AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 131, No. 1
Printed in U.S.A.

NESTED CASE-CONTROL AND CASE-COHORT METHODS OF SAMPLING FROM A COHORT: A CRITICAL COMPARISON[1]

BRYAN LANGHOLZ AND DUNCAN C. THOMAS

Langholz, B. (Dept. of Preventive Medicine, U. of Southern California, School of Medicine, Los Angeles, CA 90033-9987) and D. C. Thomas. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *Am J Epidemiol* 1990;131:169–76.

The recently developed case-cohort method of sampling from a cohort is compared with the nested case-control method. Corrected asymptotic relative efficiency results show that the case-cohort design for single "disease" outcomes offers less improvement for intervention trials for which there is no random censoring than originally suggested. Furthermore, simulation results indicate that if there is moderate random censoring or staggered entry, the case-cohort method can do substantially worse than the nested case-control method.

biometry; epidemiologic methods; follow-up studies; research design; statistics

The partial likelihood method (1, 2) for fitting the proportional hazards model has served as the theoretical foundation for most of the recent discussion of the analysis of censored survival data. The proportional hazards model specifies that the incidence rate at time t for persons with the vector of quantified risk factors z is given by

$$\lambda(t;z) = \lambda_0(t) \exp(z'\beta)$$

where $\lambda_0(t)$ is the baseline incidence and β is the vector of regression parameters.

In the partial likelihood approach to fitting the proportional hazards model, the covariate values of each case (failure) are compared with those of all subjects who were at risk at the time the case failed. This method has been widely used in the analysis of clinical trial data. Although theoretically appropriate, it has been less widely used for epidemiologic cohort studies and large scale prevention trials because the large sample sizes make the computational costs prohibitive. Alternatives based on the use of Poisson regression models with grouped data (3, 4) have been shown to offer nearly full statistical efficiency at a small fraction of the computing cost. However, covariate information must still be collected on the entire cohort, and the need for grouping in this approach to some extent limits detailed examination of covariate effects.

Another alternative is to sample from the original cohort (5). The most widely used sampling scheme, known as "nested case-control" sampling, involves selection of a (generally fixed) number of "controls" from those at risk at the failure time of each case (6). The resulting sample is then analyzed using standard techniques for the analysis of matched case-control studies.

Recently, several other sampling schemes have been proposed. The most widely dis-

Received for publication March 24, 1988, and in final form June 19, 1989.

[1] From the Department of Preventive Medicine, University of Southern California, School of Medicine, 2025 Zonal Avenue, Los Angeles, CA 90033-9987. (Reprint requests to Dr. Langholz.)

This work was supported by National Cancer Institute grants CA14089 and CA42949.

The authors thank Edward Rappaport for programming assistance and Dr. Malcolm Pike for helpful suggestions.

cussed of these sampling schemes is the "case-cohort" design of Prentice et al. (7–9). In this design, the sample consists of a random subcohort from the entire cohort plus all cases that occur outside the subcohort. The case-cohort design was originally introduced as a method of analysis for prevention trials, which typically are relatively short in duration and censoring occurs primarily because of termination of the study.

For both the nested case-control and case-cohort designs, covariate information needs to be assembled only on the cases and the selected controls so that the data collection and data management costs may be only a fraction of those of the full cohort study. Also, because there is no need to group the subjects, the limitations on covariate modeling imposed by the usual methods of person-years analysis and Poisson regression are avoided.

In this presentation, we discuss the nested case-control and case-cohort designs focusing especially on relative efficiency issues. We show that, even for the intervention trial for which one might expect that case-cohort sampling should do very well, the gain in efficiency for case-cohort over nested case-control sampling is surprisingly small.

In the presence of moderate random right censoring or when subjects can enter the study at various times, case-cohort sampling can be much less efficient than nested case-control sampling. Variants of simple case-cohort sampling need to be studied to see if this loss of efficiency can be avoided.

The case-cohort design does provide one with the ability to evaluate the relative risks for multiple disease outcomes without the need to draw a new comparison series and to estimate baseline hazards and such summary measures as standardized mortality ratios (10). We also examine these issues and other practical considerations affecting the choice of sampling design.

IDEALIZED INTERVENTION TRIALS

In this portion of the paper, we compare the efficiency of the nested case-control and case-cohort designs for "idealized intervention trials," in which all persons enter the study at time zero and are followed for a fixed period of time during which they either "fail" or are alive at the end of the study.

Asymptotic relative efficiencies

The asymptotic relative efficiency for nested case-control sampling, the ratio of the variance of the parameter estimated from the sample to that estimated from the full cohort as the cohort size increases to infinity, is given for binary exposures by Breslow et al. (3). The corresponding formula for case-cohort sampling of an idealized intervention trial is given by Self and Prentice (9). Efficiency arguments in favor of the case-cohort design are based on the assumption that the primary component to the cost of a study is the number of *distinct* persons used (8, 9).

In Appendices 1 and 2 we derive the expected number of distinct persons used in the nested case-control design and the case-cohort design for the idealized intervention trial for the null situation in which there is no effect of the intervention. Expression 4 of Appendix 2 shows the subcohort proportion required in the case-cohort design to yield the same expected number of distinct persons as the nested case-control design. Using these results to produce comparable designs, i.e., equivalent in terms of the expected number of distinct persons used, the formulas of Breslow et al. (3) and Self and Prentice (9) allow one to calculate their asymptotic relative efficiencies.

Table 1 shows the results of these calculations for samples with the expected number of distinct persons equal to the expected number that arise from nested case-control samples with 1:1, 1:3, and 1:5 matching ratios of cases to controls. The third column of the table gives the proportion of the cohort which needs to be sampled in order to obtain a case-cohort sample of the same average size as a nested case-control sample. The fourth and fifth columns give the

TABLE 1

*Asymptotic relative efficiency of nested case-control and case-cohort sampling methods controlling for the distinct number of persons in idealized intervention trials with no effect of treatment**

Matching	Overall disease probability	Subcohort fraction	Case-control	Case-cohort	Ratio†	Ratio from Self and Prentice (9)‡
1:1	0.05	0.050	0.50	0.51	1.02	1.04
	0.10	0.100	0.50	0.52	1.04	1.10
1:3	0.05	0.143	0.75	0.77	1.03	1.05
	0.10	0.271	0.75	0.78	1.04	1.11
1:5	0.05	0.226	0.83	0.85	1.02	1.06
	0.10	0.409	0.83	0.87	1.05	1.11

* Failure times are assumed to be exponentially distributed, with 0.5 probability of treatment.

† Asymptotic relative efficiency of case-cohort design/asymptotic relative efficiency of the nested case-control design.

‡ From previously published asymptotic relative efficiencies of the case-cohort design which do not account for repeated sampling of persons in the nested case-control design.

asymptotic relative efficiencies for nested case-control and case-cohort methods, respectively, and the sixth column gives their ratio. The final column of table 1 gives the ratio of the asymptotic efficiencies as reported by Self and Prentice (9). These authors failed to control for the possibility of repeated sampling of persons in the nested case-control design so their ratios are somewhat exaggerated. The asymptotic relative efficiencies are higher for the case-cohort design, but the differences are small. The incorrect large differences with increasing disease probability previously reported (9) are almost entirely explained by the fact that the larger the proportion of failures, the higher the chance that persons will be picked more than once in the nested case-control sampling since a larger proportion of the cohort is sampled.

Simulation study

We have shown above (table 1) that the asymptotic relative efficiency of the two designs is very close for the idealized intervention trial when there is no effect of "treatment." As part of a series of simulation studies, we have investigated the comparative efficiency of the two designs for the idealized intervention trials as above when the relative risk is either 1 or 2 (table 2). As in the simulations of Prentice et al. (7, 8), we simulated cohorts of size 500 persons, with 250 exposed and 250 unexposed, for relative risks of 1 and 2. Failure times were based on randomly generated exponential variates and with failures rates and censoring in each situation set to yield 50 expected failures. Nested case-control sampling was performed with 1:1, 1:3, and 1:5 matching ratios.

In order to assure that the case-cohort samples yield the same expected number of distinct persons as the nested case-control samples, we used the following strategy: 1) The number of distinct persons was tabulated for each of 200 complete simulations of nested case-control sampling, i.e., for each simulation, a cohort was generated and a nested case-control sample was drawn. 2) The expected number of distinct persons for nested case-control samples was estimated from the samples generated in step 1 of this strategy. The subcohort size which will yield the same expected number of distinct persons for case-cohort samples was calculated using the formula given in Appendix 2. 3) For each of 200 complete simulations of both designs; i.e., for each simulation, a nested case-control sample and a case-cohort sample were drawn from the cohort, the maximum likelihood estimates of β were calculated for each design. The size of the subcohort was

TABLE 2

Empirical variances of β for nested case-control and case-cohort sampling designs under differing study designs

Study design	1:3 matching			1:5 matching		
	Case-control	Case-cohort	Ratio	Case-control	Case-cohort	Ratio
	Relative risk = 1 (β = 0)					
Intervention trial						
No loss to follow-up	0.10	0.09	1.06	0.09	0.08	1.08
10% loss to follow-up	0.12	0.11	1.08	0.08	0.07	1.08
20% loss to follow-up	0.11	0.11	1.00	0.12	0.11	0.94
Clinical trial	0.13	0.16	0.82	0.13	0.13	0.99
Occupational cohort	0.12	0.19	0.64	0.09	0.14	0.62
	Relative risk = 2 (β = 0.693)					
Intervention trial						
No loss to follow-up	0.11	0.10	1.12	0.12	0.11	1.09
10% loss to follow-up	0.12	0.11	1.02	0.11	0.11	1.04
20% loss to follow-up	0.10	0.10	0.97	0.13	0.13	1.08
Clinical trial	0.12	0.17	0.71	0.11	0.13	0.86
Occupational cohort	0.13	0.23	0.57	0.10	0.14	0.74

fixed to the number calculated in step 2 of this strategy. The estimates of β from these simulations were then used to calculate the empirical variances given in table 2.

The variances computed from the simulation studies for the idealized intervention trials with 1:3 and 1:5 matching are given in the first and sixth rows of table 2. The results when the relative risk is 2 are analogous to those when the relative risk is 1, with case-cohort sampling marginally more efficient than nested case-control sampling.

For the idealized intervention trial for the 1:1 matching situation, the case-cohort method did considerably better than the nested case-control method with empirical relative efficiency of 1.33. This is considerably different from the ratio of 1.04 (table 1) predicted by the asymptotic theory. This is because, with 50 failures, $\beta = 0$, and 1:1 matching, the expected number of discordant pairs is 25 and the empirical nested case-control variance of $\hat{\beta}$ is not estimated well by the variance estimator based on asymptotic theory, the inverse information. This tendency was also observed by Prentice et al. (7). When we generated 200 cohorts of size 1,000 with $\beta = 0$, yielding an average of 100 failures, we obtained empirical variances for β of 0.093 and 0.087 for nested case-control and case-cohort sampling, respectively. These are somewhat more consistent with the theory and validate the small sample explanation of the observed difference with the theory. Given this observation, asymptotic efficiency comparisons are likely to be better represented by the 1:3 and 1:5 matching situations and only these are included in table 2. Also, it warrants a cautionary note to practitioners of nested case-control sampling that when the number of failures is small, multiple controls should be used to ensure that confidence intervals and significance tests are reliable.

OTHER COHORTS

It has been suggested that case-cohort sampling may be more efficient than nested case-control sampling for a wide range of cohort designs. In order to explore whether this is true or not, we performed a series of simulation studies for which we examined the performance of the two sampling designs for the following additional cohort types: 1) "intervention trial with loss to follow-up" in which some persons are censored before the end of the study, 2) "clin-

ical trial" in which all persons enter at time zero with random right censoring, and 3) "occupational cohort study" with staggered entry and random right censoring.

In order to investigate the effect of loss to follow-up, for a proportion of persons, chosen to yield a given expected loss to follow-up, a uniformly distributed censoring time was generated. If this censoring time was less than the failure time, the person was considered as lost to follow-up, and thus censored, at that time. We simulated cohorts for which there was 10 or 20 percent expected loss to follow-up.

For clinical trial cohorts, an exponentially distributed censoring time was generated and compared with the failure time. The person failed or was censored at the minimum of these two times.

For occupational cohorts, each person entered the cohort at an exponentially distributed entry time. The failure or censoring time was then determined by adding a time, as generated for the clinical trial, to the entry time.

Numbers of exposed and unexposed persons and the method for assuring that the expected number of distinct persons is the same for both designs were as described in simulation study. Censoring and failure parameters were chosen to yield 50 expected failures.

As anticipated by statistical theory, in every simulation, for both nested case-control and case-cohort designs, the average score was not significantly different from zero. A slight upward bias was noted in the estimates of β when the relative risk was 2 ($\beta = 0.693$) but it seemed equivalent for case-cohort and nested case-control designs and decreased with the number of controls. The empirical variances for each situation are given in table 2.

Examining table 2, one sees that loss to follow-up seems to reduce the small efficiency advantage of the case-cohort design. The clinical trial example may be viewed as an extreme case of loss to follow-up and nested case-control sampling yields somewhat lower variances than the case-cohort. The largest difference in variance is observed in the occupational cohort situation. We treated our occupational cohort study as retrospective in which all persons in the study are ascertained prior to sampling and the case-cohort subcohort is simple a random sample of persons from the entire cohort. In this situation, nested case-control sampling does substantially better than the case-cohort because, with high probability, the sampled subcohorts would yield risk sets with no controls resulting in a case being "thrown away."

Although it is unlikely that more complex or multiple covariates will change the conclusions based on table 2, more research needs to be done in order to assess how covariate configurations effect efficiency. The analysis of the Montana smelter workers data by Breslow et al. (3) is often cited as showing that, for nested case-control sampling, 20 controls per case may be needed in order to achieve good efficiency when multiple covariates are to be assessed. It should be pointed out, however, that they allowed multiple failures for each risk set so that it is not possible to apply their experience to the situation in which there is only one failure per risk set.

Discussion

For retrospective studies, the major issue in the decision as to whether to choose case-cohort or nested case-control designs is based on the relative efficiency of the designs since the major cost involved may be proportional to the number of persons for whom covariate information must be obtained. For prospective cohort studies, including intervention trials, there are two additional considerations about case-cohort sampling which should be kept in mind.

While the subcohort can be used to monitor the progress of the study, everyone in the cohort must be followed with equal rigor. In a rather obvious example, if members of the subcohort in an intervention trial were regularly screened to increase compliance with a given regimen while the

rest of the cohort was not, serious bias might arise. Cases occurring outside the subcohort would have complied to a lesser degree than subcohort members. If compliance leads to lower risk of the disease of interest, the relative risk will be underestimated. Prospective nested case-control sampling is not subject to this potential bias because it is not known in advance who will be controls.

Another problem is that of collecting time-dependent information over the course of the study. This is illustrated by a problem which a colleague encountered in using the case-cohort design in a prospective study for which blood was drawn from each participant at the beginning of the study (13). Certain of the blood products of interest deteriorate with time so that ideally one would like to analyze the blood as soon as it is obtained. Unfortunately, this is not feasible and can be done only for a sample of persons. For the case-cohort design, if one were to analyze the blood of persons in the subcohort at the time it is collected, bias would arise because the blood from cases arising outside the subcohort would have suffered some deterioration. This is the problem discussed in the preceding paragraph. To ensure that there is no differential deterioration of the blood, one needs to analyze the subcohort and cases' blood at the end of the study. This precludes interim analyses, requires a large amount of laboratory work at the end of the study, and uses blood which has deteriorated the entire length of the study. An alternative is to analyze the blood in batches and stratify by the time at which the blood was analyzed. While unbiased, this complicates the study and sacrifices efficiency. The nested case-control design does not suffer from this problem. Case and matched control blood samples are analyzed at the same time so that they are matched on deterioration of the blood. The only stipulation is that there needs to be enough blood sampled so that if a person arises in the study at multiple times, there is enough blood for multiple analyses. We stress that the only perfect solution is to analyze all the blood at the beginning of the study. If that is not possible, care must be taken to match on the time that the blood is analyzed.

Efficient estimation of a single disease relative risk may not be the only goal of a cohort study. A major advantage of the case-cohort design is the ability to analyze multiple outcomes with a single sample. This may be very valuable for cohort studies of general health, although adjustment of significance levels may be required to account for the induced correlation between outcomes.

Also, it is simple to compute standard mortality ratios relative to external rates (10) and baseline hazard estimates (9). One simply divides the person-years in each "cell" or the usual baseline hazard estimator derived from the case-cohort sample by the sampling fraction (the number in the subcohort divided by the total number in the cohort). While we are not aware of methods which provide for person-years analysis with the nested case-control design, methods are available for estimating the standard mortality ratio as a smooth function of time and covariates if one knows the number at risk in the cohort for each risk set (11, 12). These same techniques apply to estimating the baseline hazard.

Computationally, estimates, variances, and significance tests are easily obtained for nested case-control samples using any of the software packages available for the analysis of case-control studies. Case-cohort samples, however, require special and difficult to program algorithms in order to calculate quantities used in significance testing and asymptotic variance estimators of parameter estimates.

The results of our investigation lead us to conclude that single disease relative efficiency is not a reason to choose case-cohort over nested case-control sampling. In fact, efficiency should play only a minor

role in deciding which of the sampling methods is best for a given study. While we observed a substantial loss in efficiency by the case-cohort design for clinical trials and occupational-type cohorts, it is possible that variants of the design, such as "refreshing" the subcohort (8), may inhibit this loss. The actual goals of the study will generally make it clear which is the more appropriate sampling method. In studies of general health, the case-cohort design may be advantageous because it is important to have the ability to analyze multiple outcomes. If the study has one specific outcome, nested case-control sampling may be preferred because the well-developed methodology of case-control studies may be directly applied to the collection and analysis of the data.

References

1. Cox DR. Regression models and life tables (with discussion). J R Stat Soc B 1972;34:187–220.
2. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York: John Wiley & Sons, 1980.
3. Breslow NE, Lubin JH, Marek P, et al. Multiplicative models and cohort analysis. J Am Stat Assoc 1983;78:1–12.
4. Frome EL. The analysis of rates using Poisson regression models. Biometrics 1983;39:665–74.
5. Mantel N. Synthetic retrospective studies and related topics. Biometrics 1973;29:479–86.
6. Liddel FDK, McDonald JC, Thomas DC. Methods for cohort analysis: appraisal by application to asbestos mining (with discussion). J R Stat Soc A 1977;140:469–90.
7. Prentice RL, Self SG, Mason MW. Design options for sampling with a cohort. In: Prentice RL, Moolgavkar SH, eds. Modern statistical methods in chronic disease epidemiology. New York: John Wiley & Sons, 1986:50–62.
8. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 1986;73:1–11.
9. Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. Ann Stat 1988,16:64–81.
10. Wacholder S, Boivin J-F. External comparisons with the case-cohort design. Am J Epidemiol 1987;126:1198–1209.
11. Andersen PK, Borch-Johnsen K, Deckert T, et al. A Cox regression model for the relative mortality and its application to diabetes mellitus survival data. Biometrics 1985;41:921–32.
12. Breslow NE, Langholz B. Nonparametric estimation of relative morality functions. J Chronic Dis 1987;40(suppl 2):89S–99S.
13. Yeh F-S, Yu MC, Mo C-C, et al. Hepatitis B virus, aflatoxins, and hepatocellular carcinoma in southern Guanzxi, China. Cancer Res 1989;49:2506–9.
14. Wacholder S, Gail MH, Pee D, et al. Alternative variance and efficiency calculations for the case-cohort design. Biometrika 1989;76:117–23.

Appendix 1

We compute the expected number of distinct persons from nested case-control sampling for the idealized intervention trial when $\beta = 0$. Let: n = the size of the cohort; p = the probability of failure; I = the number of failures in the cohort; s = the number of sampled controls per risk set; D = the number of distinct persons in a nested case-control study; and C_k = the set of indices for controls sampled in risk set k, where $k = 1, \ldots, I$ are in time order.

$$E[D] = n \times \{1 - \Pr(\text{person } j \text{ is neither a case nor picked as a control})\}$$

$$= n \times \{1 - \sum_{i=0}^{n} \Pr(j \text{ is neither a case nor picked as a control} \mid I = i) \times \Pr(I = i)\}. \quad (1)$$

Now $\Pr(j$ is neither picked as a control nor is a case $\mid I = i)$

$$= \Pr(j \text{ is not a sampled control in any risk set} \mid j \text{ is not a case}, I = i) \times \Pr(j \text{ is not a case} \mid I = i)$$

$$= \Pr(j \notin C_1, j \notin C_2, \ldots j \notin C_i \mid j \text{ is not a case}, I = i) \times \Pr(j \text{ is not a case} \mid I = i)$$

$$= \Pr(j \notin C_i \mid j \text{ is not a case}, I = i) \times \ldots \times \Pr(j \notin C_i \mid j \text{ is not a case}, I = i) \times \Pr(j \text{ is not a case} \mid I = i) \quad (2)$$

since each risk set is sampled independently. For the idealized intervention trial, there are $n - k$ eligible controls in the kth risk set, so that

$$\Pr(j \notin C_k \mid j \text{ is not a case}, I = i) = (n - k - s)/(n - k).$$

Further,

$$\Pr(j \text{ is not a case} \mid I = i) = (n - i)/n.$$

Thus, expression 2 equals

$$\prod_{k=1}^{i} (n-k-s)/(n-k) \times (n-i)/N = (n-1-s)!/(n-i-s-1)! \times [n!/(n-i)!]^{-1}.$$

Now, with $\beta = 0$, I has a binomial distribution with parameters p and n so that the expectation of the last quantity is given by

$$\sum_{i=0}^{n} \{(n-1-s)!/(n-i-s-1)! \times [n!/(n-i)!]^{-1}\} \times n!/i!(n-i)! \times p^i(1-p)^{n-i}$$

$$= (1-p)^{s+1} \sum_{i=0}^{s-1-s} \binom{n-1-s}{i} p^i(1-p)^{s-1-s} + \sum_{i=s-s}^{n} \text{some very small quantities} \approx (1-p)^{s+1}.$$

The approximation is needed only to acknowledge the technical problem that if there are enough failures, the last few risk sets will have fewer than s controls from which to sample.

Thus, from expression 1, the expected number of distinct persons in a nested case-control sample is, to close approximation

$$n[1-(1-p)^{s+1}]. \tag{3}$$

APPENDIX 2

We calculate the proportion of a cohort to be sampled in order to obtain the same expected number of distinct persons as in nested case-control sampling. In addition to the quantities defined in Appendix 1, let α = the proportion of persons samples from the cohort to obtain the subcohort, and T = the number of persons in a case-cohort sample (subcohort and additional cases). Now the expected number of (distinct) persons for case-cohort sampling given i failures is

$$E[T \mid I = i] = \alpha n + (n - \alpha n) \times i/n$$

so that

$$E[T] = \alpha n + (1-\alpha)np$$

which implies that

$$\alpha = (E[T] - np)/n(1-p). \tag{4}$$

So, choosing subcohorts of size $(E[D] - np)/(1-p)$ will yield the same expected number of distinct persons as the nested case-control sampling.

We may also use the above formula to obtain the exact α needed to obtain subcohorts with the same number of distinct persons as nested case-control samples for the idealized intervention trial when $\beta = 0$. Setting $E[T]$ in expression 4 equal to expression 3, yields

$$\alpha = 1 - (1-p)^s$$

so that the subcohort size must be

$$n\alpha = n[1-(1-p)^s].$$