REVENTIVE MEDICINE 23, 587–590 (1994)

Nested Case–Control Studies[1]

VIRGINIA L. ERNSTER, PH.D.

Department of Epidemiology and Biostatistics, School of Medicine, University of California, San Francisco, San Francisco, California 94143-0560

The nested case–control study design (or the case–control in a cohort study) is described here and compared with other designs, including the classic case–control and cohort studies and the case–cohort study. In the nested case–control study, cases of a disease that occur in a defined cohort are identified and, for each, a specified number of matched controls is selected from among those in the cohort who have not developed the disease by the time of disease occurrence in the case. For many research questions, the nested case–control design potentially offers impressive reductions in costs and efforts of data collection and analysis compared with the full cohort approach, with relatively minor loss in statistical efficiency. The nested case–control design is particularly advantageous for studies of biologic precursors of disease. To advance its prevention research agenda, NIH might be encouraged to maintain a registry of new and existing cohorts, with an inventory of data collected for each; to foster the development of specimen banks; and to serve as a clearinghouse for information about optimal storage conditions for various types of specimens.

INTRODUCTION

Most of us were trained to think that there are basically two kinds of observational analytic study designs in epidemiology: the cohort study and the case–control study. During the past 15 to 20 years, we have seen the emergence of variants of these designs, including the nested case–control study (which is alternatively called the case–control in a cohort design) and the case–cohort study. Although the selection of the comparison group and the analytic methods for each of these two are different, both incorporate features of the classic cohort and case–control study designs and are essentially hybrids of those designs. Although the focus here is on the nested case–control study, I will comment briefly on the case–cohort approach and how the two designs compare with each other. This discussion is largely limited to issues of study design and strengths and limitations relative to other designs and does not consider issues of statistical analysis. There is now considerable literature comparing the statistical properties of nested case–control studies with full cohort as well as case–cohort studies, including simulation studies (1–6). In general, the relatively minor loss in statistical efficiency of the hybrid designs compared with the full cohort study is offset by considerable reductions in the number of study subjects and in the associated time and costs of data collection and/or analysis.

[1] Presented at the Symposium "Disease Prevention Research at NIH: An Agenda for All," October 6–8, 1993, Bethesda, MD.

STUDY DESIGN

The Case–Control Study

In the conventional case–control study, a group of individuals with the disease of interest (cases) is selected and a group of individuals without the disease (controls) is also selected. For example, if an investigator is interested in the relationship between serum cholesterol level and risk for coronary heart disease (CHD), he or she might identify all new cases of CHD in a community over a defined period as well as a group of individuals without CHD from the same community and then measure their serum cholesterol levels or perhaps attempt to determine what their cholesterol levels were prior to disease occurrence through a review of medical records or patient interviews. In the simplest analysis, cases and controls would be dichotomized into high or low cholesterol categories and the estimated relative risk of CHD determined. Although the ratio of controls to cases varies widely across studies, it is most commonly 1:1 or 2:1. A concern with this study design is that serum cholesterol levels measured at the time of the disease may be a result rather than a cause of the disease, while those obtained from review of medical records may not be available for all subjects nor have been analyzed in a standardized fashion, and those obtained by interview may also be subject to problems of accuracy or recall bias. There is a vast literature on case–control studies which concerns, among other things, the importance of selecting controls from the same underlying population as the cases and of taking potential confounding variables into account through such strategies as matching or multivariate analysis (7, 8).

0091-7435/94 $6.00

Cohort Study

In the classic cohort study, one begins with a group of individuals free of disease, for whom baseline data are collected, and then follows them over time to determine who does and who does not develop the disease—or various diseases—of interest. One might, for example, measure baseline cholesterol levels, or at least collect blood specimens, for a cohort of thousands of individuals, enabling the eventual determination of rates of developing CHD during the follow-up period by baseline cholesterol level. This design has the advantage over the case–control study of collecting exposure data before the disease occurs, an appropriate time sequence for a cause–effect relationship; there is also less concern about the accuracy of the exposure data, and individuals who develop disease come from the same population (cohort) as those who do not. Incidence rates can be calculated directly from cohort data (unlike case–control data), facilitating external comparisons. On the other hand, it is very costly, time-consuming, and relatively inefficient to measure serum cholesterol on so many individuals who do not end up developing the disease of interest.

The Nested Case–Control Study

In the nested case–control study, one also begins with a defined cohort and identifies cases that have already occurred (in the case of a retrospective nested case–control study) or as they occur (in the case of a prospective nested case–control study). Then, for each case, a specified number of controls is selected from among those in the cohort who have not developed the disease by the time of disease occurrence in the case. The number of controls selected per case may vary, but it is common in the nested case–control literature to find four or five controls per case. Several additional points to note are that time-matching is an essential feature of this design, whether controls are matched to cases on age, date of entry into the cohort, length of time in the cohort, or a combination of these measures (9); that a cohort member who serves as a control at one point in time may later become a case; and that a cohort member may be selected as a control for more than one case (10).

Several examples of nested case–control studies are instructive. One investigated the relation of serum cholesterol and large bowel cancer (11). The cohort in which the study was nested consisted of 48,314 Kaiser Permanente Medical Care Program members who had serum cholesterol data obtained at multiphasic examinations and were then followed a mean of 7.2 years (more than 348,000 person years) for the occurrence of colorectal cancer (Fig. 1). The 245 members of the cohort who developed colorectal cancer were the cases; for each case, five controls were selected from the cohort who were matched to the case on age, race, sex, and time of multiphasic examination. Instead of having to retrieve and analyze the serum cholesterol and other data for all of the cohort members, the investigators were able to confine their study efforts to the 245 cases and their matched controls (N = 1,225), a much smaller and logistically more feasible sample size. The authors compared mean serum cholesterol levels in the cases and controls and calculated odds ratios for colorectal cancer by quartile of serum cholesterol level.

Another nested case–control study examined the relationship between serum organochlorines and breast cancer (12). Study subjects were drawn from a cohort of over 57,000 female members of Kaiser Permanente Medical Care program who underwent multiphasic examinations in the late 1960s, at which time blood samples were collected and stored. The cohort was followed through 1990. One hundred and fifty women who developed breast cancer during the follow-up period were then randomly selected and individually matched to 150 women in the cohort who had remained free of breast cancer. Using the blood stored at baseline, serum levels of DDE and PCBs were compared between cases and controls. This approach resulted in considerably less cost and effort in the processing of laboratory specimens, as well as savings in the retrieval and analysis of other epidemiologic data, than had a full-blown cohort analysis been undertaken.

A third nested case–control study investigated

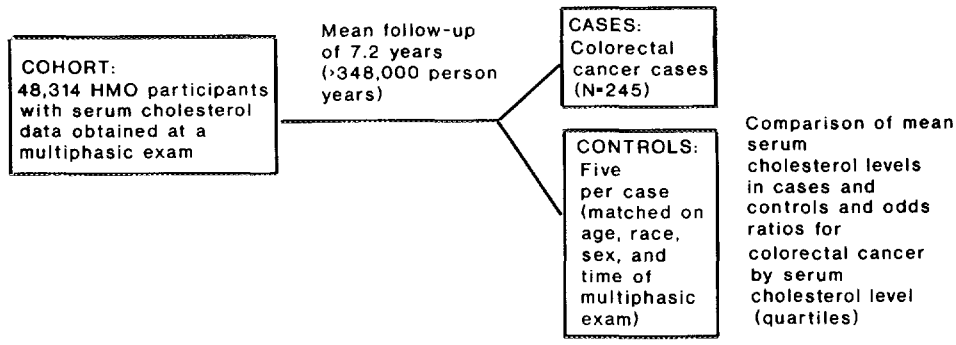

FIG. 1. Example of a nested case–control study (11) on serum cholesterol and large bowel cancer.

whether aspirin use was associated with reduced risk for colon cancer mortality (13). It was based on the American Cancer Society's prospective cohort study which enrolled 1.2 million adults in 1982, at which time baseline data, including frequency and duration of aspirin use, were collected. The paper reported colon cancer death rates by amount of aspirin use during a 6-year follow-up period for all 662,424 whites in the cohort who had completed the relevant questionnaire items. However, for multivariate analyses—which controlled for first-degree family history of colon cancer, body mass index, physical activity, and diet, the authors performed a nested case–control study. That component involved only 598 individuals who had died of colon cancer during the follow-up period and five age- and sex-matched controls per case. In other words, rather than analyzing data for over 660,000 individuals, the investigators analyzed data for about 3,600 individuals.

There are a number of advantages to the nested case–control design. First, the controls are from the same population as the cases; although, as Wacholder has observed, in theory *every* case–control study takes place within a cohort, it is often difficult to characterize the cohort (14). Second, as is clear from the examples, owing to the smaller number of study subjects than in full cohort studies, nested case–control studies are less expensive, and the collection and analysis of data are less time-consuming—which is especially important when collection of detailed data (e.g., occupational chemical exposure histories) or performing extensive laboratory tests for an entire cohort would not be feasible. Third, data on exposure are more likely to have been collected prior to diagnosis of disease than in the conventional case–control study, consistent with a cause–effect interpretation if an association is found, and recall bias is not an issue for those exposures. The nested design is especially valuable for studies for biologic precursors of disease, such as serum cholesterol, somatic mutations, and so on—not only because such tests may be costly or labor-intensive but because it is essential that predisease states be determined. However, all exposure data of interest are not necessarily collected before disease occurrence; in an occupational nested case–control study, for example, one might have excellent chemical exposure data collected prior to disease but still need to collect good smoking history data once cases and controls have been identified. Until recently, an additional advantage of the nested case–control design over the case–cohort design, which is described briefly below, was that multivariate analysis using standard techniques and available software for conventional matched case–control data was more easily accomplished.

It obviously makes the most sense to undertake a nested case–control study when a cohort exists that is appropriate for the research question and when there are real efficiencies in cost and effort to be achieved by analyzing a subset of the data, offsetting any loss in statistical power.

The Case–Cohort Study

Like the nested case–control study, the case–cohort study also takes place within a cohort. All cases of a disease of interest occurring in the cohort are included as well as a random sample (or a stratified random sample) of the entire cohort. The latter group is often referred to as the subcohort. Unlike the nested case–control design, cases in the case–cohort study are not matched to individuals in the comparison group on time or other variables. An example of a study of this design addressed the role of selenium in human breast cancer; blood samples were collected from 5,162 healthy women on the island of Guernsey, who were then followed for the development of breast cancer (15). Laboratory analysis of baseline selenium levels were performed only for the women who ultimately developed breast cancer ($N = 46$) and for the 138 noncases randomly selected from the total cohort.

The case–cohort study is essentially a variant of the nested case–control study—or an unmatched case–control study within a cohort (4). Of the two, it is the preferred design for calculating incidence rates and standardized mortality ratios and for making external comparisons. It has several other advantages: study subjects in the subcohort can be selected as soon as they are deemed eligible for the cohort, rather than having to wait until a case develops so that matching can take place; the same subcohort can be utilized for multiple disease outcomes; for intervention trials, compliance, changes in biologic parameters, and other measures can be monitored in the subcohort throughout the study; and the subcohort can even serve as a source of controls for nested case–control studies (3, 9, 10). Since only a fraction of the total cohort is studied, this design shares with the nested case–control study the advantages of economy of cost and effort; one comparison of a case–cohort analysis with a full cohort analysis using data from a cohort study designed to examine the association between treatments for Hodgkin's disease and second cancer occurrence (with a ratio of one case for every four subcohort members) reported a five-sixths saving in cost of data collection for only an 11% loss in statistical efficiency (4).

Several years ago the results of a Swedish study of the risk of breast cancer following hormone replacement therapy were published based on a cohort of 23,244 women identified from population-based medical records who had been prescribed menopausal estrogens (16). During the follow-up period (mean of 5.7 years), 253 breast cancer cases developed in the cohort. The relative risk of the disease in the cohort of women prescribed estrogens was compared with that in the rest of the female population, a classic full cohort anal-

ysis. To address the risk of breast cancer by type and duration of treatment, which required collecting more detailed data, the investigators employed a case–cohort approach, in which all of the cases were compared with a subcohort of the total cohort, selected with a 1 in 30 sampling ratio. Finally, to further address questions of dose–response while adjusting for potential confounders, the investigators employed a nested case–control approach, selecting for each case up to five controls from the subcohort matched to the case on birth and year of inclusion in the cohort.

CONCLUSION/SUMMARY

This brief review has focused on some of the key aspects of the design of nested case–control versus other types of analytic epidemiologic studies. It has not dealt with the relative statistical efficiency of the various designs, which depends on such things as numbers of controls per case or subcohort size, disease frequency, exposure frequency, relative risk, and whether more than one disease is to be studied (2, 3, 4, 14). As always, the appropriateness of a particular study design is largely dependent on the research question. Selecting a nested case–control or case–cohort design to examine variables that are easy and inexpensive to collect and analyze for the full cohort may result in little more than loss of statistical efficiency. However, these designs have many attractive features and deserve greater attention, especially now that software is available to select matched controls from large cohorts (17) and that techniques are evolving to appropriately analyze case–cohort data (9).

To take advantage of these newer study designs in advancing its prevention research agenda, NIH might be encouraged to maintain an up-to-date registry of cohorts (including those that are just being set up), with an inventory of data and specimens available from each, which could serve as a resource for addressing research questions in addition to those intended by the original investigators; foster the development of banks for such specimens as serum, tissue, and DNA; and serve as a national clearinghouse for information about storage facilities and optimal storage conditions for various types of specimens (18).

Received January 27, 1994
Revision requested May 23, 1994
Accepted May 23, 1994

REFERENCES

1. Kupper LL, McMichael AJ, Spirtas R. A hybrid epidemiologic study design useful in estimating relative risk. *J Am Stat Assoc* 1975; 351:524–528.
2. Langholz B, Thomas DC. Nested case–control and case–cohort methods of sampling from a cohort: A critical comparison. *Am J Epidemiol* 1990; 131:169–176.
3. Prentice RL. A case–cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; 73:1–11.
4. Wacholder S, Boivin J-F. External comparisons with the case–cohort design. *Am J Epidemiol* 1987; 126:1198–1209.
5. Lubin JH, Gail MH. Biased selection of controls for case–control analyses of cohort studies. *Biometrics* 1984; 40:63–75.
6. Miettinen O. Design options in epidemiologic research. *Scand J Work Environ Health* 1992; 8:7–14.
7. Schlesselman JJ. Case–Control Studies. New York: Oxford Univ. Press, 1982.
8. Kelsey JL, Thompson WD, Evans AS. Methods in Observational Epidemiology. New York: Oxford Univ. Press, 1986.
9. Wacholder S. Practical considerations in choosing between the case–cohort and nested case–control designs. *Epidemiology* 1991; 2:155–158.
10. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case–control studies. III. Design Options. *Am J Epidemiol* 1992; 135:1042–1050.
11. Sidney S, Friedman GD, Hiatt RA. Serum cholesterol and large bowel cancer. *Am J Epidemiol* 1986; 124:33–38.
12. Krieger N, Wolff MS, Hiatt RA, Rivera M, Vogelman J, Orentreich N. Breast cancer and serum organochlorines. *JNCI* 1994; 86:589–599.
13. Thun MJ, Namboodiri MM, Heath CW Jr. Aspirin use and reduced risk of fatal colon cancer. *N Engl J Med* 1991; 325:1593–1596.
14. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case–control studies. I. Principles. *Am J Epidemiol* 1992; 135:1019–1028.
15. Overvad K, Wang DY, Osen J, Allen DS, Thorling EB, Bulbrook RD, Hayward JL. Selenium in human mammary carcinogenesis: A case–cohort study. *Eur J Cancer* 1991; 27:900–902.
16. Bergkvist L, Adami H-O, Persson I, Hoover R, Schairer C. The risk of breast cancer after estrogen and estrogen-progestin replacement. *N Engl J Med* 1989; 321:293–297.
17. Beaumont JJ, Steenland K, Minton A, Meyer S. A computer program for incidence density sampling of controls in case–control studies nested within occupational cohort studies. *Am J Epidemiol* 1989; 129:212–219.
18. Petrakis NL. Biologic banking in cohort studies, with special reference to blood. *Natl Cancer Inst Monogr* 1985; 67:193—198.