American Journal of Epidemiology Advance Access published November 3, 2005

American Journal of Epidemiology
Copyright © 2005 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved; printed in U.S.A.

Vol. 162, No. 12
DOI: 10.1093/aje/kwi340

Original Contribution

Two-Stage Case-Control Studies: Precision of Parameter Estimates and Considerations in Selecting Sample Size

James A. Hanley[1,2], Ilona Csizmadi[3], and Jean-Paul Collet[1,4]

[1] Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada.
[2] Division of Clinical Epidemiology, Royal Victoria Hospital, Montreal, Quebec, Canada.
[3] Population Health and Information, Alberta Cancer Board, Calgary, Alberta, Canada.
[4] Centre for Clinical Epidemiology and Community Studies, Jewish General Hospital, Montreal, Quebec, Canada.

Received for publication December 21, 2004; accepted for publication July 7, 2005.

A two-stage case-control design, in which exposure and outcome are determined for a large sample but covariates are measured on only a subsample, may be much less expensive than a one-stage design of comparable power. However, the methods available to plan the sizes of the stage 1 and stage 2 samples, or to project the precision/power provided by a given configuration, are limited to the case of a binary exposure and a single binary confounder. The authors propose a rearrangement of the components in the variance of the estimator of the log-odds ratio. This formulation makes it possible to plan sample sizes/precision by including variance inflation factors to deal with several confounding factors. A practical variance bound is derived for two-stage case-control studies, where confounding variables are binary, while an empirical investigation is used to anticipate the additional sample size requirements when these variables are quantitative. Two methods are suggested for sample size planning based on a quantitative, rather than binary, exposure.

case-control studies; confounding factors (epidemiology); efficiency; multivariate analysis; sample size; two-stage sampling; variance inflation factor

Abbreviations: HRT, hormone replacement therapy; ln, natural logarithm; MI, myocardial infarction; *or*, empirical odds ratio: estimate of the odds ratio parameter OR; V, vasectomy.

With efficient sampling, a two-stage case-control design, in which exposure and outcome are determined for a large sample but covariates (notably confounders) are measured on only a subsample, may be much less expensive than a one-stage design of comparable power (1). This design was introduced independently by Walker (2) and White (3). Subsequent statistical developments, such as those by Cain and Breslow (4), Scott and Wild (5), and Chatterjee et al. (6), have focused on a unified data analysis approach to the various two-stage designs; efficient estimators of the parameters of interest; correct calculation of their precision; and use of routinely available regression software that allows weights or offsets, or repeated fitting of regression models, with updating of these weights or offsets between iterations.

Despite its economic advantages over traditional case-control studies, only a small number of investigators have used the two-stage case-control design. Some of the resistance may stem from a distrust of its “biased sampling,” which seems to violate a fundamental principle taught in introductory epidemiology courses. Its slow adoption may also have to do with the seemingly complex analyses, inexperience with offsets and weights, and the technical level of some of the papers that describe these analyses.

A related reason may be the lack of tools to plan the size of a two-stage case-control study. Methodological papers focus on the *relative* efficiencies of various data analysis models, using simulated and already assembled data sets, and give little guidance to those planning to collect new data

Correspondence to Dr. James A. Hanley, 1020 Pine Avenue West, Montreal, Quebec, H3A 1A2, Canada (e-mail: james.hanley@mcgill.ca).

by using this design. End users need to be able to calculate statistical precision/power in *absolute* terms. The tools currently available for doing so are no further advanced than they were when, in 1996 and 1998, we planned a series of two-stage case-control studies (7–9). For the first, we developed a method, subsequently published (10), that accommodates a binary exposure and a single binary confounder. For the second, we extended the calculations to allow for multiple confounding variables and/or covariates but were still limited to a binary exposure. Although some etiologic studies involve a natural all-or-none exposure (11, 12), most examine the amount of exposure. Thus, while our prestudy projections had to treat exposure as a binary variable, in our actual statistical analyses it was represented by either a set of indicator variates for exposure categories or a quantitative variate for tests of trend.

In this article, we extend the planning tools to accommodate multiple confounding variables and/or covariates and either a binary or a categorical exposure. We also indicate how to proceed when exposure is represented as a quantitative variate. Although sample size considerations are based on *projected* variances, the components of these variances are best understood in the context of an *existing* data set. Thus, we begin by describing a simple data analysis situation (binary exposure, one binary confounder) and calculate the variance of the natural logarithm (ln) odds ratio (OR) "by hand." We rearrange the variance formula to make it more useful for planning purposes. From our numerical investigations, and by taking advantage of the balanced structure of the stage 2 sample, we develop simple bounds for the variance in the case of multiple binary covariates and less formal ones for quantitative covariates. We conclude with some ways to approach sample sizes for a quantitative exposure factor.

```
DATA...                                                   ESTIMATE
Stage                        MI=1    MI=0                 or1 =
                                                          23x1192/238x120
1          V=1 (1.15)  23               238 (14.87)       = 0.96 ;
   2                        20    16
   2                                                      Var[ln or1]
1          V=0 ( 7.5) 120  16     20  1192 (59.6)          = 0.0569

                   Smoke = 0          Smoke = 1
                  MI=1  MI=0         MI=1    MI=0
           V=1     9     11           11      5
   2
           V=0     5     12           11      8

CORRECTIONS...
           logit[Prob[MI | S V offset ]
            = - 2.8904                                    or =
              + 1.0917 × S                                exp[0.0880]
              + 0.0880 × V
              + 1      × offset                           = 1.09
           fitted...              fitted...
              8.8  11.2             11.2   4.8
              5.2  11.8             10.8   8.2
           v = Sum{1/fitted}
             = 0.4798               v = 0.5124
           Var[ln or] | logistic ]
           = 1/( 1/0.4798 + 1/0.5124 )
           = 0.2478
           Var[ln or] | corrected]
           =0.2478 - Sum{1/n} + Sum{1/N}                  Var[ln or]
           =0.2478 -  0.2250  + 0.0569                     = 0.0797
```

FIGURE 1. Stage 1 and stage 2 data sets created from the Walker et al. (11) study of vasectomy (*V*) and myocardial infarction (MI). Shown is the point estimate of the empirical odds ratio (*or*), adjusted for confounding by smoking (*S*). Estimated variance of its natural logarithm (ln), obtained by using the Cain and Breslow (4) method. In the upper portion, frequencies from stage 1 are shown in slightly larger type, and those from stage 2 in smaller type, with inverses of sampling fractions in parentheses.

ANALYSIS OF DATA FROM A TWO-STAGE STUDY: WORKED EXAMPLE

Walker et al. (2, 11) examined the role of vasectomy ($V = 0$ or 1) in the etiology of myocardial infarction (MI), using data from the already computerized records of a health maintenance organization. Figure 1 shows the (*V*, MI) frequencies in a case-control study, with a denominator ("control," MI = 0) series 10 times the size of the case series, created from this study. These frequencies yield an empirical odds ratio (*or*) of 0.96; the estimated variance of its ln is 0.0569. Walker et al. were concerned that smoking, which is positively associated with MI, might be negatively associated with vasectomy. If it were, the 0.96 value would be too low, and adjustment for smoking might move it considerably above 1. Since it was too expensive to obtain smoking histories (via written records or direct interview) for all 1,573 men, these histories were obtained for only a sample of 72 of them. This second-stage sample was not a simple random sample; instead, to minimize the variance of the estimate of the ln odds ratio, the numbers sampled from each of the four (*V*, MI) categories were chosen to be roughly equal.

The separate (*V*, MI) frequencies for the 37 nonsmokers and 35 smokers are shown next in figure 1. As expected, among the nonvasectomized, a substantially greater proportion of those who had a history of smoking were found among those who had suffered an MI than among those who had not (11/16 vs. 8/20); furthermore, among those who had not suffered an MI, the proportion of nonsmokers was just slightly higher among the vasectomized than the vasectomized (11/16 vs. 12/20). After adjustment for the slight negative confounding by smoking, the odds ratio estimate for the *V*–non-*V* contrast is 1.09 (figure 1).

The variance to accompany the ln of 1.09 is calculated from three separate items: 1) the variance reported by the logistic regression applied to the stage 2 data, 2) the four cell frequencies in the 2 × 2 table of stage 1 data (we refer to these as the "4 *N*'s"), and 3) the corresponding sample sizes in the stage 2 data (the "4 *n*'s"). With this notation, the original expression of the variance for the ln *or* is as follows (3, 4):

$$\mathrm{Var}[\ln or] = \mathrm{Var}_{\mathrm{logistic}}[\ln or] - [\mathrm{Sum}\{1/n\} - \mathrm{Sum}\{1/N\}]. \tag{1}$$

In the above example, $\mathrm{Var}_{\mathrm{logistic}}[\ln or] = 0.2478$. If what is in effect Woolf's formula is applied to the stage 2 frequencies, $\mathrm{Sum}\{1/n\} = 1/20 + 1/16 + 1/16 + 1/20 = 0.2250$.

The stage 1 frequencies lead to Sum$\{1/N\}$ = 1/23 + 1/238 + 1/120 + 1/1,192 = 0.0569. Substituting these three items into equation 1 yields

$$\mathrm{Var}[\ln or] = 0.2478 - [0.2250 - 0.0569] = 0.0797.$$

In the next section, we take advantage of three sets of arithmetic facts. First, the 0.0797 could also have been calculated as 0.0569 + [0.2478 − 0.2250] = 0.0569 + 0.0228. Second, and most important, the 0.0228 is a difference of two variances, obtained from two different logistic regressions fitted to the stage 2 data: the 0.2478 when smoking was included in the model, the 0.2250 when it was not; and the 0.0228 is 10 percent of the 0.2250. Third, the 0.0569 is the variance associated with the ln of the crude *or* calculated from the stage 1 data.

THE VARIANCE FORMULA: REARRANGED FOR PLANNING PURPOSES

Rather than use equation 1 as it is given in the original articles (3, 4), our worked example shows that, for planning purposes, it can be written more profitably in an alternative form:

$$\begin{aligned}&\mathrm{Var}[\ln or]\\&= \mathrm{Sum}\{1/N\}\\&\quad + (\text{amount by which } \mathrm{Var}_{\mathrm{logistic}}[\ln or] \text{ exceeds } \mathrm{Sum}\{1/n\})\\&= \text{Stage 1 variance} + \text{some percentage of } \mathrm{Sum}\{1/n\}. \qquad (2)\end{aligned}$$

The advantages of this rearranged formula are threefold. First, the stage 1 variance is familiar to those who plan traditional case-control studies, and the design factors and population parameters that determine its magnitude are well understood. Second, the quantity Sum$\{1/n\}$ is easily calculated for any proposed set of stage 2 sample sizes and is also readily recognized as both the Woolf- and the logistic-based variance of the crude ln *or* in the stage 2 data set. Third, the literature already provides some guidance on the factors that influence the extent to which Sum$\{1/n\}$ is increased when additional variables are included in a logistic model. For example, table 2 of Smith and Day (13) and table 7.10 of Breslow and Day (14) deal with the analysis of a "one-stage" case-control study, with equal numbers of cases and controls. These tables give the ratio of the required sample size if the analysis incorporates (via stratification) a binary confounding variable, relative to that required if stratification is ignored. Using the same broad approach, and taking advantage of the representation in equation 2, the next two main sections of the text derive results specific to two-stage studies. First, however, to provide a template, we describe the two-stage case-control study, the planning of which prompted this work.

In our study of the role of hormone replacement therapy (HRT) in the prevention of colon cancer, we expected to have detailed, computerized stage 1 information on HRT prescriptions for a case series of 650 women diagnosed with colon cancer. We planned to obtain similar data in a denominator ("control") series of 2,600. We were concerned that covariates, not in the databases and available only through interview, could confound the comparison. From our guesstimates of trends in HRT use (subsequently documented by Csizmadi et al. (15)), we calculated that, if 15 percent of controls had long-term HRT exposure, then the 650 women would include almost 100 so exposed. Thus, in "stage 2," we planned to interview as many of the 100 as possible, together with a random sample of 150 of the 550 "less- or unexposed" cases, 150 of the expected 390 highly exposed controls, and 150 of the 2,210 less- or unexposed controls. These numbers, as close as possible to "balanced," were chosen to optimize precision.

Our projections of power were based on null and nonnull versions of equation 2. The stage 1 variance was calculated from the anticipated frequencies in the 2 × 2 table (Sum$\{N\}$ = 3,250). We calculated what the variance of the long-term HRT regression coefficient (for now, long-term HRT was taken to be a binary variable) would be if we omitted the covariates and fit the reduced model to the Sum$\{n\}$ = 550 stage 2 observations

$$\mathtt{logit}[\mathtt{Prob}[\mathtt{Colon\,Cancer} \mid \mathtt{offset\,HRT}]].$$

This variance, not a function of the offsets, is simply {1/100 + 1/150 + 1/150 + 1/150} = 0.03. Thus, it remained to anticipate by how much the variance obtained under the larger model

$$\mathtt{logit}[\mathtt{Prob}[\mathtt{Colon\,Cancer} \mid \mathtt{offset\,HRT\,covariates}]]$$

would exceed the 0.03 obtained from the reduced one.

ONE BINARY CONFOUNDER

When a covariate is added to a logistic regression model, the estimated variance of the regression coefficient [ln *or*] of interest is increased. In the case of a binary covariate C and a binary "exposure" variable E, the increase is a function of six factors: the prevalences of C and E, how strongly each one is associated with the outcome, how correlated they are with each other, and how common the outcome is. In a case-control study with incidence density sampling, this last factor is fixed by the investigator. Moreover, as evident in appendix 1, provided that rates are multiplicative in E and C, the "stage 2" variance component of equation 2 does not depend on how strongly E is associated with the outcome. Of the remaining four factors, we show the two most important in figure 1 and deal with the remaining two (the prevalences of E and C) by calculating the maximum value of [Var$_{\mathrm{logistic}}$[ln *or*] − Sum$\{1/n\}$]/Sum$\{1/n\}$, expressed as a percentage, over a wide range of possible prevalence configurations. These maxima are shown in figure 2.

As expected, the percentage by which Var$_{\mathrm{logistic}}$[ln *or*] in the stage 2 regression exceeds Sum$\{1/n\}$ is a strong function of the two features that make C a confounder, namely, the degree to which it is associated with the outcome and is correlated with E. The percentages are slightly lower than those obtained by subtracting 100 from each of the entries for "p = 0.5" in table 2 of Smith and Day (13) and table 7.10 of Breslow and Day (14). For example, for $\mathrm{OR}_C = 5$ and $\mathrm{OR}_{CE} = 5.4$, our figure 2 predicts that the second term in equation 2 equals 43 percent of Sum$\{1/n\}$, while their tables, with $p = 0.5$ and $\mathrm{OR}_E = 2$, predict 49 percent. That they

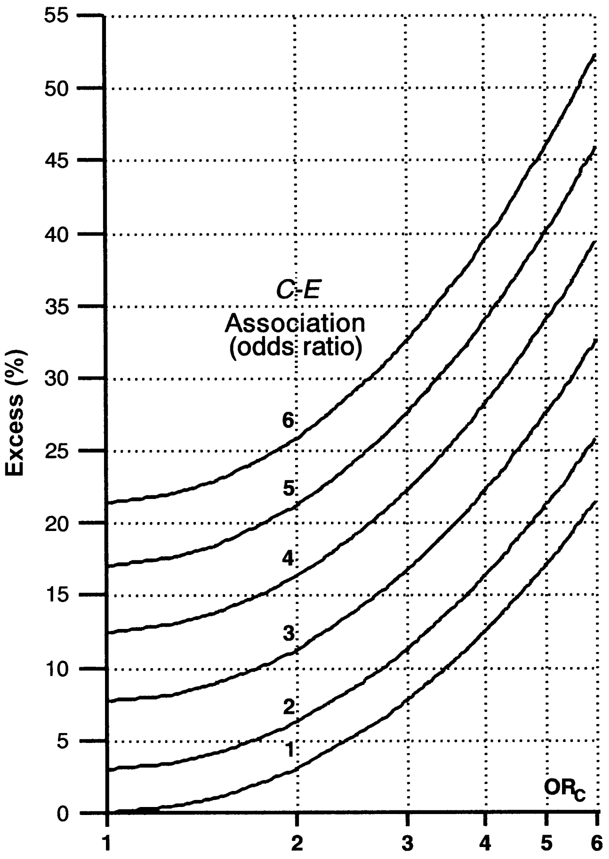

FIGURE 2. Percentage by which the variance of the natural logarithm of the odds ratio, obtained from a logistic regression model (with one binary exposure variable *E*, one binary covariable *C*, and offsets) fitted to stage 2 data, exceeds the sum of the reciprocals of the stage 2 sample sizes. The strength of *C* is given by the odds ratio OR$_C$ (horizontal axis), and the degree of association of *C* with *E* is given by the *C–E* odds ratio. Each plotted value is the maximum excess within the range $0.05 \leq \text{Prob}[E+]$, $\text{Prob}[C+] \leq 0.95$.

should differ is not surprising because the two settings, and specific calculation methods, are somewhat different. The cited tables refer to a traditional one-stage case-control study, while the values in figure 2 are derived from a two-stage case-control study, where the prevalence of *E* in the stage 2 data set is *designed* to be as close as possible to 0.5. The tabulated increases show a small dependence on OR$_E$, while our percentages are independent of this parameter. The two sets of calculations do have several features in common: the table is based on equal numbers of cases and controls, while the control:case ratio in the second-stage data set is designed to be close to 1:1; in both settings, the OR$_C$ and OR$_{CE}$ parameters refer to those in the source population, and Woolf's method is used to calculate the variance.

A general rule of thumb

For a potential confounding variable only weakly associated with the outcome and the exposure, that is, when $\text{OR}_C < 2$ and $\text{OR}_{CE} < 2$, figure 2 suggests that the second component in equation 2 would be less than 7 percent of $\text{Sum}\{1/n\}$. The upper bound would be approximately 12 percent if one of the odds ratios was 3 and the other was 2. This symmetry in the impact of OR$_C$ and OR$_{CE}$ leads us to round up each percentage to the next 5 percent and to arrive at simple upper bounds for the stage 2 component of the variance:

$\text{OR}_C + \text{OR}_{CE}$:	4	5	6	7	8	9	10
Upper bound:	10%	15%	20%	25%	30%	35%	40%

Doing so brings these percentages closer to those in Smith and Day's table and their observation that, for a weak-confounder scenario, it is only "necessary to increase the sample size by about 15%" (13, p. 358). From the pattern above, we can, for OR$_C$ and OR$_{CE}$ both being less than 5, give a general upper bound

$$\text{Var}[\ln or] \leq \text{Sum}\{1/N\} + 0.05 \times [\{\text{OR}_C - 1\} + \{\text{OR}_{CE} - 1\}] \times Sum\{1/n\}. \tag{3}$$

Applications

It is first of interest to check this inequality by an "after-the-fact" application to the data in the worked example considered in figure 1, where $\text{Sum}\{1/n\} = 0.2250$ and the variance from the logistic regression that included $C =$ smoking was 0.2478, an increase of 0.0228 or 10 percent over that from the model in which *C* was omitted. The estimate, exp[1.0917], of OR$_C$ is approximately 3, whereas the OR$_{CE}$ value, calculated from the controls, was $(5 \times 12)/(8 \times 11) = 0.68$. Since the influence of an OR$_{CE}$ of less than 1 is the same as that of one of magnitude 1/OR$_{CE}$, we substitute 1/OR$_{CE}$ in equation 3 to obtain

$$5\% \times [\{3 - 1\} + \{(1/0.68) - 1\}],$$

$$\text{that is, an upper bound of } 12\% \text{ of Sum}\{1/n\}.$$

The values in figure 2 were calculated by assuming a balanced design (note that the values in parentheses in each row of the appendix 1 figure add to unity). In the worked example, the 4 *n*'s were 20, 16, 16, and 20. However, this imbalance is already reflected in the $\text{Sum}\{1/n\}$, so small deviations from a perfect balance should not have a serious impact on the percentage.

When planning our own study, we were unable to postulate any strong confounder of the HRT–colon cancer association. Therefore, we relied on Smith and Day's advice (13), so—to account for a single binary covariate—we projected that the second component in equation 2 would be less than 15 percent of the Woolf variance obtained if one omitted such a confounder from the stage 2 model.

SEVERAL CONFOUNDERS

Confounders represented as binary variates

Although it is more difficult to project the additional variance when there are multiple confounders, the multivariable extensions of equations 2 and 3 are a useful point of

TABLE 1. Statistical power (%) for a two-stage case-control study with 650 cases and 2,600 controls providing stage 1 data and xx/150/150/150* of these persons providing stage 2 data†

Prevalence (%) of long-term exposure	Rate ratio					
	0.50	0.55	0.60	0.65	0.70	0.75
25	>99	99	96	89	77	60
20	99	98	94	85	72	55
15	98	95	89	78	64	47
10	94	87	78	66	52	38
5	76	67	56	45	34	25

* Denotes a plan to interview xx cases with long-term exposure, 150 less- or unexposed cases, 150 controls with long-term exposure, and 150 less- or unexposed controls at stage 2. xx will be a function of the prevalence of long-term exposure in the 650 cases, and it is assumed that 50% of these cases with long-term exposure will be interviewed.

† Power is given as a function of the prevalence of long-term exposure and the true rate ratio.

departure. We focus first on confounders represented by *binary* variates. Since it is difficult to study several, we focus, heuristically, on two, C_1 and C_2. The way in which we created these two variables and calculated the "percentage excess" is described in appendix 2. The fitted equation linking the "percentage excess" to the influential parameters suggests that we can omit the small effect of OR_{C1-C2} and round the other coefficients up slightly to obtain, for two confounders, a relation similar to that in equation 3:

$$\begin{aligned}\mathrm{Var}[\ln or] \le \mathrm{Sum}\{1/N\} + 0.05 \times [\{\mathrm{OR}_{C1} - 1\} \\ + \{\mathrm{OR}_{C2} - 1\} + \{\mathrm{OR}_{E-C1} - 1\} \\ + \{\mathrm{OR}_{E-C1} - 1\}] \times \mathrm{Sum}\{1/n\}. \qquad (4)\end{aligned}$$

The common form, and the same "5% per unit of excess OR" rule for equations 3 and 4, suggests that it can also be used in the case of $k > 2$ covariates. Since investigators will be unable to precisely anticipate what the parameters' values will be, they will probably base their plans on their sense of how strong the overall confounding is likely to be.

Application

We did not expect strong confounders of the HRT–colon cancer association, so we calculated the projected power based on the variance from the first stage, plus 20 percent of that from the second stage, that is, "Var[N]" + 0.20 "Var[n]". Doing so allows for 1) a single confounder with $\mathrm{OR}_C \le 3$ and $\mathrm{OR}_{C\text{–}E} \le 3$ or 2) two confounders, each with $\mathrm{OR}_C \le 2$ and $\mathrm{OR}_{C\text{–}E} \le 2$. Table 1 gives the range of statistical power for a number of scenarios, and table 2 shows the sequence of hand calculations for a specific scenario.

Confounders measured quantitatively

We first considered the literature on planning the size of a single-stage case-control study with, say, k such covariates to be adjusted for by logistic regression. Smith and Day (13) suggested that, when the correlation of E with each C equals r, and the correlation of each C with each other C equals r^2, the excess variance can be expressed as $k \times \{r^2/(1 - r^2)\}$.

TABLE 2. Illustration of power calculation for the prevalence = 20%, rate ratio = 0.70 entry in table 1

Stage(s)	Null		Nonnull	
	Cases (no.)	Controls (no.)	Cases (no.)	Controls (no.)
1				
Long-term HRT*	130	520	97	520
Less- or unexposed	520	2,080	553	2,080
V^*_1: Variance[ln* *or**]†	0.0120		0.0145	
2				
Long-term HRT	65‡	150	48	150
Less- or unexposed	150	150	150	150
V_2: Variance[ln *or*]†	0.0354		0.0407	
1 and 2				
$V = V_1$ + (20% of V_2)	0.0191		0.0227	
$-1.96 \times$ square root of $[V_{\mathrm{null}}]$ − ln 0.7			(a): 0.0858	
$[V_{\mathrm{nonnull}}]^{1/2}$			(b): 0.1528	
Z^*_β			(a) ÷ (b): 0.57	
Power: Prob[$Z < Z_\beta$]			0.72	

* HRT, hormone replacement therapy; *V*, variance; ln, natural logarithm; *or*, an estimate (empirical) of the odds ratio parameter OR; *Z*, deviate in a Normal distribution.

† V_1 and V_2 were calculated by using Woolf's formula.

‡ Assuming half of the exposed cases would be interviewed.

TABLE 3. Variance inflation factors in stage 2 logistic regression: four studies with varying degrees of confounding

Outcome/contrast and details	$or*_1$	or_{adj}*,†	V_{2L}*	V_{2L} – Sum{1/n} ["Excess"]	Excess as % of Sum{1/n}
1. Nonfatal MI*/vasectomized vs. not‡	0.96				
N: 23/238/120/1,192; Sum{1/N} = 0.0569					
n: 20/16/16/20; Sum{1/n} = 0.2250					
Covariate: smoking		1.09	0.2478	0.0228	10
2. Autism/MMR* vaccinated vs. not§	1.53				
N: 936/4,242/1,780/3,605; Sum{1/N} = 0.0260					
n: 200/200/200/200; Sum{1/n} = 0.0404					
Covariates: age, age squared, year of birth		1.06	0.0260	0.0094	23
3. Lung cancer death/female vs. male smokers¶	0.45				
N: 936/4,242/1,780/3,605; Sum{1/N} = 0.0021					
n: 200/200/200/200; Sum{1/n} = 0.0200					
Covariates: age, education, smoking intensity and duration		0.76	0.0021	0.0091	46
4. Coronary heart disease/males vs. females#	2.39				
N: 176/99/1,177/1,584; Sum{1/N} = 0.0173					
n: 150/99/150/150; Sum{1/n} = 0.0301					
Covariates: smoking (three categories), age, body mass index, serum cholesterol, systolic and diastolic blood pressures		2.24	0.0173	0.0141	48

* *or*, an estimate (empirical) of the odds ratio parameter OR; adj, adjusted; V_{2L}, Var[natural logarithm (ln) or_{adj}] calculated from multivariable logistic regression, with offsets, of second-stage data (V, vasectomy); MI, myocardial infarction; MMR, measles-mumps-rubella.

† Although not shown, the correct variance for ln or_{adj} is Sum{1/N} + [V_{2L} – Sum{1/n}].

‡ Study analyzed in table 1.

§ Data set created by the authors to closely match findings in Madsen et al. (18).

¶ Created from data in Gillespie et al. (19) from 6 years of follow-up in the American Cancer Society's Cancer Prevention Study II. Analysis here was restricted to those classified as "current smokers" at study entry.

# New cases in the first 10 years of the Framingham study (Massachusetts).

They recommend using this variance inflation factor only if the covariates are relatively weak, for example, when "considering the effect on sample size in a case-control study of breast cancer in which adjustment will be necessary for, say, age at first birth, age at menarche, parity and socioeconomic status" (13, p. 358). Their "variance inflation factor" is derived from regression models for a quantitative dependent variable and the usual identity link and normal (and homoskedastic) error, thus ignoring the fact that, in logistic regression, the variance is itself a function of the mean. Moreover, our investigations indicated that, depending on the value of prob[E+], the correlations in the stage 2 data can sometimes be larger than those in the source population. For these two reasons, we are unable to extend this to a general rule of thumb for the variance in stage 2 studies in which covariates are mostly quantitative.

In the absence of general expression for bounds on the variance inflation factor, we examined empirically the "cost of adjustment" by using examples with varying degrees of confounding. The results are shown in table 3. In example 1, discussed above, the "price" of adjusting for the one covariate, smoking, was relatively minor despite its large influence on incidence rates, because its association with the factor of interest was weak. Example 2 is of particular interest because, although it appears that there is one variable, age, both linear and quadratic age terms are required to properly describe the onset (diagnosis) rate as a function of age. Adjustment for the quite different child-years distribution in the vaccinated and unvaccinated reduced the excess risk of 53 percent $(\mathrm{OR}_{\text{crude}} - 1\}$ to just 6 percent $\{\mathrm{OR}_{\text{adjusted}} - 1\}$, indicating considerable confounding. This confounding is reflected in the variance inflation of 23 percent, comparable to that produced by a single binary variable C where, say, $\mathrm{OR}_C = 4$ and $\mathrm{OR}_{C\text{–}E} = 3$. Example 3 focuses on the hypothesis that women are more susceptible than men to tobacco carcinogens, a hypothesis that has also been examined with other designs and more recent data (16). It too shows considerable confounding, involving several factors. Not surprisingly, the ratio of the variances from logistic regressions that included and excluded these factors was 1.46; that is, the stage 2 variance was 46 percent of Sum$\{1/n\}$. In example 4, the adjusted *or* was only slightly lower than the unadjusted one. Nevertheless, inclusion of several important covariates added substantially to the second-stage variance.

From these investigations, we suggest that, unless confounding is extreme, an amalgam of 50 percent of the "Woolf" variance calculated from the stage 2 frequencies, together with 100 percent of the Woolf variance based on stage 1 frequencies, provides a useful upper bound for many multivariable analyses of two-stage case-control studies.

QUANTITATIVELY MEASURED EXPOSURE

In many studies, exposure is recorded quantitatively (e.g., *duration* of HRT in our study of colon cancer). Before proceeding to stage 2, one must divide the exposure scale into $(K + 1) > 2$ categories. The closer to equal (balanced) the sizes of the separate second-stage samples are from each of the two {case,control} × $(K + 1)$ cells, the smaller the variance of the estimated coefficient(s). In the second-stage analysis, odds ratios for the K index categories can be estimated by including K indicator variables in the logistic regression and exponentiating the estimates of the K regression coefficients $\gamma_1, \gamma_2, \ldots, \gamma_K$. The precision of each estimated γ can be anticipated from a form of equation 1, where the four N's and four n's are now the first- and second-stage frequencies in the reference category and in the index category in question. However, the K estimates are (positively) correlated, since each one represents a contrast with a common reference category. Their covariances and correlations can be calculated by using the expression given in Cain and Breslow (4, p. 1200).

In practice, at the time of the analysis, one might instead—for greater statistical efficiency and parsimony—represent the exposure as a single *quantitative* variable with coefficient β (the offsets must be included irrespective of the representation of the exposure). Unfortunately, with such an analysis, the variance of $\hat{\beta}$ is no longer expressible in the same way as in equation 1, so it is more difficult to anticipate its magnitude. As a first approximation, one might anticipate the average value of E in each category and the proportions of the source population (controls) that would be classified into these categories. For example, if exposure duration categories had midpoints $0, e_1, e_2, \ldots$, one could treat the fitted $\hat{\gamma}_j$ for category j as an estimator of $\beta \times e_j$. Thus, one could use a weighted average of the (correlated) estimates $\{\hat{\gamma}_1/e_1, \hat{\gamma}_2/e_2 \ldots \hat{\gamma}_k/e_K\}$ as an estimator of β.

This approach worked well in a two-stage case-control study we created from the same Framingham data as those used in table 3, but now with the focus on the coefficient of the quantitative variate representing the reported average number of cigarettes smoked per day. We selected the second-stage sample by using (and, in the logistic regression, included offsets for) the three categories (0, 1–15, and $\geq$25) but used the *quantitative* representation in the regression equation. The coefficient of this variable was $\hat{\beta} = 0.0214$ (standard error, 0.0062). Had we represented smoking by two indicator variables, their fitted coefficients would have been 0.2563 (variance = 0.02999) and 0.8159 (variance = 0.0567). Dividing these estimates by the mid-values of 14.7 and 36.1 cigarettes, respectively, yields two estimates of β: 0.0174 (variance $= 0.0299/14.7^2 = 0.000138$) and 0.0226 (variance $= 0.0567/36.1^2 = 0.000044$). A precision-weighted weighed average of these two (positively correlated, $r = 0.48$) estimates yields a single estimate $\hat{\beta} = 0.0223$ (standard error, 0.0065), very close to the ones obtained directly. The variances of the 0.2563 and 0.8159 could have been projected by using equation 2 and their covariance by using the expression in Cain and Breslow (4).

An alternative is to follow the suggestion of Vaeth and Skovlund (17). They approximate the power against a non-null value β_{ALT} for the coefficient of a quantitatively represented exposure by the power achievable with a specially constructed binary exposure variable. To do so, they imagine two groups, one situated 1 standard deviation below and the other 1 standard deviation above the mean exposure level. The log of the odds ratio arising from the contrast of these two groups is $\Delta = \beta_{\mathrm{ALT}} \times 2$ standard deviations.

However, given the many uncertainties in anticipating the exact distribution of the exposures and the complexities in calculating the variances, it may be more practical to plan the sample size by using a *binary* version of E and to keep the gains from using the quantitative version of E as insurance against overly optimistic projections.

DISCUSSION

In this paper, we restricted our attention to one special case of the two-stage design, namely, a case-control study in which exposure information is readily available on cases and controls but information on covariates (notably confounders) is obtained on only a subsample. We did not consider other applications, such as those to investigations of interaction and to studies involving surrogate exposures.

We focused on an *exposure* represented as a single *binary* variable, whose associated coefficient is the ln odds ratio of interest. For this situation, we showed how one can project the magnitude of the variance for its estimator and determine the expected statistical precision and power with various sample size configurations and various amounts of confounding. These calculations can be done by 1) rearranging the variance expression, 2) using Woolf's formula with the first- and second-stage frequencies, and 3) using upper bounds for the variance inflation that occurs when covariates are added to a logistic regression. The effect of including a single binary covariate is quantified in figure 1 and the simple rule of thumb given by equation 3; the rule appears to extend naturally to multiple binary covariates, as is evident in inequality 4. For quantitative covariates, we could offer only the general suggestions gleaned from the studies in table 3. Ironically, our inability to provide a definitive approach for this type of covariate has as much to do with logistic regression per se as it does with two-stage design itself.

We also dealt, but in less detail, with an *exposure* measured on a quantitative scale. Analysts often categorize such a variable and represent it by indicator variables. As shown by Cain and Breslow (4, p. 1200), if they do so by using the same exposure categories for the analyses as were used in the second-stage sampling and include the obligatory offsets, the correct variance for each ln *or* can be computed from the one provided by the logistic regression by using a simple closed expression similar to equation 1. Thus, the sample size calculations shown in figures 1 and 2 and expression 3 are easily adapted. When linearity is justified, analysts will want to summarize the effect of the exposure by using a single linear coefficient. To project the power for this analysis, we suggest two possible methods: either the method of Vaeth and Skorlund (17) or the use of what is in effect a meta-analysis of the separate estimates obtained from the categorical approach.

Our aim was to bring out the broad principles, so that those who consider using the two-stage design will have a first approximation of the precision/power achievable with various sample size configurations of stage 1 and stage 2 sample sizes. As with any sample size/precision calculations, even for simpler and better understood designs and analyses, they are merely projections, using several approximations and uncertain inputs. They should be treated accordingly.

ACKNOWLEDGMENTS

This work was supported by an operating grant to J. H. from the Natural Sciences and Engineering Research Council of Canada; fellowships to I. C. from Fonds de la recherche en santé du Québec, the Canadian Institutes of Health Research, and the Alberta Cancer Board; and an R01 grant to all three authors from the US National Cancer Institute (CA78698).

Conflict of interest: none declared.

REFERENCES

1. Breslow NE. Two-phase case-control studies. In: Armitage P, Colton T, eds. Encyclopedia of biostatistics. Vol 1. Chichester, United Kingdom: John Wiley & Sons, 1998:532–40.
2. Walker AM. Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. Biometrics 1982;38:1025–32.
3. 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.
4. Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. Am J Epidemiol 1988;128: 1198–206.
5. Scott AH, Wild CJ. Fitting regression models to case-control data by maximum likelihood. Biometrika 1997;84:57–71.
6. Chatterjee N, Chen YH, Breslow NE. A pseudoscore estimator for regression problems with two-stage sampling. J Am Stat Assoc 2003;98:158–68.
7. Sharpe CR, Collet JP, Belzile E, et al. The effects of tricyclic antidepressants on breast cancer risk. Br J Cancer 2002;86: 92–7.
8. Sharpe CR, Collet JP, McNutt M, et al. Nested case-control study of the effects of non-steroidal anti-inflammatory drugs on breast cancer risk and stage. Br J Cancer 2000;83:112–20.
9. Csizmadi I, Collet JP, Benedetti A, et al. The effects of transdermal and oral oestrogen replacement therapy on colorectal cancer risk in postmenopausal women. Br J Cancer 2004;90: 76–81.
10. Schaubel D, Hanley J, Collet JP, et al. Two-stage sampling for etiologic studies: sample size and power. Am J Epidemiol 1997;146:450–8.
11. Walker AM, Jick H, Hunter JR, et al. Vasectomy and non-fatal myocardial infarction. Lancet 1981;1:13–15.
12. Engels EA, Chen J, Viscidi RP, et al. Poliovirus vaccination during pregnancy, maternal seroconversion to simian virus 40, and risk of childhood cancer. Am J Epidemiol 2004;160: 306–16.
13. Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. Int J Epidemiol 1984;13:356–65.
14. Breslow NE, Day NE, eds. Statistical methods in cancer research. Vol 2. The design and analysis of cohort studies. Lyon, France: International Agency for Research on Cancer, 1987. (IARC scientific publication no. 82).
15. Csizmadi I, Benedetti A, Boivin JF, et al. Use of postmenopausal estrogen replacement therapy from 1981 to 1997. CMAJ 2002;166:187–8.
16. Henschke CI, Miettinen OS. Women's susceptibility to tobacco carcinogens. Lung Cancer 2004;43:1–5.
17. Vaeth M, Skovlund E. A simple approach to power and sample size calculations in logistic regression and Cox regression models. Stat Med 2004;23:1781–92.
18. Madsen KM, Hviid A, Vestergaard M, et al. A population-based study of measles, mumps, and rubella vaccination and autism. N Engl J Med 2002;347:1477–82.
19. Gillespie BW, Halpern MT, Warner KE. Patterns of lung cancer risk in ex-smokers. In: Lange N, Billard L, Conquest L, et al, eds. Case studies in biometry. Somerset, NJ: Wiley-Interscience, 1994:385–408. (Data accessible from the following website: http://www.stat.cmu.edu/).

APPENDIX 1

Basis for Figure 1

This example describes how an "excess" value was calculated, using a source population where the prevalence of $E+$ is Prob[$E+$] = 0.15 and that of $C+$ is Prob[$C+$] = 0.20, and the codistribution of confounder (C) and exposure (E) yields an odds ratio $\mathrm{OR}_{CE} = 2$. In the $E-$, the outcome is three times as common in those $C+$ as it is in those $C-$, that is, $\mathrm{OR}_C = 3$. The calculation was repeated for all combinations within the range of 0.05 × Prob[$E+$] × 0.95 and 0.05 × Prob[$C+$] × 0.95, and the *maximum* excess over this range is the value 11.2 percent shown in figure 1.

(i) *E-C* distribution in Source				**(ii)** *C* distribution in Controls*		
	C-	*C+*		*C-*	*C+*	Total
E-	0.696	0.154	0.85	0.696 (0.82)	0.154 (0.18)	0.850 (1.00)
E+	0.104	0.046	0.15	0.104 (0.69)	0.046 (0.31)	0.150 (1.00)
	0.800	0.200	1.00			

(iii) Event Rates in Source†			**(iv)** *C* distribution in Cases‡		
E-	1	3	0.696 (0.60)§	0.462 (0.40)	1.158 (1.00)
E+	4	12	0.416 (0.43)	0.552 (0.57)	0.968 (1.00)
Sum{reciprocals}			6.66¶	13.04#	4.00**

$$\frac{1}{1/6.66 + 1/13.04} \qquad 4.41^{\dagger\dagger}$$

$$100 \times \frac{4.41 - 4.00}{4.00} \qquad \text{excess} = 10\%$$

* Same as in source.

† Event rates are multiplicative and are relative rather than absolute. Since the selected value does not affect the calculations, we arbitrarily assumed that, in the *C*–, the outcome is four times as common in those *E*+ relative to the rate in those *E*–, that is, $\mathrm{OR}_E = 4$.

‡ Relative frequencies were obtained by multiplying source *E–C* frequencies by event rates, for example, 0.696 × 1, 0.154 × 3, 0.104 × 4, and 0.046 × 12.

§ Proportions (in parentheses) scaled to add to 1 within each row to reflect separate stage 2 sampling within each row and a balanced design.

¶ 1/0.82 + 1/0.69 + 1/0.60 + 1/0.43.

1/0.18 + 1/0.31 + 1/0.40 + 1/0.57.

** 1/1.00 + 1/1.00 + 1/1.00 + 1/1.00.

†† Woolf variance of ln *or*, used to approximate variance calculated by logistic regression.

APPENDIX 2

Two Binary Confounders

We created the trivariate distribution of a binary exposure, E, and two binary confounders, C_1 and C_2, by supposing that they arise from two bivariate normal distributions of $\{C_1, C_2\}$— one for those who are "$E-$" centered on $\{0,0\}$ and one for those who are "$E+$" centered on $\{\delta,\delta\}$. The value $\delta/2$ was used to dichotomize the C_1 and C_2 values to create eight "cells" in all; the frequencies in the source were obtained from the bivariate normal density functions and the marginal frequency prob[$E+$]. The expected frequencies in cases and controls were then calculated in the same way as in appendix 1, but with four possible values of the (now) two confounders. The eight source frequencies served as the frequencies for the controls. The (relative) frequencies of cases in these eight cells were obtained by multiplying the source frequencies by the corresponding event rates (taken to be multiplicative). To mimic the balanced stage 2 sample, the 16 frequencies were then scaled so they summed to unity within each of the four ($\{E+/E-\} \times \{\text{Case/Control}\}$) combinations sampled from. A data set with 16 observations, one for each of the $\{E+/E-\} \times \{\text{Case/Control}\} \times \{C_1+/C_1-\} \times \{C_2+/C_2-\}$ configurations, was then created and analyzed by using multiple logistic regression, with the associated (scaled) frequency serving as the weight for each observation and the appropriate quantity serving as an offset (inclusion of the latter has a large effect on the point estimate of ln OR_E, but no effect on its variance). The amount by which this variance exceeded the $\text{Sum}\{1/n\} = (1/1 + 1/1 + 1/1 + 1/1) = 4$ was calculated and expressed as a percentage.

This process was carried out for 324 combinations of values of prob[E+] (0.1–0.5), OR_{C1}, OR_{C2}, $\text{OR}_{E\text{–}C1}$, $\text{OR}_{E\text{–}C2}$, and $\text{OR}_{C1\text{–}C2}$ (each OR from 1 to 5) and yielded values for excess variance that varied from 0 percent to 72 percent. The percentage excess showed virtually no relation with prob[E+] or OR_E. Its dependence on the amount by which the odds ratios exceeded their null values was estimated from the linear regression model ($r^2 = 0.99$):

$$\begin{aligned}\% = {} & 4.7\{OR_{C1} - 1\} + 4.7\{OR_{C2} - 1\} \\ & + 3.9\{OR_{E-C1} - 1\} + 3.9\{OR_{E-C1} - 1\} \\ & + 0.6\{OR_{C1-C2} - 1\}.\end{aligned}$$