AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 128, No. 6
Printed in U.S.A.

LOGISTIC REGRESSION ANALYSIS AND EFFICIENT DESIGN FOR TWO-STAGE STUDIES[1]

KEVIN C. CAIN AND NORMAN E. BRESLOW

This commentary concerns epidemiologic studies in which the disease status and the exposure of primary interest are ascertained for a large group of subjects. However, information about covariables, needed to control the relation between exposure and disease for confounding effects, is collected only for a smaller sample that may be limited in size by the costs of subject selection and covariable measurement or because the requisite data are missing from the records. Two questions are raised: 1) How does one conduct the statistical analysis so as to make use of all available information for the estimation of covariable-adjusted exposure effects? and 2) How does one select the subsample, if indeed it is within the investigator's power to do so, so as to maximize the amount of information it provides? Four examples will illustrate the fundamental problem.

1. *Occupational exposures.* Suppose that employment records and basic medical records are available for all employees of a large chemical company. From these records, it is relatively easy to determine whether an employee has had substantial exposure to a particular toxic chemical and whether or not he or she has developed the disease in question. However, personal interviews are needed to determine cigarette and alcohol consumption, and because of cost considerations, it is only possible to conduct such interviews (in person or by proxy) for a sample of the chemical workers.

2. *Secondary analysis.* Suppose that data from a large case-control study show what appears to be a significant relation between a risk factor and the disease. However, some important confounding risk factors were not measured in the original study, either by accidental omission or because the data were initially gathered for another purpose. We wish to go back and collect the needed information on a subsample of the original study group.

3. *Two-stage case-control design.* Suppose that the cost of ascertaining exposure and outcome is inexpensive (e.g., telephone interviews) relative to the cost of obtaining detailed covariate measurements (e.g., in-person interviews). Then a two-stage design in which exposure and outcome are determined for a large sample but covariates are only measured on a smaller sample may be much cheaper than a one-stage design of comparable power.

4. *Missing data.* Suppose that the data come from a large disease register and that the outcome and the variable of interest are known for virtually all subjects. However, one of the important covariables is missing for the majority of the subjects. If we are willing to make certain assumptions about the randomness of the missing observations which are considerably less stringent than that they be missing "at random" for the sample as a whole, then we can use the methods described below to analyze the data.

A valid strategy to follow in the first three examples would be to select random samples of the diseased and nondiseased subjects from the original sample and then analyze the sampled data using logistic regression as if these were the only data available. This will result in unbiased estimates of exposure effects, but it may be

[1] Department of Biostatistics, SC-32, University of Washington, Seattle, WA 98195. (Reprint requests to Dr. Kevin C. Cain.)

This work was supported in part by USPHS grant no. RO1-CA40644.

The authors thank the Coronary Artery Surgery Study investigators for allowing use of their data.

highly inefficient, for two reasons. First, it ignores the extra information from the first-stage sample, namely, the disease and exposure status of all study subjects, including those for whom the values of the covariables are unknown. Second, this type of second-stage sample (referred to below as the case-control sample) may not be the most efficient design. For example, if exposure is relatively rare, this type of sample will give a very small number of exposed cases and controls, and a more powerful analysis may be achieved by oversampling the exposed cases.

White (1) addressed the question of how to analyze such two-stage data for the restricted situation in which there is a single binary exposure variable and the confounding variables are discrete with a finite number of values (see also Walker (2)). Her solution is based on a weighted least-squares approach. The solution presented here, based on likelihood methods using logistic regression, allows multiple exposure and confounding variables, some or all of which can be continuous-valued. If the exposure variable is discrete and a separate relative risk is to be estimated for each exposure category, this method may be implemented easily by making some simple adjustments to the output from a standard logistic regression program.

Notation and analysis

At the first stage of data collection, N_1 cases (with $D = 1$ denoting the presence of disease) and N_0 controls (with $D = 0$) are sampled without knowledge of exposure and are then classified into J exposure classes or strata ($S = 1, 2, \ldots J$). Let N_{ij} denote the number of subjects with disease status $D = i$ and exposure stratum $S = j$. The marginal or unadjusted relation between disease and exposure stratum can be estimated from the data available at this stage as $OR_j = (N_{1j}N_{01})/(N_{0j}N_{11})$, where OR_j denotes the odds ratio for disease in exposure stratum j compared with exposure stratum 1. These odds ratios are not adjusted for possibly confounding covariables, however, and a second-stage sample is drawn in which these covariables can be measured.

At the second stage, n_{ij} of the N_{ij} observations are sampled randomly for each disease-exposure cell ($i = 0, 1; j = 1, 2, \ldots J$). The sampling fraction n_{ij}/N_{ij} can be different in the different cells.

For each observation in the second-stage sample, we measure the covariables and perhaps even obtain a more detailed measurement of exposure. Let x_{ijk} denote the vector of regression variables for the kth observation within the ijth disease-exposure cell ($k = 1, 2, \ldots n_{ij}$) that will be included in our logistic regression model: a constant term, an exposure variable or variables (continuous or else indicators to represent discrete categories), the covariables, and any interaction terms, including any interactions between exposure and covariables. Our model for the population is $\Pr(D = 1 \mid x) = 1/(1 + \exp(-x'\beta))$, which is the standard logistic regression model relating the probability of disease to a vector of covariates x.

Methods of analysis of data from a single case-control sample using the logistic model are well known to biostatisticians and epidemiologists (3). The situation here is more complicated since the sampling fractions at the second stage may depend on both disease and exposure. Methods of analysis for such stratified samples have been explored mainly by econometricians (4, 5). An extension of this approach that is applicable to the present two-stage sampling problem was developed by Breslow and Cain (6). We present here a description and illustration of the analysis method derived therein, referring readers who wish a rigorous presentation of the mathematical derivation to the original article. A similar method is presented by Fears and Brown (7), but their analysis does not account for the extra information from the first-stage sample, and hence the estimated variances are incorrect.

Let us start with the simple situation in which exposure is modeled as a categorical variable with categories corresponding to strata from which the second-stage samples

were drawn. The first J components of the vector x for a given subject are $x_1 = 1$, the constant term; $x_2 = 1$ if the subject is in exposure stratum 2, and $x_2 = 0$ otherwise; ... $x_J = 1$ if the observation is in stratum J, and $x_J = 0$ otherwise. Thus, for $j = 2, \ldots J$, β_j is the log odds ratio for exposure category j compared with exposure category 1. In this situation, one may use a standard logistic regression program with the data collected at the second stage to estimate $\hat{\beta}$ and its covariance matrix and then make simple adjustments to get the corrected values, namely,

$$\hat{\beta}_j^{(\text{Adj})} = \hat{\beta}_j^{(\text{Unadj})} + \log\left(\frac{N_{1j}N_{01}n_{11}n_{0j}}{N_{11}N_{0j}n_{1j}n_{01}}\right)$$

for $j = 2, \ldots J$, and

$$\hat{\beta}_1^{(\text{Adj})} = \hat{\beta}_1^{(\text{Unadj})} + \log\left(\frac{N_{11}n_{01}}{N_{01}n_{11}}\right),$$

although this coefficient for the constant term is not really meaningful if the first-stage sample is a case-control sample. As shown by Breslow and Cain (6), no adjustment is necessary to the other components of $\hat{\beta}$, that is, those that correspond to covariables in the model or to any interactions between exposure effects and covariables.

The estimated standard error of $\hat{\beta}_j$ can also be adjusted easily. Define

$$c^* = (1/n_{01} - 1/N_{01}) + (1/n_{11} - 1/N_{11})$$

and

$$c_{jj} = c^* + (1/n_{0j} - 1/N_{0j}) + (1/n_{1j} - 1/N_{1j}),$$

for $j = 2, \ldots J$. Let V_{jj} denote the square of the estimated standard deviation of $\hat{\beta}_j$ obtained from the logistic regression program. Then, for $j = 2, \ldots J$,

$$V_{jj}^{(\text{Adj})} = V_{jj}^{(\text{Unadj})} - c_{jj}.$$

If there are more than two exposure categories, the covariance between the components of $\hat{\beta}$ corresponding to two of the indicator variables for exposure categories (call them $\hat{\beta}_j$ and $\hat{\beta}_{j'}$) is adjusted as

$$V_{jj'}^{(\text{Adj})} = V_{jj'}^{(\text{Unadj})} - c^*.$$

The adjusted variance of the coefficient of the constant term is

$$V_{11}^{(\text{Adj})} = V_{11}^{(\text{Unadj})} - (c^* + 1/N_0 + 1/N_1),$$

and the covariance between $\hat{\beta}_1$ and $\hat{\beta}_j$ is

$$V_{1j}^{(\text{Adj})} = V_{1j}^{(\text{Unadj})} + c^*, \quad \text{for} \quad j = 2, \ldots J.$$

The standard errors for the components of $\hat{\beta}$ corresponding to covariables measured at the second stage do not need adjustment, nor do the covariances between exposure variables and covariables.

The standard error of $\hat{\beta}_j$ (for $j = 2, \ldots J$) is smaller after adjustment because the unadjusted value uses only data from the smaller second-stage sample, while the adjusted $\hat{\beta}_j$ also uses information from the first stage to give a more accurate estimate. It is plausible that the other components of $\hat{\beta}$ do not require adjustment, since only the second-stage sample contains relevant information about these parameters.

If exposure is modeled as a continuous-valued variable or variables, it is unfortunately no longer possible to make simple adjustments to the output from a standard logistic regression program. All components of $\hat{\beta}$ and its estimated covariance matrix must be adjusted by means of matrix calculations. The correct method of analysis for this situation is described in the Appendix.

Examples

Fisher et al. (8) used data from the Coronary Artery Surgery Study to examine the relation between sex and operative mortality in patients undergoing coronary artery bypass surgery. Female sex appeared to be a strong risk factor, with operative mortality being 1.9 per cent and 4.5 per cent for men and women, respectively. However, after adjustment for other covariables which measured the severity of disease and the size of the patient, sex was no longer statistically significant.

We use these data from the Coronary Artery Surgery Study to illustrate the two-

stage design, with operative mortality as the outcome and sex as the exposure variable. Our analysis is clearly artificial in that it only uses covariable information from the patients in the second-stage sample, although in fact the covariables are known for the entire first-stage sample. Such two-stage data could easily arise in practice, however. Operative mortality and the exposure of interest (e.g., sex of patient, use of internal mammary or vein grafts, type of anesthesia) may be easily available for a large number of patients, while the covariables are expensive to measure.

Table 1 shows the number of subjects in the first-stage sample and in two different second-stage samples. In one second-stage design, referred to as the case-control sample, 100 cases (i.e., who died during surgery) and 100 controls are randomly selected without regard to sex. The fraction of cases (and also controls) who are female is approximately the same as in the first-stage sample. In the second-stage design, referred to as the balanced design, 50 subjects are randomly selected from each of the four sex-mortality categories.

The unadjusted estimate of $\hat{\beta}_2$ from the case-control sample gives an unbiased estimate of the true β_2 since unbiased samples of cases and controls were taken (notice that the adjustment to $\hat{\beta}_2$ is small). The unadjusted $\hat{\beta}_2$ from the balanced sample, however, gives a biased estimate of β_2 since biased samples of cases and controls were taken (notice that the adjustment to $\hat{\beta}_2$ is large). Thus, a valid analysis could be obtained from the unadjusted estimates from the case-control sample shown in table 2. However, a great deal of power would be lost if the information from the first-stage sample were not used. Notice that the adjusted standard errors are much smaller than the unadjusted ones, although they are not as small as the standard error from the full first-stage sample.

We next present an example in which there are six exposure strata, defined by sex and three weight categories, as shown in table 3. A balanced second-stage sample was selected with 20 subjects in each of the 12 mortality-weight-sex categories. Some categories have fewer than 20 subjects since

Table 2 shows the results of fitting a logistic regression model to each of the second-stage samples and, for comparison, to the first-stage sample. For each second-stage sample, the column labeled "unadjusted" is the output from a standard logistic regression program run on the second-stage sample. Only the values in the first two rows, corresponding to the constant term and to sex, need to be adjusted to account for the information from the first-stage sample. For example, consider the adjustment to the coefficient for sex in the case-control sample:

$$\hat{\beta}_2^{(\text{Adj})} = \hat{\beta}_2^{(\text{Unadj})} + \log((N_{12}N_{01}n_{11}n_{02})/(N_{11}N_{02}n_{12}n_{01}))$$

$$= 0.650 + \log((58 \times 6{,}666 \times 67 \times 19)/(144 \times 1{,}228 \times 33 \times 81))$$

$$= 0.650 + 0.040 = 0.690.$$

The adjustment to the estimated variance is

$$V_{22}^{(\text{Adj})} = V_{22}^{(\text{Unadj})} - c_{22}$$

$$= V_{22}^{(\text{Unadj})} - ((1/n_{01} - 1/N_{01}) + (1/n_{11} - 1/N_{11}) + (1/n_{02} - 1/N_{02}) + (1/n_{12} - 1/N_{12}))$$

$$= (0.348)^2 - ((1/81 - 1/6{,}666) + (1/67 - 1/144) + (1/19 - 1/1{,}228) + (1/33 - 1/58))$$

$$= 0.121 - 0.085 = 0.036,$$

and the adjusted standard error is $(0.036)^{1/2} = 0.190$.

TABLE 1

Sample sizes for Coronary Artery Surgery Study (8) data relating operative mortality to sex in patients undergoing coronary artery bypass surgery

	Male ($S = 1$)	Female ($S = 2$)
First-stage sample		
Alive ($D = 0$)	6,666	1,228
Deceased ($D = 1$)	144	58
Second-stage sample: case-control		
Alive	81	19
Deceased	67	33
Second-stage sample: balanced		
Alive	50	50
Deceased	50	50

the number of subjects in the first-stage sample with no missing covariables was less than 20. Table 4 shows the results of a logistic regression analysis in which weight is modeled as a linear term. Since the model does not include five indicator variables for the six exposure strata, it is not possible to use the simple calculations shown above to get the adjusted values. Instead, the matrix calculations shown in the Appendix must be used. The adjusted standard error is smaller than the unadjusted one for the two exposure variables, sex and weight. The adjustment also changes slightly the coefficient and standard errors of the other covariables.

In this second example, a crude measure of exposure (i.e., weight category) is used

TABLE 2

Logistic regression coefficients (and standard errors) for the data in table 1

	First-stage sample	Second-stage sample: case-control		Second-stage sample: balanced	
		Unadjusted	Adjusted	Unadjusted	Adjusted
Constant	−3.271 (0.285)	−0.167 (0.615)	−3.812 (0.594)	0.990 (0.637)	−2.845 (0.606)
Female sex	0.634 (0.171)	0.650 (0.348)	0.690 (0.190)	−0.061 (0.301)	0.722 (0.189)
Average diameter of coronary arteries	−0.065 (0.016)	−0.030 (0.034)		−0.080 (0.033)	
Congestive heart failure score	0.445 (0.072)	0.395 (0.160)		0.348 (0.165)	
Priority of surgery relative to elective					
Urgent	0.706 (0.181)	0.631 (0.365)		0.412 (0.350)	
Emergency	2.004 (0.232)	2.605 (1.072)		1.853 (0.800)	

TABLE 3

Sample sizes for Coronary Artery Surgery Study (8) data relating operative mortality to sex and weight

	Weight <60 kg		Weight 60–70 kg		Weight >70 kg	
	Male	Female	Male	Female	Male	Female
First-stage sample						
Alive	160	440	1,083	407	5,418	378
Deceased	8	18	33	26	103	14
Second-stage sample: balanced						
Alive	20	20	20	20	20	20
Deceased	8	16	20	20	20	11

TABLE 4

Coefficients (and standard errors) from a logistic regression analysis using a continuous-valued exposure variable, based on data in table 3

	Second-stage sample: balanced	
	Unadjusted	Adjusted
Constant	−4.552 (1.502)	−6.316 (1.428)
Female sex	−0.122 (0.305)	0.311 (0.233)
Weight	−0.001 (0.014)	−0.002 (0.011)
Age	0.037 (0.018)	0.038 (0.018)
Unstable angina	0.315 (0.262)	0.321 (0.265)
Congestive heart failure score	0.253 (0.148)	0.292 (0.153)
Left ventricular end diastolic blood pressure	0.041 (0.021)	0.041 (0.021)
Urgency of surgery	0.741 (0.251)	0.727 (0.256)

to define the strata in the first-stage sample, while a more detailed measure (i.e., weight) is used in the analysis of the second-stage data. This same approach can be used if the crude measure is known for all subjects in the first-stage sample but the detailed measure is only known for subjects in the second-stage sample. However, the analysis will be appropriate only if the relation between risk and exposure is correctly modeled and risk does not depend on the crude measure, conditional on the detailed measure being known. In the second example above, this will be true if the relation between risk and weight is in fact linear on the logistic scale and there is no interaction between sex and weight.

EFFICIENCY OF SECOND-STAGE DESIGNS

An efficient design is one that gives accurate estimates of the parameters in the model. Thus, the efficiency of a given design can be measured as the reciprocal of the standard error of the parameters of interest. An example of relative efficiencies can be seen in table 2, which shows that the standard error of the coefficient for sex is much smaller in the adjusted analysis of the case-control design (0.190) than in the unadjusted analysis of this design (0.348). Thus, the adjusted analysis is much more efficient than the unadjusted analysis, even though, as noted above, both analyses are valid. It is clear that failure to use the information in the first-stage sample can lead to a large loss of efficiency. How large this loss is will depend on how large the first-stage sample is relative to the second-stage sample. Extreme loss of efficiency will only occur if N_{ij}/n_{ij} is very large for all i and j. With a rare disease, it is probably more common to have a design in which the second-stage sample consists of all the cases from the first stage, plus a small fraction of the controls. Thus, N_{0j}/n_{0j} is large but N_{1j}/n_{1j} equals 1 for all j. In this situation, the ratio of standard errors between the adjusted and the unadjusted analyses will be approximately $(\frac{1}{2})^{1/2}$. The example in tables 1 and 2 shows an efficiency ratio that is somewhat better than this since N_{1j}/n_{1j} is somewhat bigger than 1 for $j = 1, 2$.

Suppose that n, the total number of subjects to be selected for the second-stage sample, is fixed by budgetary constraints. How should these n subjects be allocated among the $2J$ disease-exposure categories? Define the balanced design as follows: Choose equal numbers in each category (i.e., $n_{ij} = n/2J$ for $i = i, 2; j = 1, \ldots J$) if possible. If $N_{ij} < n/2J$ for any i, j, choose $n_{ij} = N_{ij}$ and increase the other n_{ij}'s approximately equally. We propose that this balanced design should usually be used since it is easy to define and has good efficiency.

Since we are most interested in evaluating the relation between exposure and disease, the main parameters of interest are those corresponding to the exposure variables. Of secondary interest are the parameters corresponding to the other covariables in the model and any possible interactions between these covariables and exposure. The balanced design is now compared with two alternative designs: the case-control design and the optimal design.

The adjusted standard errors in table 2 show that the case-control and balanced designs have approximately equal efficiency with respect to the parameter for

sex. The two designs also have similar efficiencies for estimating the other covariables in the model. Thus, in this example, the two designs have very similar efficiencies, but is this true in general? Breslow and Cain (6) examine this question by comparing the large sample efficiencies of the balanced design and the case-control designs for the simple situation with two exposure categories and one binary covariate. The results are shown in table 5. If the second-stage sample contains all of the cases from the first stage, then with respect to the exposure coefficient, the balanced design is either somewhat more efficient or approximately the same as the case-control design, depending on the relative risks associated with exposure and the confounder and the correlation between exposure and the confounder.

If the second-stage sample contains only a small fraction of both cases and controls from the first-stage sample, the balanced design is usually much more efficient than the case-control design. The only situation in which the case-control design is somewhat more efficient is if the relative risk for the confounder is one and there is a large negative correlation between exposure and the confounder.

With respect to the coefficient on the covariable, the case-control design is usually somewhat more efficient than the balanced design. However, in terms of the standard error of an interaction term (if one were to be included in the model), the balanced design is always much more efficient than the case-control design. Therefore, unless one is willing to make the very strong assumption that no interaction is present, the balanced design is to be preferred.

Breslow and Cain (6) also compare the balanced design with the design which has smallest standard error for estimating the relative risk of exposure and show that, in most circumstances of practical interest, the balanced design is not seriously inefficient compared with the theoretical optimum. In addition, the optimal design is usually degenerate in that $n_{ij} = 0$ for at least one i and j. (This is not as unreasonable as it at first seems, since the purpose of the second-stage sample is to estimate the effect of the covariable, and not all four of the disease-exposure categories are necessary to do this in a model with no interaction term.) Consequently, a degenerate optimal design would never actually be used in practice, especially since it has no power for testing whether an interaction term should be included in the model.

TABLE 5

Large sample efficiencies of the balanced design relative to the case control design[*,†,‡]

e^{β_2}	θ	Relative efficiencies		
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Second-stage sample contains all the cases but only a fraction of the controls from the first-stage sample				
0.2	0.2	1.02	0.83	1.43
0.2	1.0	1.18	0.88	1.45
0.2	5.0	1.34	0.93	1.65
1.0	0.2	0.99	0.74	2.30
1.0	1.0	1.00	0.81	2.09
1.0	5.0	0.98	0.82	2.06
5.0	0.2	1.14	0.76	2.94
5.0	1.0	1.22	0.81	2.12
5.0	5.0	1.00	0.76	1.74
Second-stage sample contains a small fraction of both cases and controls from the first-stage sample				
0.2	0.2	1.37	0.68	4.41
0.2	1.0	4.35	1.01	3.51
0.2	5.0	3.56	1.47	3.30
1.0	0.2	0.71	0.71	5.77
1.0	1.0	1.00	1.01	4.08
1.0	5.0	1.05	1.05	3.90
5.0	0.2	1.84	0.78	6.48
5.0	1.0	3.72	1.01	4.10
5.0	5.0	1.36	0.83	3.44

* Abstracted from table 3 of Breslow and Cain (6).

† The coefficients $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ correspond to exposure, the covariable, and the interaction, respectively, and θ is the degree of confounding between exposure and the covariable, as measured by the odds ratio in the control group.

‡ Efficiencies are calculated for $\exp(\beta_1) = 2$, $\Pr(x_1 = 1|D = 0) = 0.05$, and $\Pr(x_2 = 1|D = 0) = 0.3$. Efficiencies for $\hat{\beta}_1$ and $\hat{\beta}_2$ are based on fitting the model with no interaction term (x_3).

CONCLUSIONS

We have proposed a method of analysis for use with data in which the exposure of interest and disease outcome are known for a large number of subjects but information on confounding variables is only known for a smaller subsample. This type of data arises frequently in practice, particularly in studies of occupational health. The main advantage of the proposed analysis method is that it incorporates information from the first-stage sample as well as the second-stage sample. This can lead to a very large improvement in efficiency relative to an analysis that uses only the information from the second-stage sample. This method also makes possible the analysis of biased sampling designs in which the sampling fraction depends on both exposure and outcome. It is thus possible to use designs such as the balanced design which are potentially more efficient than the case-control design. By making simple adjustments to the output from a standard logistic regression program, the analysis is easy to implement if the exposure variable is categorical.

This method can also be used in the analysis of data in which some important covariables are missing for a large fraction of the observations. The standard analysis used in this situation is to fit a logistic regression model using only those observations with no missing values. The parameter estimates from such an analysis will be unbiased only if the odds ratios for exposure and covariables are the same for the subsample with no missing values as they are for the entire sample. An alternative analysis would use the method presented in this paper to incorporate information from the entire sample. The necessary assumption is that the probability that an observation has missing values does not depend on the value of the covariables, although it can depend on exposure and outcome. Any researchers considering this approach should first think carefully about whether this assumption is reasonable for their data.

REFERENCES

1. White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. Am J Epidemiol 1982;115:119–28.
2. Walker AM. Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. Biometrics 1982; 38:1025–32.
3. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika 1979; 66:403–11.
4. Manski CF, McFadden D. Alternative estimators and sample designs for discrete choice analysis. In: Manski CF, McFadden D, eds. Structural analysis of discrete data with econometric applications. Cambridge, MA: The MIT Press, 1981.
5. Hsieh DA, Manski CF, McFadden D. Estimation of response probabilities from augmented retrospective observations. J Am Stat Assoc 1985;80: 651–62.
6. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. Biometrika 1988;75:11–20.
7. Fears TR, Brown CC. Logistic regression methods for retrospective case-control studies using complex sampling procedures. Biometrics 1986;42: 955–60.
8. 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;85:334–41.
9. Baker RJ, Nelder JA. GLIM release 3. Cambridge, England: Oxford University, Numerical Algorithms Group, 1978.

APPENDIX

Users of the computer package GLIM (9) may be familiar with the use of offsets in logistic regression. An offset can be thought of as a variable that is included in the model but whose coefficient $\hat{\beta}$ is not estimated but rather is forced to be equal to 1. It can be shown (6) that the correct values of $\hat{\beta}$, adjusted for the biased sampling at the second stage, can be obtained by using an offset term of $\log(n_{1j}N_{0j}) - \log(n_{0j}N_{1j})$ for observations in stratum j. If exposure is modeled via indicator variables corresponding to strata, only the first J components of $\hat{\beta}$ will be affected by this offset term. Use of an offset does not, however, give correct values for the standard deviations. These are obtained as follows.

Let x_{ijk} denote the p by 1 vector of covariates for observation ijk (the kth observation in the jth stratum in the ith disease category). Let X denote the n by p matrix whose ijkth row is x'_{ijk}. Suppose we fit a logistic regression model to the observations in the second-stage sample, making adjustments for the biased sampling by means of an offset. Let $\hat{d}_{ijk}$ denote the predicted probability that observation ijk is diseased, based on this model. Let V denote the n by n diagonal matrix, whose diagonal elements are $v_{ijk} = \hat{d}_{ijk}(1 - \hat{d}_{ijk})$. The covariance matrix from the logistic regression model, ignoring the information from the first-stage sample, is $(X'VX)^{-1}$.

For $j = 1, 2, \ldots J$, let e_j denote the n by 1 vector which has 1 in rows corresponding to observations from stratum j, and 0 elsewhere. Define the p by 1 vectors $W_j = X'Ve_j$ for $j = 1, \ldots J$, and $W = \sum_{j=1}^{J} W_j$. Define the p by p matrix

$$C = \sum_{i=0}^{1} \sum_{j=1}^{J} (1/n_{ij} - 1/N_{ij}) W_j W_j' + (1/N_0 + 1/N_1) WW'.$$

Breslow and Cain (6) show that an estimate of the adjusted covariance matrix which takes account of information from the first-stage sample is

$$(X'VX)^{-1}(X'VX - C)(X'VX)^{-1} = (X'VX)^{-1} - C^*,$$

where $C^* = (X'VX)^{-1}C(X'VX)^{-1}$. This covariance matrix estimator is not always positive-definite. When this occurs, an alternative estimator based on the within-stratum sample covariance matrices of the scores can be used (6).

It can be shown that if exposure is included in the model as $J - 1$ indicator variables, corresponding to the exposure strata, the matrix C^* is zero except for the J by J elements in the upper left corner (see proposition 3 of Breslow and Cain (6)). Thus, only the variances and covariances corresponding to the constant term and the exposure indicator variables need to be adjusted; those corresponding to the covariables need no adjustment. This makes sense intuitively, since all of the information concerning the covariables is contained in the second-stage sample, and the first-stage sample only contains additional information about the exposure variables.