith component of the score vector for an individual is $X_i(Y - \hat{\pi})$. The sample variance-covariance matrix of the score is calculated separately for each group of 25 subjects to give the $V(Z,Y)$ matrices in the weighting factors W_{ZY}. Assuming that the objective is to maximize the precision of the coefficient of sex (i.e., $[k,k] = [2,2]$), the optimal sampling fractions of equation 5 subject to $n = 8{,}096$ and $n_2 = 1{,}000$ can be calculated from the estimates of β, ρ_{ZY}, and $[W_{ZY}]_{22}$. The values of $[W_{ZY}]_{22}$ found from the above calculations are displayed in table 2. The denominator in the optimal sampling fractions of equation 5 is now seen to be

$$0.823\mathrm{X}\sqrt{0.0058} + 0.152X\sqrt{0.0203} + 0.018X\sqrt{13.392} + 0.007X\sqrt{15.659} = 0.1779.$$

The optimal sampling fractions can now be found simply by dividing $(n_2/n)\sqrt{[W_{ZY}]_{22}}$ by 0.1779, and these fractions are displayed in table 2. Two of these

TABLE 1. Distribution, by sex and operative mortality, of 8,096 patients from the Coronary Artery Surgery Study who underwent bypass surgery between 1974 and 1983

	Males (Z = 0)	Females (Z = 1)
Number of subjects		
Alive (Y = 0)	6,666	1,228
Deceased (Y = 1)	144	58
Proportion of subjects		
Alive (Y = 0)	0.823	0.152
Deceased (Y = 1)	0.018	0.007

TABLE 2. Calculation of optimal sampling fractions from a pilot sample of 100 subjects from the Coronary Artery Surgery Study*

	Males (Z = 0)	Females (Z = 1)
	Alive (Y = 0)	
n	6,666	1,228
ρ_{ZY}	0.823	0.152
$[W_{ZY}]_{22}$	0.0058	0.0203
ρ'_{ZY}	0.053	0.099
	Deceased (Y = 1)	
n	144	58
ρ_{ZY}	0.018	0.007
$[W_{ZY}]_{22}$	13.392	15.659
ρ'_{ZY}	2.541	2.748

* ρ_{ZY} and ρ'_{ZY} indicate the cell prevalences and sampling fractions, respectively.

values are greater than 1. The largest value is set equal to 1, and the remaining fractions are recalculated as explained above in "Optimal Sampling." The resulting value for ρ'_{01} = 2.838, so the cycle is repeated once more to obtain the final design shown in table 3. We see that the optimal strategy is to sample all of the subjects in the sparse cell and small proportions of the others. A second-stage sample of 1,000 subjects was selected from the cohort using this sampling strategy, and a simple random sample was also selected for comparison. The achieved sample sizes for the simple random sample are included in table 3. Table 4 presents the estimates and standard errors obtained from a mean-score analysis of these data, where an analysis of the random subsample only is included for comparison. As would be expected, inclusion of the first-stage data in the analysis of the random second-stage sample results in an improvement in the precision of the estimates. A further marked reduction in all standard errors is achieved by optimal sampling of the second-stage data. In fact, when optimally sampling 1,000 subjects at the second stage, the precision achieved is almost as good as that obtainable from an analysis of the full "population" of 8,096 subjects (standard errors = 0.540, 0.009, and 0.161).

TABLE 3. Optimal sampling fractions ρ'_{ZY} and sample sizes n_2^{opt} for a second-stage sample of 1,000 subjects from the Coronary Artery Surgery Study*

	Males (Z = 0)	Females (Z = 1)
	Alive (Y = 0)	
n	6,666	1,228
ρ'_{ZY}	0.089	0.166
n_2^{opt}	593	204
n_2^{ran}	842	127
	Deceased (Y = 1)	
n	144	58
ρ'_{ZY}	1.00	1.00
n_2^{opt}	144	58
n_2^{ran}	20	11

* The achieved sample sizes for a simple random sample of 1,000 are denoted by n_2^{ran}.

Example 2

As a less trivial example, we shall use the same data and consider the setup discussed by Cain and Breslow (4) where, in addition to mortality and sex, a surrogate covariate (categorical weight) is available at the first stage, while the true covariate (weight) is to be ascertained at the second stage, together with a number of other risk factors of interest. The breakdown of the first-stage sample by outcome and the six exposure strata (defined by sex and weight category) is presented in table 5, and as before, the proportion of subjects in each cell will be used below to estimate the prevalence ρ_{ZY}.

Let us assume that our objective is to fit the same model as in Cain and Breslow's paper, and that resources are available with which to sample 1,000 second-stage subjects in order to ascertain the detailed covariate information (weight, age, unstable angina, congestive heart failure score, left ventricular end diastolic blood pressure, and urgency of surgery). We will also assume that we are principally interested in blood pressure as an independent risk factor, so we wish to maximize the precision of this component of the vector of coefficient estimates. Let us suppose that it had been decided a priori to take a pilot sample of 10 subjects from each of the strata defined by the first-stage data. In other work (12, 13), samples of this size have been shown to give reasonable estimates. Taking all eight subjects from the stratum that has only eight and 10 subjects from each of the remaining strata, estimates of the coefficients in the logistic model are obtained using the mean-score method. These estimates, together with estimates of the W_{ZY}, are used to estimate the optimal sampling fractions for the second stage (see table 6), subject to an overall second-stage sample size of 1,000.

The optimal sampling strategy in this example is to sample at the second stage all of the cases and a varying proportion of controls in the various exposure strata. A second-stage sample was selected using this design, and the results of analysis are presented in table 7. The analysis of a simple random sample of 1,000 second-stage subjects is included for comparison. We see that the inclusion of the first-stage data in the analysis of the random second-stage sample results

TABLE 4. Coefficients (β) from 1) logistic regression analysis of the random subsample only, 2) mean score analysis of the random subsample, and 3) mean-score analysis of the optimal subsample

	Random subsample (n = 1,000)		Mean-score: random (n = 8,096, n_2 = 1,000)		Mean-score: optimal (n = 8,096, n_2 = 1,000)	
	β	SE*	β	SE	β	SE
Constant	−8.431	1.385	−8.734	1.400	−7.360	0.595
Age	0.081	0.023	0.085	0.023	0.061	0.010
Sex	1.030	0.401	0.496	0.192	0.645	0.164

* SE, standard error.

TABLE 5. Numbers of subjects from the Coronary Artery Surgery Study, stratified by operative mortality, sex, and weight

Sex and weight (kg)	Alive	Deceased
Males		
<60	160	8
60–70	1,083	33
≥70	5,418	103
Females		
<60	440	18
60–70	407	26
≥70	378	14
Total	7,886	202

TABLE 6. Optimal sampling fractions for a second-stage sample of 1,000 subjects from Coronary Artery Surgery Study

Sex and weight (kg)	Alive	Deceased
Males		
>60	0.092 (14)*	1.00 (8)
60–70	0.40 (418)	1.00 (30)
≥70	0.05 (262)	1.00 (101)
Females		
<60	0.118 (50)	1.00 (17)
60–70	0.101 (40)	1.00 (25)
≥70	0.064 (23)	1.00 (12)

* Numbers in parentheses, sample size.

in an improvement in the precision of all but one of the coefficients. Almost all of these standard errors can, however, be halved by an optimal choice of the 1,000 subjects at the second stage. In addition, the optimal sampling scheme enables us to detect a significant effect due to left ventricular end diastolic blood pressure which was not apparent from the random second-stage data.

Example 3

In the above examples, the acquisition of the second-stage sample is a matter of subsampling from a fixed number of available first-stage subjects. Another realistic setting is one where the design of the study (both first and second stage) is determined prior to data collection and is conditional on budgetary constraints. To examine such a situation, let us consider again the model in example 2 and derive the total study size and second-stage sampling fractions that will optimize the precision of the effect of left ventricular end diastolic blood pressure. The optimal study design for a total budget of $10,000 will be compared under two cost structures:

1. Assume that the cost of each first-stage observation is $1.00, while gathering the second-stage data costs an additional $0.50 each. This represents a situation where access to the study subjects is the main cost, while the additional cost per covariate ascertained is comparatively low.
2. Assume that the cost of a first-stage observation is $1.00 but the cost of a second-stage observation is an additional $5.00. This represents a situation where first-stage data are readily available (e.g., in a database), but the collection of second-stage data involves considerable effort and expense.

Using the same pilot sample of 10 observations per stratum as in the previous example, the optimal sampling schemes for these two cost structures are found from the equations in expression 6 and are presented in table 8. We see that for situation 1, the optimal design requires that a total of 8,854 subjects be enrolled in the study and that full covariate information be obtained on all cases and most (91 percent) of the male controls in the 60- to 70-kg weight category, with a lower sampling intensity (11–27 percent) for the remaining controls. In contrast, if the detailed covariate information is expensive as in situation 2, the optimal scheme is for a smaller study (n = 6,602) and less intensive sampling of controls at the second stage (in this example, the sampling intensity of controls is approximately one third that found in situation 1).

TABLE 7. Coefficients (β) from 1) logistic regression analysis of the random subsample only, 2) mean score analysis of the random subsample, and 3) mean-score analysis of the optimal subsample

	Random subsample (n = 1,000)		Mean-score: random (n = 8,088, n_2 = 1,000)		Mean-score: optimal (n = 8,088, n_2 = 1,000)	
	β	SE*	β	SE	β	SE
Sex	0.745	0.569	0.200	0.265	0.279	0.216
Weight	0.0002	0.021	−0.014	0.014	−0.012	0.008
Age	0.064	0.030	0.055	0.026	0.052	0.012
Unstable angina	1.391	0.527	1.260	0.416	0.170	0.311
CHF* score	0.275	0.259	0.401	0.198	0.262	0.097
LVEDBP*	0.002	0.033	−0.013	0.024	0.020	0.012
Urgency of surgery	0.947	0.481	0.791	0.496	1.053	0.208

* SE, standard error; CHF, congestive heart failure; LVEDBP, left ventricular end diastolic blood pressure.

TABLE 8. Optimal first-stage sample sizes and second-stage sampling fractions for a budget of \$10,000, where C_1, the cost per first-stage observation, is \$1.00 and C_2 denotes the cost per observation at the second stage

Sex and weight (kg)	$C_2 = \$0.50$ ($n = 8,854$)		$C_2 = \$5.00$ ($n = 6,602$)	
	Alive	Deceased	Alive	Deceased
Males				
<60	0.21	1.0	0.07	1.0
60–70	0.91	1.0	0.30	1.0
≥70	0.11	1.0	0.04	1.0
Females				
<60	0.27	1.0	0.09	1.0
60–70	0.23	1.0	0.08	1.0
≥70	0.14	1.0	0.05	1.0

DISCUSSION

The methods presented in this paper will enable investigators to design two-stage studies that will estimate the effect of a risk factor as precisely and economically as resources allow. Both case-control and cohort data are accommodated at the first stage. In addition to study design, this method could also be useful in a completed (or ongoing) study where important covariate information is missing for a large number of subjects and the information in the sample is inadequate to answer the research question of interest. By considering the incomplete observations as the second-stage sample, optimal sampling methods could be used to select a subsample of subjects on whom to ascertain data for the missing covariates. In this way, the study can be made to yield a useful result for a minimal additional cost.

Although a logistic regression model will often be of interest for epidemiologic applications, the above methods can accommodate any likelihood function for a categorical outcome variable, and thus are applicable to a wide variety of estimation problems. Further work is required to extend the methods to accommodate continuous covariates at the first stage, and to derive designs that will offer simultaneous optimal estimation of more than one parameter. It is encouraging to note in the data examples shown that the designs which were optimal with respect to one parameter achieved an improvement in the precision of almost all parameters.

REFERENCES

1. Meydrech EF, Kupper LL. Cost considerations and sample size requirements in cohort and case-control studies. Am J Epidemiol 1978;107:201-5.
2. Nam JM, Fears TR. Optimum allocation of samples in strata-matching case-control studies when cost per sample differs from stratum to stratum. Stat Med 1990;9:1475–83.
3. Zhao LP, Lipsitz S. Designs and analysis of two-stage studies. Stat Med 1992;11:769–82.
4. Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. Am J Epidemiol 1988;128: 1198–206.
5. Flanders WD, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. Stat Med 1991;10:739–47.
6. Buonaccorsi JP. Double sampling for exact values in some multivariate measurement error problems. J Am Stat Assoc 1990;85:1075–82.
7. Tosteson TD, Ware JH. Designing a logistic regression study using surrogate measures for exposure and outcome. Biometrika 1990;77:11–21.
8. Reilly M, Pepe MS. A mean-score method for missing and auxiliary covariate data in regression models. Biometrika 1995;82:299–314.
9. Kahn HA, Sempos CT. Statistical methods in epidemiology. New York, NY: Oxford University Press, 1989.
10. Vlietstra RE, Frye RL, Kronmal RA, et al. Risk factors and angiographic coronary artery disease: a report from the Coronary Artery Surgery Study (CASS). Circulation 1980;62: 254–61.
11. Fisher LD, Kennedy JW, Davis KB, et al. Association of sex, physical size, and operative mortality after coronary artery bypass surgery in the Coronary Artery Surgery Study (CASS). J Thorac Cardiovasc Surg 1982;84:334–41.
12. Pepe MS, Reilly M, Fleming TR. A nonparametric method for dealing with mismeasured covariate data. J Stat Plann Infer 1994;42:137–60.
13. Reilly M. Semi-parametric methods of dealing with missing or surrogate covariate data. (Ph.D. thesis). Seattle, WA: University of Washington, 1991.
14. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc [B] 1977;39:1–38.

APPENDIX 1

Asymptotic Results

If complete information on X was available for all subjects, the maximum likelihood estimator of β would be found by solving the score equation

$$\sum_{i=1}^{n} S_\beta(Y_i|X_i) = 0,$$

where $S_\beta(Y|X) = (\partial/\partial\beta) \log P_\beta(Y|X)$ and n is the total number of subjects in the study. The expectation-maximization (EM) algorithm (14) for finding the maximum likelihood estimator from incomplete data involves the maximization (conditional on the available data) of the expected log likelihood—i.e., the iterative maximization of

$$\sum_{i\in S_2} \log P_\beta(Y_i|X_i) + \sum_{j\notin S_2} E[\log\{P_\beta(Y_j|X)\}|\beta^c, Y_j, Z_j],$$

where S_2 denotes the second-stage sample and β^c the current estimate. The mean-score method estimates the expectation in the second term by the sample average, so the mean-score estimate solves the equation

$$\sum_{i\in S_2} S_\beta(Y_i|X_i) + \sum_{j\notin S_2}\left\{\sum_{i\in S_2^{Z_jY_j}} \frac{S_\beta(Y_j|X_i)}{n_2(Z_j,Y_j)}\right\} = 0,$$

where $S_2^{Z_jY_j}$ denotes the members of S_2 with $Z = Z_j$ and $Y = Y_j$ and $n_2(Z_j,Y_j)$ denotes the number of such subjects. This estimating equation can also be written as

$$\sum_{i\in S_2} \frac{n(Z_i,Y_i)}{n_2(Z_i,Y_i)} S_\beta(Y_i|X_i) = 0, \tag{1}$$

where $n(Z_i,Y_i)$ is the total number of subjects with $Z = Z_i$ and $Y = Y_i$. In the paper by Reilly and Pepe (8), the asymptotic distribution of the mean-score estimator, $\hat{\beta}$, is shown to be normal with mean β and variance

$$V(\beta) = I^{-1} + I^{-1}\mathcal{V}I^{-1}, \tag{2}$$

where the information $I = E[-(\partial^2/\partial\beta^2)\log\{P_\beta(Y|X)\}]$ and

$$\mathcal{V} = \sum_{ZY} \frac{\rho_{ZY}(1-\rho'_{ZY})}{\rho'_{ZY}} \operatorname{Var}\{S_\beta(Y|X)|Y, Z\}, \tag{3}$$

with summation being over distinct (Z,Y) strata. The proportion of subjects in the stratum (Z,Y) is denoted by ρ_{ZY}, and ρ'_{ZY} denotes the second-stage sampling fraction of those subjects. A consistent estimator of the variance of the estimate is given by

$$\hat{V}(\hat{\beta}) = \frac{1}{n}(\hat{I}^{-1} + \hat{I}^{-1}\hat{\mathcal{V}}\hat{I}^{-1}), \tag{4}$$

where

$$\hat{I} = \frac{1}{n}\sum_{i\in S_2}\frac{n(Z_i,Y_i)}{n_2(Z_i,Y_i)} I_\beta(Y_i|X_i),\ \hat{\mathcal{V}} = \frac{1}{n}\sum_{Z,Y}\frac{n(Z,Y)n_1(Z,Y)}{n_2(Z,Y)}\operatorname{Var}[S_\beta(Y|X)|Y,Z],$$

and $n_1(Z,Y) = n(Z,Y) - n_2(Z,Y)$. "Var" indicates the sample variance, and $I_\beta(Y_i|X_i)$ is the usual information, $-(\partial^2/\partial\beta^2)\log P_\beta(Y|X)$.

APPENDIX 2

Optimal Designs

Fixed *n* and n_2

Using the derivative of $V(\beta_k)$ with respect to ρ'_{ZY} and a Lagrangian multiplier to accommodate the condition $\sum \rho_{ZY}\rho'_{ZY} = (n_2/n)$, the value of ρ'_{ZY} which minimizes $V(\beta_k)$ subject to fixed n and n_2 is found to be

$$\rho'_{ZY} = \frac{\frac{n_2}{n}\sqrt{[W_{ZY}]_{kk}}}{\sum_{Z,Y}\rho_{ZY}\sqrt{[W_{ZY}]_{kk}}}, \tag{5}$$

where $[\ ]_{kk}$ denotes the (k,k) element of a matrix, $W_{ZY} = I^{-1}V(Z,Y)I^{-1}$, and $V(Z,Y) = \mathrm{Var}[S_\beta(Y|X)|Y,Z]$.

Fixed budget *B*

Minimization of $V(\beta_k)$ with respect to n and ρ'_{ZY} subject to $B = n(C_2 \sum \rho_{ZY}\rho'_{ZY} + C_1)$ results in three equations in n, ρ'_{ZY} and the Lagrangian multiplier. Solving for n and ρ'_{ZY} yields

$$n = B\left\{C_1 + \frac{\sqrt{C_1C_2}\sum_{Z,Y}\rho_{ZY}\sqrt{[W_{ZY}]_{kk}}}{[I^{-1}]_{kk} - \sum_{Z,Y}\rho_{ZY}[W_{ZY}]_{kk}}\right\}^{-1} \tag{6}$$

and

$$\rho'_{ZY} = \left(\frac{B - nC_1}{nC_2}\right)\frac{\sqrt{[W_{ZY}]_{kk}}}{\sum_{Z,Y}\rho_{ZY}\sqrt{[\mathrm{W}_{\mathrm{ZY}_{kk}}}}.$$

Fixed precision

Denoting the fixed standard deviation of β_k by δ, from expression 2

$$\delta = \frac{1}{n}[I^{-1}]_{kk} + \frac{1}{n}\sum \rho_{ZY}\frac{\rho_{ZY}(1 - \rho'_{ZY})}{\rho'_{ZY}}[W_{ZY}]_{kk}.$$

Differentiation with respect to n and ρ'_{ZY} yields two further equations, where again a Lagrangian multiplier accommodates the restriction on δ. Solving for n and ρ'_{ZY} gives

$$n = \frac{[I^{-1}I_1I^{-1}]_{kk}}{\delta} + \sqrt{\frac{C_2}{C_1\delta^2}}\sum_{Z,Y}\frac{n(Z,Y)}{n}\sqrt{[W_{ZY}]_{kk}}\sqrt{[I^{-1}I_1I^{-1}]_{kk}} \tag{7}$$

and

$$\rho'_{ZY} = \sqrt{\frac{C_1}{C_2}}\sqrt{\frac{[W_{ZY}]_{kk}}{[I^{-1}I_1I^{-1}]_{kk}}},$$

where $I_1 = I - \sum_{Z,Y}\rho_{ZY}V(Z,Y)$, the information in the first-stage data.