Increased Precision Using Countermatching in Nested Case-Control Studies

Kyle Steenland and James A. Deddens

Nested case-control studies in occupational cohorts are often used to estimate exposure effects when development of detailed exposure estimates for all cohort members is too costly. Duration of exposure, which can act as a surrogate for cumulative exposure, is often readily available for all cohort members. Langholz and others have recently proposed a method of control selection called countermatching, which uses data on the surrogate to determine which controls are selected from the risk set for a given case. This method may increase precision relative to the usual random sampling of the risk set. We compare countermatching with random sampling in a nested case-control study of silicosis among miners. Data on cumulative exposure were in fact available for all cohort members, enabling estimation of the parameter of interest in the full cohort. We conducted nested case-control analyses using 100, 20, 10, and 3 controls per case using random sampling and additional analyses using 3 controls per case with two different methods of countermatching. All analyses were replicated 50 times to explore the statistical properties of the estimated exposure parameter. We found that one of the countermatching methods markedly increased efficiency compared with random sampling. Countermatching using 3 controls per case yielded an approximate 25% increase in relative efficiency compared with random sampling; it was approximately equivalent to random sampling using 10 controls. (Epidemiology 1997;8:238–242)

Keywords: epidemiologic methods, case-control studies, study design, countermatching.

Occupational nested case-control studies are often conducted within cohorts to (1) develop detailed estimates of exposure for a limited number of cases and their controls, or (2) collect additional data on an important confounder, such as smoking, for a limited number of cases and controls. Often, the cost of obtaining this information (detailed exposure estimates or data on a confounder) for the entire cohort is prohibitive.

To obtain an unbiased (or consistent) estimate of the exposure effect in the nested case-control study, proper control selection is required.[1] The usual method of selection is to sample controls randomly within risk sets for each case, defined as the case and all study subjects who survived past the time of the case's failure and had entered the study before the case's failure. Failure time is often measured on the age scale.[2] Exposures within the risk set are time dependent, and exposures for noncases must be truncated, when necessary, at the time, or age, of the case's failure. The precision of the estimated effect is improved when more controls are chosen per case. Relative efficiency, defined as the ratio of the variance of the exposure effect in the full cohort to the variance of the exposure effect in the nested case-control study, has been shown in the null case (no exposure effect) for a dichotomous exposure variable to be $(m - 1)/m$, where m represents the number of controls chosen per case.[2] This relative efficiency decreases as the exposure effect departs from the null.

Recently, Langholz and others[3–5] have proposed another method of control sampling for nested case-control studies, called countermatching. In one scenario for countermatching, some surrogate of the exposure variable is available on the entire cohort, but exposure itself still would be too costly to obtain for the whole cohort. An example would be duration of exposure or employment as a surrogate for cumulative exposure in an occupational cohort. Duration of exposure may be available for the whole cohort, but reconstruction of exposure level (intensity) for each job over time, which is necessary to derive cumulative exposure, often is not available. In this case, countermatching may result in a gain in efficiency of the exposure estimate relative to random sampling. This is the scenario we consider here.

Countermatching is somewhat counterintuitive. For the scenario we consider, it requires that each case's risk set be stratified by the surrogate of exposure and then that controls for that risk set be selected from the strata other than the case's stratum. For example, if duration of exposure were divided into quartiles within a risk set, and three controls were to be selected for each case, then

From the National Institute for Occupational Safety and Health, Cincinnati, OH.

Address correspondence to: Kyle Steenland, National Institute for Occupational Safety and Health, 4676 Columbia Parkway, Cincinnati, OH 45226.

Submitted July 2, 1996; final version accepted October 9, 1996.

Editors' note: See related editorial on page 227 of this issue.

three controls would be randomly selected, one each from each of the quartiles except the case's quartile. The logic of such a selection can perhaps be better understood in the situation when only one control is chosen per case, and exposure is dichotomous. The result of countermatching is then to maximize the number of discordant pairs; the exposure estimate is derived from the proportion of discordant pairs in which the case is exposed compared with pairs in which the control is exposed. It is well known that, in pair-matched case-control studies, no information is obtained from concordant pairs,[1] so the idea of maximizing the number of discordant pairs would appear intuitively to increase efficiency (precision). The gain in precision will be a function of how predictive the surrogate is for the true exposure (or how correlated it is with the true exposure, for continuous variables).

Although the above example assumes a single control chosen per stratum, this restriction is not necessary, and the method is generalizable to a varying number of controls; moreover, the strata within a risk set do not have to be of the same size.

The partial likelihood used to estimate the exposure effect in countermatching is the same as the usual partial likelihood, except that the data in the numerator (the case) and denominator (the risk set) are weighted by the inverse of the probability that the corresponding subjects were selected within strata (the case is actually chosen with probability of one, but, like a control, it is assigned a weight-based size of the stratum from which it is selected).

Using the notation of Langholz and Clayton,[3] suppose a risk set for a case has n members and is stratified by a surrogate of exposure into a number of strata, each with n_i subjects ($n_1, n_2, \ldots$). From each stratum I, the investigator then chooses m_i subjects. If the case is found in stratum I, then $m_i - 1$ controls may be chosen from that stratum, and m controls may be chosen from all other strata. Each risk set contributes a term to the log partial likelihood, which has the form

$$\log(\theta_{(\text{for case})}/\sum_{\text{risk set}}\theta),$$

in which θ is an exponential function of risk factors including exposure. In countermatching, weights are assigned, and the term in the log partial likelihood has the form

$$\log(W\theta_{(\text{for case})}/\sum_{\text{case-control set}}W\theta),$$

where $W = n_i/m_i$.

Langholz and Borgan[4] have developed the above weighted partial likelihood and shown that it provides consistent estimates of the parameters and their variances and that the standard error of the estimated exposure parameter is reduced relative to random sampling of the risk set. Here, we investigate further the relative efficiency of countermatching compared with random sampling in nested case-control studies, for the situation in which the effect of exposure is measured by the regression coefficient for a continuous exposure variable (a measure of the exposure-response trend). We have used empirical data from a cohort study of gold miners in which the outcome of interest was silicosis and the variable of interest was cumulative exposure; as a surrogate for cumulative exposure, we used duration of exposure (years underground). Both of these variables were time dependent. The actual value of cumulative exposure was known for all cohort members from a previously developed job-exposure matrix, but, for illustrative purposes, we can assume that these values were unknown except for cases and controls in a nested case-control study.

Methods

The data used here have been described previously.[6] Briefly, we studied a cohort of 3,300 gold miners who had worked underground for at least 1 year between 1940 and 1965 and who had been exposed to high levels of crystalline silica. Complete work histories were available for all cohort members. Vital status was traced through 1990, and 1,551 deaths were ascertained. Sampling data for respirable dust existed back as far as 1937 (respirable dust counts were converted to gravimetric measures of silica, assuming that 13% of the dust was silica). A job-exposure matrix was created to estimate dust levels by job category across time. Levels before 1937 were estimated on the basis of 1937 levels and industrial hygienists' information. Cases of silicosis (N = 170) were identified by death certificate and via two cross-sectional surveys in 1960 and 1976. Prior Poisson regression analyses of rates of silicosis indicated a clear trend of increased risk with increased cumulative exposure to silica; risk increased monotonically until the highest category, with rate ratios of 1.9, 9.8, 22.0, 54.4, 234.8, and 216.9 by ascending cumulative exposure category *vs* the lowest category (the categories were 0 to <0.2, 0.2 to <0.5, 0.5 to <1.0, 1.0 to <2.0, 2.0 to <3.0, 3.0 to <4.0, ≥4.0 mg per m^3-years).[6] Cumulative exposure (or the log of cumulative exposure) was a better predictor of silicosis risk than either average intensity of exposure or simple duration of exposure. Cumulative exposure was defined as usual as the sum across all jobs of the product of duration in job and exposure level for that job. No lag was applied.

For our purposes, silicosis had the advantage of being an outcome that is associated with exposure alone; there is no background rate of silicosis in a nonexposed population. Hence, we chose a simple model of silicosis as a function of cumulative exposure without other covariates. In Cox regression analyses, the log of cumulative exposure fit the data better than simple cumulative exposure and also appeared to have a generally good fit based on inspection of the categorical results; therefore, we used the log of cumulative exposure as our exposure measure.

We first conducted Cox regression for the entire cohort to determine the value of the parameter that we wished to estimate in a nested case-control approach. The time variable for these Cox regression analyses was

age, and risk sets consisted of all those who survived past the index case's age. Exposures for members of the risk set were truncated when individuals reached the age of the index case's failure.

We then randomly sampled controls from the risk sets, according to the usual method for nested case-control analyses. We chose 50 sets each of random samples of 3, 10, 20, and 100 controls. The use of 3 controls is probably more typical than 10, 20, or 100 for case-control analyses in which there is a substantial cost for each control, but we used 10, 20, and 100 controls to illustrate the gain in precision obtained by using more controls with simple random sampling and to have some referent points in comparing the relative efficiency of countermatching with random sampling. We restricted our analyses with countermatching to the more common situation of 3 controls per case, again conducting 50 separate case-control analyses.

Countermatching in our analysis required stratification of each risk set by some cutpoints of the surrogate variable. We used duration as the surrogate; the correlation between duration and cumulative exposure was 0.69 in the full cohort. There are a number of possibilities for choosing these cutpoints.[5] We used two methods. For method 1, we stratified each risk set by quartiles of simple duration of exposure, with these quartiles determined from all subjects within that risk set. The cutpoints for the quartiles varied between risk sets. Three controls were chosen for each risk set from the strata other than the stratum in which the case was found. The use of quartiles within each risk set assured equal numbers in each stratum within the risk set. With the sampling of one control per stratum, this meant that all weights in the partial likelihood were equal and thus canceled out (resulting in the usual unweighted likelihood). This procedure allowed use of standard software for analyzing nested case-control studies, without any weighting of the likelihood; this was a motivation for the use of method 1. We used SAS's PHREG procedure for analysis.[7]

We also used a second method of countermatching (method 2), as suggested by Langholz and Goldstein.[5] Using this method, we chose quartiles of duration based on the distribution of duration for the 170 cases. These cutpoints were then applied to all risk sets. Because risk sets are formed on the basis of age, and age is correlated with duration, following the procedure suggested by Langholz and Goldstein,[5] we first checked to see whether these distributions of duration varied by the age of the cases by dividing cases into young and old, using the median age at failure. Finding that the quartiles for duration were approximately equivalent for young and old, we used a single set of cutpoints based on these quartiles across all risk sets. This procedure resulted in stratum sizes which differed across risk sets and therefore required the use of weights in the partial likelihood. We used SAS's PHREG procedure, which, in newer versions of SAS (version 6.10 and above), allows for an offset, by which we incorporated the weights. By way of example, suppose stratification of a risk set using the quartiles of distribution of all cases' duration of exposure led to stratum sizes of 100, 80, 60, and 40, and suppose the case was found in the last stratum. One control would be sampled from each of the first three strata. The weights for the three controls would be 100, 80, and 60, whereas the weight for the case would be 40. In SAS PHREG, the offset is set to equal to the log of the weight.

We also reconstructed the data to dampen the marked size of the exposure effect to observe differences in relative efficiency for the usual random sampling *vs* the countermatching strategy under less "extreme" conditions (conditions in which the exposure distribution of the cases was not so markedly different from that of the noncases). This reconstruction was done by re-allocating some cases within their risk set from their original quartile of cumulative exposure to a lower one (quartiles based on the distribution of cumulative exposure for all subjects in the risk set) and assigning these re-allocated cases a new cumulative exposure and new duration. These new values were equal to the midpoints of the values for quartiles of cumulative exposure and duration, with quartiles again calculated for all subjects within each risk set.

The number of cases re-allocated from higher quartile to lower quartile was somewhat arbitrary; the goal was to obtain a regression coefficient for the log of cumulative exposure about half the size of the original coefficient, while retaining approximately the shape of the original exposure-response (increasing monotonically until tailing downward at the last point in a categorical analysis). The numbers of cases falling within quartiles 1–4 of cumulative exposure, based on distribution of cumulative exposure within each risk set, were originally 3, 10, 26, and 131, respectively. After re-allocation, 5, 19, 69, and 77 cases fell into quartiles 1–4 of cumulative exposure with their risk sets, which succeeded in decreasing the exposure effect by about half.

For all case-control analyses, across the 50 analyses, we calculated (1) the average point estimate of effect, (2) the average standard error of that point estimate, and (3) the average likelihood ratio test for the exposure variable.

We compared different case-control analyses by calculating the average relative efficiency for each. The average relative efficiency was calculated by dividing the variance of the estimated parameter for log cumulative exposure from the cohort study by the average variance

TABLE 1. Results for the Full Cohort Analysis (Cox Regression)

Exposure Variable	Parameter (Standard Error)	Wald Test	Likelihood Ratio Test
Original dataset			
Log cumulative exposure	1.5557 (0.1100)	199.7	344.3
Cases assigned lower exposures to dampen exposure-response gradient			
Log cumulative exposure	0.844 (0.0778)	117.7	148.0

of the estimated parameter for log cumulative exposure parameter in the 50 case-control samples. Although the estimated exposure parameters in the case-control studies are unbiased asymptotically, sampling variation will result in some bias for any given estimate, based on case-control data. We took this into consideration by also calculating a mean squared error, defined as the square of the average bias from the full cohort estimated parameter plus the average variance of the case-control estimated parameter.

Results

Table 1 gives the parameter estimates, standard errors, and likelihood ratio tests for log cumulative exposure for the full cohort using the original data, and also using data reconstructed to dampen the exposure-response gradient.

Table 2 shows the results for case-control analyses, with the original data. Use of random sampling with 100, 20, and 10 controls yields estimates almost identical to the "true" parameter estimated from the full cohort, with increasing average standard error. With 3 controls, estimates of the parameter are more unstable. The average standard error of random sampling and countermatching method 1 are about the same, indicating no gain in relative efficiency. The average standard error of countermatching method 2 shows a marked decrease, however.

Table 3 shows the results for case-control analyses when some cases have been assigned lower exposures to lessen the steep exposure-response gradient. In this situation, the distribution of cumulative exposure for the cases is not so skewed as the original data, and with 3 controls, countermatching by method 1 shows some decrease in average standard error compared with random sampling. Countermatching by method 2 shows an even greater relative decrease in average standard error (greater gain in relative efficiency).

Table 4 shows the relative efficiency for random sampling *vs* countermatching for the case-control studies. A second statistic, mean squared error, combines the information on the efficiency of the estimated parameter with the degree of bias from the true parameter. For the original data with steep exposure-response, countermatching using method 1 provided no increase in relative efficiency but some improvement in mean squared error. Countermatching with method 2 was more effective, with a marked gain in relative efficiency compared with random sampling (from 50% to 78%), and with a corresponding decrease in mean squared error. For the derived data with a lower exposure effect, and using 3 controls, both methods of countermatching showed gains in relative efficiency and mean squared error compared with random sampling, but, again, method 2 was notably superior to method 1. For both the original data and the derived data with a lower exposure effect, countermatching using method 2 with 3 controls was approximately equivalent to random sampling with 10 controls.

TABLE 2. Case-Control Analyses for Original Data

Number of Controls Chosen from Risk Set	Average Parameter Estimate*	Average Standard Error*	Average Likelihood Ratio Test*
100	1.557	0.1121	339.1
20	1.537	0.1190	314.3
10	1.540	0.1270	293.8
3	1.456	0.1552	212.4
3 (countermatching 1)	1.526	0.1570	259.1
3 (countermatching 2)	1.609	0.1247	273.9

* For each row of the table, 50 different sets of controls were selected, and 50 analyses were conducted.

TABLE 3. Case-Control Analyses of Reconstructed Data with Lower Exposure-Response Gradient

Number of Controls Chosen from Risk Set	Average Parameter Estimate*	Average Standard Error*	Average Likelihood Ratio Test*
100	0.843	0.0783	146.0
20	0.845	0.0812	139.4
10	0.840	0.0840	131.7
3	0.876	0.1031	108.1
3 (countermatching 1)	0.842	0.0929	127.0
3 (countermatching 2)	0.865	0.0853	133.3

* For each row of the table, 50 different sets of controls were selected, and 50 analyses were conducted.

Discussion

Countermatching can be thought of as one example of a family of methods for selection of a subset of controls (and possibly cases), called "two-stage" designs. Breslow[8] has recently provided a general discussion of such "two-stage" designs, in which an original large number of cases and controls is identified at stage 1, and, to reduce costs, a subset of them is chosen for more detailed data collection at stage 2. Rather than choose the subset at random, case and controls may be sampled nonrandomly to increase the precision of the final estimator of exposure effect (for example, when the exposure is rare exposed cases and controls might be oversampled). The oversampling must then be taken into account at the analysis stage to obtain unbiased estimates of the exposure effect.

In countermatching as used here, a surrogate of cumulative exposure (that is, duration of exposure) is used for control selection within risk sets to increase variation in exposure between cases and controls, which in turn increases the precision of the estimator of exposure effect. Controls within risk sets are purposefully selected to have different durations of exposure than cases. Appropriate weighting of the likelihood is necessary to account for the departure from random sampling, to obtain unbiased estimates of exposure effect. The resulting estimates are more precise than those obtained by random sampling.

In the context of any given study, there is a variety of methods by which countermatching can be performed. Our data were obtained from an occupational cohort that we analyzed using a Cox model, with the formation of risk sets for each case, using age as the time variable. For countermatching, the risk sets were stratified by duration of exposure (our surrogate for cumulative ex-

TABLE 4. Relative Efficiency and Mean Squared Error (MSE) of Case-Control *vs* Full Cohort Estimate*

Design and Number of Controls	Relative Efficiency (%)	MSE
Original data		
Random sampling		
100	96.4	0.0126
20	85.4	0.0145
10	73.9	0.0166
3	50.0	0.0342
Countermatching 1		
3	49.0	0.0256
Countermatching 2		
3	77.7	0.0184
Reconstructed data with lower exposure effect		
Random sampling		
100	98.8	0.0061
20	91.9	0.0066
10	85.7	0.0071
3	59.1	0.0103
Countermatching 1		
3	71.6	0.0085
Countermatching 2		
3	83.1	0.0077

* Average relative efficiency = variance full cohort parameter for exposure/average variance case-control parameter for exposure. MSE = square of average bias (from full cohort parameter) + average variance of case-control parameter.

posure), and controls were then selected from strata other than the case's stratum. There are a number of ways in which the stratification could be carried out. We chose to use two methods. Method 1 stratified by quartiles of duration calculated within each risk set and was motivated by its simplicity and the fact that weights were not required in the analysis (as all weights were equal, because stratum sizes are equal within risk sets). In method 2, the cutpoints were chosen on the basis of the quartiles of distribution of duration for all cases; stratum sizes differed within risk sets, and weighting was required in the analysis. Although we know of no theoretical justification for method 2 *vs* method 1, there is an intuitive one. Because there is an exposure effect, and because duration and cumulative exposure are correlated, the cases' durations (and cumulative exposures) will tend to be clustered in the upper end of the distribution in any given risk set. Because the goal is to maximize variation in the exposures of cases *vs* controls in each risk set (achieving more "discordance"), it is best to form strata based on the cases' durations rather than durations within a risk set (note that under the null hypothesis, with no exposure effect, cases are not clustered, and method 1 and method 2 are equivalent). In fact, method 2 did prove to be superior for increasing precision, and incorporation of weights into the likelihood was not a problem using available software. Method 1 resulted in some gain in precision over random sampling in the "re-allocated" dataset in which the exposure effect was not as strong, so that clustering of cases in upper percentiles of duration and exposure distributions was not so extreme.

The gain in precision using countermatching *vs* random sampling was appreciable. In our data, the use of countermatching with 3 controls was equivalent to the use of random sampling using 10 controls. In situations where the cost of obtaining detailed exposure information for each control is important, countermatching would certainly be recommended.

Acknowledgments

John Bailar and Bryan Langholz kindly provided comments on the manuscript.

References

1. Breslow NE, Day NE. Statistical Methods in Cancer Research. vol. 2. The Design and Analysis of Cohort Studies. IARC Scientific Pub. No. 82. Lyon: International Agency for Research on Cancer, 1987.
2. Breslow NE, Lubin H, Marek P, Langholz B. Multiplicative models and cohort analysis. J Am Stat Assoc 1983;78:1–12.
3. Langholz B, Clayton D. Sampling strategies in nested case-control studies. Environ Health Perspect 1994;102(suppl 8):47–51.
4. Langholz B, Borgan O. Counter-matching: a stratified nested case-control sampling method. Biometrika 1995;82:69–79.
5. Langholz B, Goldstein L. Risk set sampling in epidemiologic cohort studies. Stat Sci 1996;11:35–53.
6. Steenland K, Brown D. Silicosis among gold-miners: exposure-response analyses and risk assessment. Am J Public Health 1995;85:1372–1377.
7. SAS Institute. SAS User's Guide: Statistics. Version 6.07. Cary, NC: SAS Institute, 1991.
8. Breslow NE. Statistics in epidemiology: the case-control study. J Am Stat Assoc 1996;91:14–28.