Sampling Strategies in Nested Case-Control Studies

Bryan Langholz[1] and David Clayton[2]

[1]Department of Preventive Medicine, University of Southern California School of Medicine, Los Angeles, California; [2]Medical Research Council Biostatistics Unit, University Forvie Site, Cambridge, UK

A stratified version of nested case-control sampling which we call "countermatching" is presented. This design uses data available for all cohort members to obtain a sample for collecting additional information in a case-control substudy. Hitherto the only stratified sampling design for such studies has involved matching of controls to cases with respect to confounding variables. However, in some situations, rather than sampling to make controls as similar as possible to cases, we might wish to make them as different as possible. This is achieved by the counter-matched design. Statistical analysis of counter-matched studies is straightforward using existing computer software. We investigate the use of the design when a surrogate measure of exposure is available for the full cohort, but accurate exposure data is to be collected only in a nested case-control study, and when exposure data are available for the whole cohort but data concerning important confounders are not. Asymptotic relative efficiency calculations indicate that a substantial efficiency gain relative to simple random sampling of controls can be expected in these situations. We also illustrate how the design might be implemented in practice. — Environ Health Perspect 102(Suppl 8):47–51 (1994)

Key words: asymptotic efficiency, cohort study, case-control study, counter-matching, design of medical study, matching, epidemiology, survival analysis, partial likelihood

Introduction

Epidemiologic cohort studies are considered the most reliable design for studying risk factors for disease incidence for two reasons. First, prospective collection of biologic marker, demographic, and environmental data avoids the information bias that may occur in retrospective data collection. Second, case-control studies are prone to selection bias attributable to flawed sampling of base populations.

Ideally, complete and accurate risk factor data would be acquired for every subject in the cohort. However, since incidences of diseases such as cancer are rare events, study cohorts usually must be very large and fewer resources usually can be devoted to collecting risk factor data for any one subject than would be possible in a retrospective case-control study. Thus the choice between the two study methods is often a choice between higher reliability with sparse data and lower reliability with more complete data.

A cost effective compromise is to identify all risk factors that would be subject to possible bias in retrospective data collection and, at the time of recruitment of a subject into a cohort study, to collect the source materials necessary to generate such data in a nested case-control study. When the processing of such source materials is costly relative to their collection, this design achieves the economy of the case-control method while avoiding the problems of selection bias and retrospective data collection which compromise the retrospective case-control method.

There are many instances in which the processing of source materials for risk factor measurements is costly. These include, for example, coding of records of diet, exercise or occupational exposure, and laboratory analysis of blood and urine samples. In all these cases, the source materials, whether paper records or biologic samples, can be banked when the subject is recruited but analyzed in detail only for a relatively small number of cases and controls.

Another situation in which nested case-control studies can be valuable is when a study cohort can be identified opportunistically from records collected for another purpose and mechanisms exist for registering disease events of interest, but where risk factor data for individual subjects are scant. These data may be augmented in cases and a sample of controls drawn from the study cohort. Although the problem of information bias due to retrospective data collection remains, selection bias is avoided in such studies.

This article discusses the design of nested case-control studies with particular emphasis on new possibilities arising out of recent advances in analytical methods.

Sampling Risk Sets

Although the use of nested case-control studies goes back at least to the 1960s (*1*) and the idea was more formally proposed by Mantel (*2*), their analysis and logical basis have been considerably clarified by the idea of partial likelihood. This was introduced by Cox (*3*) and further developed by a number of others throughout the 1970s, culminating in the work of Anderson and Gill (*4*). The main idea is illustrated in Figure 1, which represents a small study of 11 subjects. Each horizontal line represents the observation of a single subject through time. Termination of the observation line by a vertical bar represents end of study without occurrence of the disease of interest (censoring) and termination by a filled circle represents disease inci-

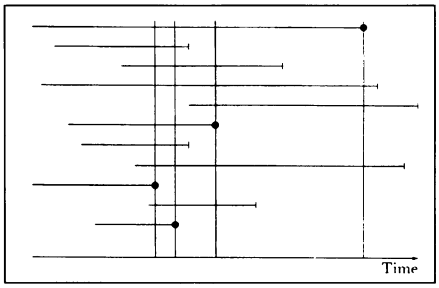

Figure 1. Definition of risk sets.

This paper was presented at The 4th Japan-US Biostatistics Conference on the Study of Human Cancer held 9–11 November 1992 in Tokyo, Japan.

Bryan Langholz acknowledges the hospitality of the members of the Medical Research Council Biostatistics Unit for providing superb working facilities during his sabbatical year. His work on this project was supported by National Cancer Institute grant CA14089.

Address correspondence to Bryan Langholz, Department of Preventive Medicine, University of Southern California, School of Medicine, 2025 Zonal Ave, Los Angeles, CA 90033-9987. Telephone (213) 342-1212. Fax (213) 342-3211.

dence. Thus, in the study shown, four cases of disease were observed. Corresponding to each case is a risk set, consisting of all subjects under study at the moment of event occurrence—in the diagram all subjects whose observation lines cross the relevant vertical. The partial likelihood method compares risk factor profiles of cases with the profiles in the corresponding risk sets.

If, for each subject, we express the probability rate of disease incidence as a multiple, θ, of a "baseline" time trajectory, each disease event contributes a term of the form

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

to the log partial likelihood. Standard maximum likelihood methods may then be used to fit models for θ as a function of risk factors.

A nested case-control study is generated by sampling the risk set—the risk set is replaced by a set containing the case and a random sample of all the remaining subjects in the set (the case-control set). The statistical analysis is otherwise unchanged, the log partial likelihood contribution for each event being

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

which is identical to that for "conditional" logistic regression analysis of individually matched case-control studies.

The loss of power for testing association when this stategy is adopted is surprisingly small. If $m-1$ controls are selected for each case, the relative efficiency, the ratio of the (large sample) variance of the estimated coefficient from the nested case-control study to that using the full cohort, is $(m-1)/m$ (*5,6*). This relative efficiency refers to the test of association between disease and a single explanatory factor; after adjustment for confounding, the efficiency of the nested case-control study may not be so good.

The above discussion assumes that no information is available about the subjects in the study cohort except for their observation periods and their disease status. Almost always, however, some incomplete risk factor data will be available for every subject. There is now the possibility of stratified sampling of controls—we classify members of each risk set into strata according to this incomplete data, and sample the different strata independently. We shall consider two such stratified sampling schemes.

Matching

In matched nested case-control studies, controls are drawn from the same stratum as the case; the number of controls drawn from each stratum may be chosen to vary according to where the case falls. The partial likelihood is unchanged and comparisons between cases and controls are made within strata. Thus, stratifying variables are treated as confounders. This design is well known and its aim is efficient control for confounding. Its limitation is that effects of stratifying variables, after controlling for any variables in the sample, cannot be estimated, although their role as effect modifiers may be studied. Matching for variables that are not to be regarded as confounders is to be avoided, first because its effect cannot be estimated and second because the study loses power with respect to variables associated with it (the problem of overmatching).

Counter-matching

An alternative stratified design with quite different aims has recently been proposed (*7*). This design employs stratified sampling of controls with strata determined not by confounders but by variables of interest in the analysis, or by proxies for such variables. The aim of the design is, again, improvement in efficiency; but to achieve efficiency in this case, we require a design that is almost directly opposite to matching. Whereas the aim of stratified sampling by confounders is to yield controls that are as similar as possible to the cases, stratified sampling by exposures of interest aims to maximize the variation of exposure within case-control sets.

We will denote the total number of subjects in a risk set by n, and the numbers falling into exposure categories 0, 1,... by n_0, n_1,... respectively. We require a different distribution of subjects in our nested case-control study—m_0, m_1,... let us say. This requires that, if stratum i contains the case, we draw at random a sample of size m_i-1 from the n_i-1 eligible subjects and, if stratum i does not contain the case, we draw m_i from n_i. The name "counter-matching" for the above design is suggested by the special case of the 1:1 study when there are two sampling strata and we require $m_0 = m_1 = 1$. In this case the design requires that the control be drawn from the opposite stratum to the case.

The partial likelihood method of analysis may be simply adapted to the counter-matched design. The contribution of each case-control set to the log partial likelihood becomes

$$\log[(W\theta)_{(\text{for case})} / \sum_{\text{case-control set}} (W\theta)]$$

where W are risk weights that depend upon the sampling stratum from which the subject is drawn. For a subject drawn from sampling stratum i, the risk weight is

$$W = \frac{n_i}{m_i} .$$

Note that risk weights apply not only to controls but also to cases.

This method of analysis allows us simultaneously to estimate effects of the stratifying variables, which are measured in the whole cohort, and of further variables available only in the nested case-control study. Computer implementation of the analysis is straightforward, using existing programs for conditional logistic regression or for Cox's life table regression. Some packages explicitly allow the definition of prior risk weights while in others the same effect can be achieved by use of an offset in a log-linear regression model.

When to Use Counter-matching

The need for matching has been extensively debated in the epidemiologic literature and is now well understood. If there are strong confounders whose effects must be controlled in the analysis, failure to match for them in the case-control study means that the efficiency of the study relative to a full cohort study may fall well short of the $(m-1)/m$ rule indicated above. This relative efficiency can be recovered, however, by drawing matched controls from the same confounder stratum as the case.

The counter-matched design, however, is novel. In this section we shall indicate some possible uses of the design and present some illustrative calculations of its relative efficiency using formulas derived in Langholz and Borgan (*7*).

Counter-matching by Surrogate Exposure

It may be that a crude measure of exposure is easily obtained for all subjects in a cohort study, but good exposure measurements are expensive to collect. In these circumstances it can be advantageous to stratify risk sets according to the surrogate exposure measure and to sample controls according to the counter-matched design.

Example 1. Breast Cancer and Oral Contraceptives (OCs). In pharmicoepidemiologic studies, subjects can accurately remember if and when medication was

used but are much less likely to remember the exact name of the drug. The control selection would then be stratified by whether the subject took any of the drugs within the class. For instance, suppose it is of interest to assess the risk of breast cancer from a specific formulation of oral contraceptive as part of a larger cohort study of breast cancer. Women could be asked, on a mailed questionnaire, the dates that they took oral contraceptives. This information is likely to be quite accurate (granting that some women may not wish to answer the question). However, the mailed questionnaire is much less likely to provide accurate information about the exact brand names of OCs used since subjects will have a much poorer memory of such details. Thus it is natural to consider conducting a nested case-control study in which medical records are obtained or subjects are interviewed personally to obtain accurate brand name information. While a simple nested case-control sample could be considered, it would seem wasteful of the preliminary data held for the full cohort. Counter-matching makes use of this information to obtain a more informative case-control study.

Example 2. Pancreatic Cancer and Occupational Exposure. In occupational cohort studies, personnel records may contain enough data about workplace history to provide a surrogate exposure measure. An example is the actual investigation of the role of occupational exposures in the risk of pancreatic cancer in a cohort of 5886 chemical manufacturing workers (*8*). Basic information about each employee was available from the company's personnel records, including job titles and workplace at the plant. There were 28 cases of pancreatic cancer and, as stratified sampling was not known at the time, a simple 1:4 matched nested case-control study was drawn and detailed chemical exposure information was obtained for the 140 subjects in the sample. A counter-matched design would have been well suited to this situation. Each of the 28 cases and the corresponding risk sets would be stratified according to the estimated level of general chemical exposure calculated from job titles and workplace location histories, and counter-matched sampling would be carried out as indicated above.

Relative efficiency calculations indicate that the efficiency advantage of counter-matching over simple sampling in these situations can be substantial. Let Z represent a dichotomous exposure and $\tilde{Z}$ be a crude measure of it, also dichotomous, and denote the rate ratio for exposure by ψ:

$$\psi = \frac{\text{Incidence rate } (Z = 1)}{\text{Incidence rate } (Z = 0)}$$

Tables 1 and 2 investigate the asymptotic relative efficiency (ARE) of the 1:1 counter-matched design as compared with a simple 1:1 unstratified design defined by

$$\text{ARE} = \frac{\text{Variance of } \log\hat{\psi} \text{ from simple study}}{\text{Variance of } \log\hat{\psi} \text{ from counter–matched study}}$$

in relation to the rate ratio ψ, the probability of exposure $Pr(Z=1)$, the sensitivity of the surrogate measure $Pr(\tilde{Z}=1|Z=1)$, and its specificity $Pr(\tilde{Z}=0|Z=0)$. At the null hypothesis, $\psi=1$, the ARE of the counter-matched design depends on the sensitivity and specificity of the surrogate exposure measure but not on the probability of exposure. Denoting the false positive and false negative probabilities by α and β respectively, so that the sensitivity is $1-\beta$ and the specificity is $1-\alpha$,

$$\text{ARE}_{\psi=1} = 2[(1-\alpha)(1-\beta)+\alpha\beta].$$

This relationship is illustrated in Table 1. Note that the ARE obtained by measuring true exposure in the full cohort relative to a simple 1:1 nested study is 2. As we might expect, this is achieved when the sensitivity and specificity are both 1 ($\alpha=\beta=0$). Table 2 investigates the efficiency of the design well away from the null ($\psi=4$). In this case the ARE of the counter-matched design depends on the probabilities of exposure in the cohort, and three values are investigated in the table (0.05, 0.1, and 0.2). The counter-matched design has an efficiency advantage in all the circumstances investigated, this advantage increasing with the rarity of exposure. Even when the crude measure is not very predictive of exposure, the counter-matched design can be substantially more efficient than a simple sampling design.

Without using counter-matching, the only way to achieve greater efficiency in a nested case-control study is to increase the number of controls drawn for each case. However, this can be much less cost effective than counter-matching. Table 1 shows (in parentheses) the number of controls, $m-1$ required in a simple $1:m-1$ design to achieve the same efficiency as the counter-matched 1:1 design. For example, if the surrogate measurement has specificity of 1.0 but a sensitivity of 0.8, a simple unstratified study would require four controls for each case to achieve the same efficiency as a 1:1 counter-matched design. In this case the counter-matched design achieves 80% of the full cohort efficiency.

Counter-matching by Exposure

Suppose that exposure information has been collected and coded for all cohort subjects and there is an observed association between this exposure and disease. The next natural step in assessing whether this association is causal is to see whether it can be explained by the confounding effect of other factors, as yet unmeasured. To answer this question it is natural to collect detailed data concerning potential confounders in a nested case-control study.

If the exposure of interest is rare, a simple nested case-control design might include few exposed controls and the distribution of confounders in this group might be poorly estimated. A more efficient

Table 1. Asymptotic relative efficiencies for 1:1 counter-matching by a surrogate exposure measure.

	Specificity		
Sensitivity	1.0	0.9	0.8
0.40	0.80	0.84	0.88
0.50	1.00 (1)	1.00	1.00
0.70	1.40 (2)	1.32	1.24
0.80	1.60 (4)	1.48	1.36
0.90	1.80 (9)	1.64	1.48
0.95	1.90 (19)	1.72	1.54
1.00	2.00 (∞)	1.80	1.60

Table 2. ARE for 1:1 counter-matching by surrogate exposure, $\psi=4$.

	$Pr(Z=1)=0.05$ Specificity			$Pr(Z=1)=0.1$ Specificity			$Pr(Z=1)=0.2$ Specificity		
Sensitivity	1.0	0.9	0.8	1.0	0.9	0.8	1.0	0.9	0.8
0.40	1.89	1.42	1.16	1.79	1.38	1.14	1.61	1.30	1.11
0.50	2.33	1.72	1.36	2.17	1.65	1.33	1.92	1.53	1.28
0.70	3.17	2.31	1.76	2.89	2.17	1.69	2.46	1.94	1.57
0.80	3.57	2.59	1.95	3.23	2.41	1.86	2.70	2.12	1.70
0.90	3.97	2.87	2.14	3.54	2.64	2.02	2.92	2.29	1.82
0.95	4.16	3.00	2.23	3.70	2.76	2.10	3.03	2.37	1.87
1.00	4.35	3.14	2.33	3.85	2.87	2.17	3.12	2.45	1.92

design would draw more equal-sized samples of exposed and unexposed controls. This is possible using counter-matching.

Example 3. The Beaverlodge Uranium Miners Cohort. In this study, radon exposure histories were assembled for 10,908 miners ever employed at the Beaverlodge Mine between 1950 and 1980. After an initial analysis of the cohort to describe associations between radon exposure and lung cancer mortality (*9*), the extent to which smoking confounded the observed relative risks for radon exposure was investigated in a nested case-control sample (*10*). Each of 89 cases was matched to three controls matched as closely as possible for birth year, province of death, and year of death. The final study group consisted of 46 cases and 95 controls for whom next of kin were found who agreed to be interviewed regarding the subject's smoking and occupational histories and providing demographic information. The results indicate that radon exposure was "negatively confounded" by smoking, i.e., the adjusted relative risks for radon are higher than the unadjusted. A counter-matched sample would have used the radon exposure information to stratify the sample. Risk-set members would be classified into strata defined by their radon exposure at the age of the death of the case. When exposure is a continuous measure such as this, subjects should be classified so that about equal numbers of cases appear in each stratum. If the distribution of exposure does not change much with age, classification could be done by dividing at median exposure of the cases and taking equal numbers from the two groups. Otherwise, the method described in Langholz and Borgan (*7*) could be used.

Example 4. Effectiveness of Cancer Screening. Nested case-control studies have proved extremely useful in providing evidence for the effectiveness of cancer screening programs. For example, Verbeek et al. (*11*) employed the population register used to generate breast cancer screening appointments to define a study cohort. The outcome of interest was death from breast cancer and the exposure of interest was the screening history—it was hoped that screened women would have a lower mortality rate from breast cancer than unscreened women. For each death, a set of controls was drawn from the corresponding risk set and the screening histories of cases and controls were compared. Studies such as this can be carried out using only the screening program records plus linked mortality data; and, in principle, the nested case-control design is unnecessary since screening histories are available for the whole cohort. However, an apparent benefit of screening may be attributable to differences between program compliers and noncompliers with respect to other aspects of behavior. To exclude such confounding, a more detailed nested case-control study could be carried out, in which information about possible confounders would be sought. However, simple sampling of risk sets might yield rather few unscreened controls, and a counter-matched design could be more efficient.

We have investigated the efficiency of counter-matching for this purpose analytically. Let Z_1 be the exposure known for the full cohort and Z_2 be the confounder to be collected for the sample. Both are assumed dichotomous. The associations between Z_1 and Z_2 and disease, when they are mutually adjusted, are given by the relative risks ψ_1 and ψ_2) and the strength of association between the exposure and confounder is measured by the odds ratio

$$\theta = \frac{Pr(Z_1 = 1, Z_2 = 1)Pr(Z_1 = 0, Z_2 = 0)}{Pr(Z_1 = 1, Z_2 = 0)Pr(Z_1 = 0, Z_2 = 1)}$$

Table 3 shows the relative efficiencies of the counter-matched design compared to a simple nested case-control design for estimating the exposure log relative risk $\log\psi_1$ after controlling for a confounder, Z_2. A range of values for the effects ψ_1 and ψ_2 and for the exposure:confounder odds ratio θ are investigated when both exposure and confounder are rare and when both are common. In addition to 1:1 matched samples, a "balanced" 1:3 counter-matched sample, i.e., with two exposed and two unexposed subjects in each case-control set, is compared to a 1:3 simple nested case-control sample. As expected, the counter-matched study has a marked efficiency advantage, especially in the rare exposure situation. In this case, the relative efficiency increases sharply with increasing relative risk of exposure (ψ_1). When exposure is common, this increase is much less pronounced and, in fact, may decrease, reflecting the difference in efficiency behavior of nested case-control sampling for rare and common exposures (*5*). The last two columns of the table give the relative efficiencies of the 1:1 counter-matched sample to the full cohort. In the rare exposure situation, the efficiency is quite high except when the Z_2 is strongly associated with both disease and exposure. (Note that in this case it is still much better than simple nested case-control sampling.) In the situation where Z_2 is associated with neither disease nor exposure ($\psi_2 = \theta = 1.0$), counter-matching is fully efficient. This behavior is to be expected, since stong confounders are highly relevant to assessing the

Table 3. ARE of counter-matched study for estimating the effect of exposure after controlling for a confounder.

		$Pr(Z_1=1)=0.01$, $Pr(Z_2=1)=0.05$					
		1:1 counter-matched (vs simple 1:1)		1:3 counter-matched (vs simple 1:3)		1:1 counter-matched (vs full cohort)	
ψ_2	θ	ψ_1=1	ψ_1=4	ψ_1=1	ψ_1=4	ψ_1=1	ψ_1=4
0.2	0.10	1.97	4.80	1.35	2.33	0.95	0.95
	1.00	2.00	4.82	1.35	2.31	0.97	0.97
	10.0	1.93	4.36	1.32	2.20	0.99	0.98
1.0	0.10	1.93	4.68	1.33	2.27	0.96	0.96
	1.00	2.00	4.85	1.33	2.28	1.00	1.00
	10.0	1.56	3.83	1.31	2.15	0.78	0.81
5.0	0.10	1.95	4.49	1.32	2.17	0.99	0.99
	1.00	2.00	4.73	1.43	2.41	0.89	0.89
	10.0	1.33	3.52	1.64	2.58	0.38	0.42
		$Pr(Z_1=1)=0.5$, $Pr(Z_2=1)=0.5$					
		1:1 counter-matched (vs simple 1:1)		1:3 counter-matched (vs simple 1:3)		1:1 counter-matched (vs full cohort)	
ψ_2	θ	ψ_1=1	ψ_1=4	ψ_1=1	ψ_1=4	ψ_1=1	ψ_1=4
0.2	0.10	1.70	1.81	1.25	1.21	0.66	0.84
	1.00	2.00	1.93	1.33	1.29	0.78	0.78
	10.0	1.70	1.45	1.25	1.22	0.66	0.46
1.0	0.10	1.57	1.48	1.22	1.19	0.79	0.71
	1.00	2.00	2.00	1.33	1.27	1.00	1.00
	10.0	1.57	1.48	1.22	1.19	0.79	0.71
5.0	0.10	1.70	1.45	1.25	1.22	0.66	0.46
	1.00	2.00	1.93	1.33	1.29	0.78	0.78
	10.0	1.70	1.81	1.25	1.21	0.66	0.84

Table 4. ARE of counter-matched study for estimating the effect of exposure after controlling for a confounder.

		$Pr(Z_1=1)=0.05$, $Pr(Z_2=1)=0.01$					
		1:1 counter-matched (vs simple 1:1)		1:3 counter-matching (vs simple 1:3)		1:1 counter-matching (vs full cohort)	
ψ_2	θ	$\psi_1=1$	$\psi_1=4$	$\psi_1=1$	$\psi_1=4$	$\psi_1=1$	$\psi_1=4$
0.2	0.10	0.26	0.63	0.90	0.93	0.21	0.52
	1.00	0.30	0.74	0.91	0.97	0.24	0.58
	10.0	0.55	1.23	0.95	1.12	0.46	0.80
1.0	0.10	0.11	0.37	0.70	0.78	0.05	0.19
	1.00	0.19	0.62	0.73	0.90	0.09	0.29
	10.0	0.67	1.87	0.89	1.42	0.34	0.60
5.0	0.10	0.09	0.28	0.50	0.59	0.02	0.05
	1.00	0.29	0.74	0.58	0.84	0.03	0.12
	10.0	1.31	3.03	0.97	1.97	0.08	0.35
		$Pr(Z_1=1)=0.5$, $Pr(Z_2=1)=0.5$					
		1:1 counter-matched (vs simple 1:1)		1:3 counter-matching (vs simple 1:3)		1:1 counter-matching (vs full cohort)	
ψ_2	θ	$\psi_1=1$	$\psi_1=4$	$\psi_1=1$	$\psi_1=4$	$\psi_1=1$	$\psi_1=4$
0.2	0.10	1.08	0.85	1.02	1.01	0.51	0.41
	1.00	1.00	0.83	1.00	0.97	0.50	0.36
	10.0	1.08	0.93	1.02	0.98	0.51	0.31
1.0	0.10	1.00	0.78	1.00	0.99	0.50	0.32
	1.00	1.00	0.78	1.00	0.99	0.50	0.32
	10.0	1.00	0.78	1.00	0.99	0.50	0.32
5.0	0.10	1.08	0.93	1.02	1.01	0.51	0.31
	1.00	1.00	0.83	1.00	0.99	0.50	0.36
	10.0	1.08	0.85	1.02	1.05	0.51	0.41

true association between exposure and disease, whereas factors that are neither related to disease nor exposure are irrelevant, and there is nothing to be gained by efficient sampling designs to study such a factor.

Discussion

In this article we have shown that matching and counter-matching have opposite roles and both, when properly used, can lead to gains in efficiency. Matching should be used for stratified sampling with respect to a confounder while counter-matching should be used for stratified sampling with respect to an exposure of interest or a proxy.

Reversal of these roles is not advantageous. Matching for a variable closely associated with exposure is known to lead to reduced efficiency and is called overmatching. Likewise, counter-matching with respect to a confounder should not be considered. Table 4 shows the relative efficiencies for estimating the rate ratio for exposure, ψ, in this situation; counter-matching may be seen usually to be very inefficient.

Nested case-control studies are very useful for assessing the determinants of variations of cancer rates within a population. Because subjects in the cohort are identified without regard to disease status, there is little risk of the selection biases that may occur in a purely retrospective case-control study. For data generated from material collected as part of the cohort study, information bias is less likely because it is collected prior to the onset of disease. Substantial cost savings often may be realized by restricting data extraction and coding to the nested case-control sample. Other source data may also be gathered "retrospectively" on the sampled subjects, although the danger of information bias would need to be carefully considered.

Simple nested case-control sampling is generally quite cost effective relative to the analysis of the entire cohort. The increase in efficiency per additional control depends upon what is being investigated in the substudy and the level of efficiency is determined by the number of controls per case. In many important situations, such as assessing the relationship between disease and a rare exposure, evaluation of the role of confounding, or variation in relative risk with a potential effect modifier, many controls may be required to achieve a specified level of efficiency. Matching and counter-matching provide alternative means to this end by using information collected for the entire cohort to stratify the sampling in such a way as to increase the efficiency per sampled control compared to simple nested case-control sampling. The cost effectiveness of these strategies will depend on the and the cost of collecting and coding stratifying variables.

REFERENCES

1. Morris JN, Chave SPW, Adam C, Sirey C. Vigorous exercise in leisure-time and the incidence of coronary heart-disease. Lancet 1:333–339 (1973).
2. Mantel N. Synthetic retrospective studies and related topics. Biometrics 29:479–486 (1973).
3. Cox DR. Regression models and life-tables (with discussion). J R Stat Soc Ser B 34:187–220 (1972).
4. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. Ann Stat 10:1100–1120 (1982).
5. Breslow NE, Lubin JH, Marek P, Langholz B. Multiplicative models and cohort analysis. J Am Stat Assoc 78: 1–12 (1983).
6. Goldstein L, Langholz B. Asymptotic theory for nested case-control sampling in the Cox regression model. Ann Stat 20:1903–1928 (1992).
7. Langholz B, Borgan Ø. Counter-matching: a stratified nested case-control sampling method. Biometrika (in press).
8. Garabrant D, Held J, Langholz B, Peters J, Mack T. DDT and related compounds and the risk of pancreatic cancer. J Natl Cancer Inst 84:764–771 (1992).
9. Howe G, Nair R, Newcombe H, Miller A, Abbatt J. Lung cancer mortality (1950–1980) in relation to radon daughter exposure in a cohort of workers at the Eldorado Beaverlodge mine. J Natl Cancer Inst 77:357–362 (1986).
10. L'Abbe K, Howe G, Burch J, Abbatt J, Band P, Choi W, Du J, Feather J, Gallagher R, Hill G, Matthews V. Radon exposure, cigarette smoking and other mining experience in the Beaverlodge miners cohort. Health Phys 60:489–495 (1991).
11. Verbeek ALM, Hendricks JHCL, Holland R, Mravunac M, Sturmans F, Day NE. Reduction of breast cancer mortality through mass screening with modern mammography. Lancet 1:1222–1224 (1984).