Parallels, Convergences, and Departures in Case-Control Studies and Clinical Trials

Marvin A. Schneiderman and David L. Levin

National Cancer Institute, NIH, Bethesda, Maryland 20014

Summary

In studies of the etiology of disease one usually must rely on nature’s experiments, observing populations both with and without a disease, then determining how these populations differ with respect to characteristics that some people think could be related to the disease. These unplanned natural experiments can provide meaningful information, or they can mislead. The key to unlocking the useful information the experiments may contain lies in the intelligent, informed choice of “controls,” the nondiseased individuals, for comparison with diseased cases. Some principles for the choice of controls are given in this paper. We show how some of the principles and problems of the controlled trial parallel the principles and problems of epidemiological studies. Examples of matching and cautions concerning possible overmatching are given. Relationships are drawn with controlled and uncontrolled clinical trials and between prospective and retrospective comparisons. The two major principles that can be laid down are: (a) know the subject matter, but do not choose controls in a given way merely because other people have done it that way before; (b) know the specific question(s) the study is designed to answer. This helps avoid overmatching.

Study Descriptions

Case-control studies are performed to answer questions about the effect of a known variable (e.g., something specific in the environment or past history) on the occurrence (cause?) of a disease. They contrast with clinical trials which are concerned with the course of the disease or, more specifically, the effects of intervention (i.e., treatment) on the course of the disease. In a case-control study, we start with the existence of the disease and look backwards in time to discover differences between cases and controls which we think may relate to the appearance of the disease. We use this method to help us to answer questions such as the following. (a) Is smoking related to excess development of lung cancer? (b) Is age at 1st pregnancy related to subsequent development of breast cancer? (c) Is diet (i.e., saturated fat levels) related to death from arteriosclerosis? (d) Does a specific kind of viral infection predispose to development of cervical cancer?

In case-control studies, we do not assign persons to specific groups. Nature assigns them for us, having already done the experiment for us. Cases are cases because the persons already have the disease: controls are controls primarily because they do not. In some sense, the people we study are self-selected. In a prospective (follow-up) study, we also look at persons who are self-selected or previously selected (smokers versus nonsmokers; men versus women) and observe them for periods of time to see how each group behaves with respect to development of a disease. In a retrospective clinical study, we compare patients treated one way with patients treated another way. We hope that the differences we see in patients after their treatment are due to the treatment. We can only rarely be sure this is true.

In a controlled, prospective clinical trial, we select those who will receive the study treatment and those who will receive some contrasting treatment(s). Usually, we assign patients at random to one of several treatment schedules and then observe the result. Clinical trials help us to answer questions such as (a) what is the effect of postoperative radiation on survival following mastectomy? (b) does the use of hexachlorophene soaps reduce the incidence of infection in newborn nurseries? and (c) are Papanicolaou smears that are administered to a well population effective in reducing mortality of cervical cancer?

In a case-control study, we usually have cases with a given disease on hand and must then find our controls, being careful (and this is critical) that the population from which we select our controls indeed reflects the group about which we wish to draw inferences when our study is over. Then retrospectively compare our 2 groups with respect to some limited list of characteristics we think important.
This involves looking through records to see how patients have responded in the past. From these historical data, we can do a cooperative trial in which each physician would be encouraged to report his results. The local herb and poltergeist theories are not indicated in treating this infection. The fact that the experiment was not done prospectively didn't really mean that it was automatically bad. Uncontrolled trials can be prospective, but that usually doesn't help. For example, if we are cautious, we create a clinical impression. Clinical experience is an excellent thing. It is a fine way of formulating hypotheses to be tested. Often the thoughts a doctor has about the cause of a disease are correct. However, some people do not treat these data as clinical impressions but rather they treat them as sound, solid things, to which statistical significance tests are added as the final word in "demonstrating" something. What we often cannot tell is whether the patients on one treatment did better because the treatment was better or because those patients were better to begin with. Detailed examination to "equate" patients becomes the equivalent of careful choosing of controls in a case-control study.

### Study Selection

Why do (or not do) a case-control study? In a prospective study of cancer or other diseases of low incidence, it is necessary to observe a large population to get enough cases on which to draw valid conclusions. Cancer has a long latent period. Having identified a possible cause, we would have to wait for development of the disease. This could take a long time. In a case-control study, we start by identifying known cases. This allows us sometimes to study a long time-period in only a short time.

The case-control study also allows us to study many variables at the same time. In studying cervical cancer, we can ask about age at 1st intercourse, number of sexual partners, number of pregnancies, urban-rural status, etc. If we were doing a prospective study, we would probably have to segregate our populations on the basis of 1 characteristic (or, at most, very few characteristics), observe the people for disease development, and then probably look at other characteristics retrospectively. A prospective study is often prospective in only a few characteristics.

Since we want to be able to say that differences between cases and controls are disease related, we want our controls to be more similar to our cases as possible, but to differ with respect to the disease. Investigators usually match cases and controls on demographic variables such as age, race, and sex, or on some other factor such as hospital or residence. Since we cannot perform analysis on matched variables, we must select our matching criteria carefully. If we have matched on race, we are most unlikely to detect
etiological factors closely associated with race. If race is
highly correlated with other variables (for example, socio-
economic factors), we may also lose the effects of these
other variables. Statistical aspects of matching are dis-
cussed in detail elsewhere (2, 3, 9–11, 15). We present
other theoretical considerations of these types of studies.

What are some of the problems of case-control studies
and how can we avoid them? Of primary importance is the
representativeness of our study, both of cases and of con-
trols. How do the patients studied compare to the next
patient being treated? To patients in general? To the pop-
ulation from which the patients came? We can increase the
 comparability of our patient population to a “standard”
population by also examining controls selected from our
reference population and by looking for differences be-
tween cases and controls. We can increase the sensitivity
of our comparison by matching cases to controls for vari-
ables we feel are important for a particular disease. Some
of these matchings are very obvious. If we match by sex
in breast cancer, we assure ourselves that we won’t hap-
pen to have mostly female cases and mostly male controls.

It is not necessary, and often not desirable, to find a
matching control for every case on a 1-to-1 basis. We
may wish to compare our patient population with various
groups by selecting multiple controls from populations of
other patients with related disease, other patients with un-
related diseases, a healthy population, and/or the general
population. Which group or groups we select as our con-
trols will affect our ability to generalize our conclusions
with regard to various other classes of persons. If con-
trasts with several different control groups all lead to the
same conclusion, we can have more confidence that the
conclusion is really correct.

It is possible to overmatch. If we stratify on several
variables, (e.g., 2 races, 2 sexes, 5 age groups, and 3 hos-
pitals) we may end up with many strata (2 \times 2 \times 5 \times 3 =
60). It then becomes difficult to find a control for each
case. One solution is to match not by patient but by group.
Thus, if we have 4 cases in 1 group, say white males age
60 at Hospital No. 3, it is not necessary to find 4 controls.
Perhaps we will have 1 or 2 or 10. Mantel and Haenszel
(9) have examined the benefits of using group controls
and have developed a very effective computing procedure.
Using the Mantel-Haenszel procedure, we can control for
the effect of several variables without having to match
each case with a single control.

In order to choose the variables on which to match, we
must be aware of the specific characteristics of the disease
that we feel are most important. Feinstein (5), writing on
disease classification, states “By studying the evolution of
the various clinical forms of a disease, clinicians can
select those clinical entities that are distinctive enough to
warrant separate categorical designations and those that
may be clustered together in a combined category. En-
tities that are of no prognostic or therapeutic significance
can often be ignored; entities that are clinically important
can be specifically noted and labeled.” In an analogous
manner, the epidemiologist must make use of information
about the particular disease and patient population under
study in order to select those entities on which to match
and those that may be clustered together in a combined
stratification category.

Representativeness for one trait does not imply repre-
sentativeness for another. In the American Cancer
Society’s prospective study on the relationship between
smoking and lung cancer (6), the men studied were not
representative of all men in the United States. The group
contained a preponderance of white men, persons with
high incomes, and white-collar workers. The study showed
that more smokers than nonsmokers develop lung can-
cer. Because it was felt that the sample studied was repre-
sentative in terms of the effects of smoking, the American
Cancer Society could validly generalize their findings to all
men and eventually to women, too. It is, of course, not
necessary that the samples be representative with regard to
all possible characteristics, which is likely to be impossible
to achieve, but only with regard to those of consequence. If
we cannot eliminate selection bias from our study, we must
evaluate how it would affect our results.

Can we set up a “universal” control and use it for several
different studies? This technique can be useful, and we
unconsciously use it whenever we compare something to
“past history.” However, we must be alert to changes with
time which could render our control obsolete. The mortality
rate for cervical cancer is falling. If we compare current
cases with previously defined controls, we may draw erro-
neous conclusions. On the other hand, we can use changes
with time to our advantage. By comparing present cases
with earlier cases, we can gain information about trends;
e.g., if cervical cancer rates are falling, coincident with
urbanization of the population, we might tend to consider
that urbanization was not an important etiological factor
in cervical cancer.

Retrospective studies impose “cart-horse” limitations
on us. We wish to know the state of the patient at the time
his disease was initiated or earlier, but we usually do not
see him until well after the disease process has begun. Thus,
when we identify a case and then perform tests on that pa-
tient, we may find things that are a result rather than a
cause of the disease. One problem with studying the possi-
ble viral etiology of cancer is that, by the time the patient
develops clinical symptoms, there may be no residual evi-
dence of infection or the viral infection may have come after
the appearance of the disease, with perhaps the cancer
tissue providing a more fertile ground for viral growth than
does normal tissue.

The long latent period of cancer leads to still other prob-
lems. We may believe we are comparing diseased to well
persons, but if one of our controls later develops cancer, we
may have been making a case-to-case comparison. These
kinds of errors will dilute any real effects that exist. We can
partially avoid the problem by group matching, but there
is really no way to completely avoid this problem.

Examples

It is worthwhile to see how these procedures operate in
a few actual or proposed studies. We will refer to both some
effective case-control techniques and to some situations in
which the procedures leave us uneasy.
Rosenthal et al. (12) recently reported an association between BCG\(^4\) vaccination in infants and reduced leukemia mortality among children (in Chicago). In their study, they identified (through records of the Chicago Board of Health) 22 leukemia deaths among black children under the age of 7 who died between 1964 and 1969. They compared the names of those children with a list of 54,414 infants reported to have received BCG at Cook County Hospital in the time interval, 1958 to 1969. They found 1 name on both lists. These data yield a mortality rate of 0.31 per 100,000 vaccinated children per year. To determine a rate for nonvaccinated children, they estimated the total population of black children under the age of 7 from the 1970 census and subtracted the number vaccinated. This they assume to be the number of children not vaccinated during this time. The resultant figure (172,986) led to the projection of an annual mortality rate of 2.02 per 100,000 “nonvaccinated” children per year. What are some possible problems with this method of choosing controls? (There are also possible errors of ascertainment, but we will not discuss them here.)

The birth rate has been declining for some time, so that the 1970 census could easily be an understatement of the number of children born between 1958 and 1969 (who could then have been 0 to 6 years old at any time from 1964 to 1969). A more appropriate control might have been all the unvaccinated black children born in Chicago during the same years. A quick review of birth data for the years 1958 to 1969 gave us about 360,000 “nonwhite” births, about one-third more than the number Rosenthal used.

Since the “control” population was not specifically identified, it is not possible to know how many (if any) of them received BCG elsewhere. Thus the control group might actually include some BCG-vaccinated cases. We also do not know how many of the infants vaccinated or unvaccinated left the Chicago area or died of other causes and thus were not “available” as possible leukemia deaths, or how many other children not born in Chicago migrated to Chicago during the years studied. Knowledge of the place of birth of all 22 cases would, of course, be essential. Finally, since leukemia is a disease related to social class, one would also have to attempt some matching by social class.

Much of the report is based on city and hospital records. A test is available to the authors. They could have tried class.

The Abbreviation used is: BCG, Bacillus Calmette-Guérin.

Of leukemia was more likely due to lack of clustering than to insensitivity of their test procedure.

In a proposed search for etiological factors in lung cancer among Chinese, the International Agency for Research on Cancer plans to use 2 sets of controls. The 1st group, used to test the association between smoking and cancer, will be matched by sex, age, and dialect since, for the population under study, “smoking is believed to vary by dialect group, at least in women, and failure to match could result in a spurious association between smoking and lung cancer in women.” Thus, by controlling for the effect of dialect group, the study will be able to eliminate any confounding effect of the variable on the effect of smoking. The 2nd control group, used to test for factors other than smoking, will be matched by age, sex, and smoking history.

The use of dual controls has 2 advantages. Since you cannot test on a matched variable, the use of 2 controls allows testing in one control on the variable which was matched in the other. In addition, information will be provided on the interaction of factors. If the multiple controls in this study lead to the same conclusion, one can have more confidence in the result.

Adam et al. (1) described a computer-based algorithm for matching controls and patients in a study of herpesvirus and cervical cancer. This is an example of an attempt at close matching. The algorithm is applied to 2 sets of cases. In the 1st set, 27 cases were identified, and a potential control pool of 193 women was selected, with no attempt at matching. Each of the 220 women was classified according to age (at onset of cancer), race, socioeconomic level, age at 1st intercourse, age at 1st pregnancy, and number of live births. The algorithm was applied to determine a “match score” for all case-control pairs (27 × 193 = 5211). Allowable limits were placed on the scores. The cases were arranged in random order. The best match for the 1st case was found, and that control was eliminated from the control pool. The best match for the 2nd case was then found, and so on, until each case was matched. Analysis was performed on the 27 pairs, using only 54 persons out of a total 220 interviewed (less than 25%). The difficulties in matching on many categories are seen; when they matched by all 6 of the rated characteristics at 1 time, controls could not be found for 5 cases, and only 22 pairs could be compared. Much potential information was lost. Perhaps collapsing 1 or 2 categories and then matching by group instead of on a 1-to-1 basis would have allowed the authors to use more of the available data.

The 2nd set studied was similar, except that controls were partially prematched to assure better final matching. Here 42 cases and 190 potential controls were interviewed. Again a 1-to-1 control was used so that only 36% (84 of 232) of the available information was used. As a further minor criticism, even if one felt strongly about 1-to-1 matching, perhaps a better method would have been to pick the best case-control match among all cases, then the 2nd best, etc.
The Adam paper has been criticized as a possible example of overmatching (8). Since the algorithm matched on attributes highly correlated with virus infection, it may have had the effect of matching on the variable under study itself and thus might have missed the effect of viral infection on the etiology of cervical cancer.

These 2 studies by Dr. Adam used previously collected data as a source of matching controls. Is this satisfactory? If so, are we opening the door to “historical controls,” which many writers (7, 13) have warned against? Is this argument relative to “data banks”? It depends on the nature of the persons in the pool. For example, if pooled controls came largely from New York hospitals such as the Bellevue Medical or Flower & Fifth (2 hospitals with few Jewish patients), whereas diseased cases came from Memorial Hospital (a cancer center with a large Jewish population), we could find a spurious association between religious groups and cancer.

Smith and Pike (P. G. Smith and M. C. Pike. Space-Time Clustering: A Case-Control Method of Examining Diseases with Long Latent Periods, in preparation) have recently suggested a case-control method for studying space-time clusters of Hodgkin’s disease and other rare diseases with postulated long latent periods and possible spread by contagious mechanisms.

A list is compiled of all patients with Hodgkin’s disease in a defined geographical area during a specified time period. For each patient, one or more controls are chosen matched for age, sex, socioeconomic status, and area of residence. The names of the controls are added to the list of cases. The amount of contact between all persons on the list is calculated, and the resultant measure of contact is analyzed. If the disease is indeed contagious, patients should have had more contact with other patients than with controls.

There are many technical problems to such an approach (e.g., defining a measure of contact for all combinations of pairs of persons on the list), but the method does present a novel use of case-control methodology. The determination of “suitable controls” is carefully considered: Smith and Pike remark “How closely should one match for place of residence? Nearest neighbour controls might be appropriate but this is probably ‘overmatching’ since neighbours may tend to have the same contacts. If however we reject neighbours as controls and select a person, chosen at random, living in the study area, we will also be testing for mere spacial clustering, since, if there was clustering of the patients’ homes then the patients would tend to have more contact with each other, because they lived closer to each other, than the controls who were selected from a wider area. Thus the test may yield a ‘false’ positive result. Of course, spacial clustering of cases may be due to a contagious process but might also depend on purely environmental factors and the statistical test would not distinguish between these two.”

Vaillant et al. (14) reported a study of psychological vulnerabilities of physicians. Theirs was a retrospective-prospective study. For the past 30 years, a population of 268 men (college sophomores at the start of the study) has been followed. From this population, 47 physicians were identified as “cases” and compared with a control group of nonphysicians selected randomly from the population. No matching was performed 30 years ago. This allowed Vaillant and his coworkers to look into the effect of variables in childhood developmental and medical history on the present psychological state of the men. This is a case in which lack of knowledge forced an “open” approach to a problem which permitted later comparison on certain important measures. Usually, we are not so fortunate, and many “open” studies collapse because the items that are later seen to be of possible importance were not recognized when the data were first collected.

Discussion

A few principles have been suggested for the selection of controls in epidemiological studies. By and large, the choice of controls hinges on special knowledge of the disease under study. One often uses this special knowledge when deciding whether it is worthwhile to seek “matches” on more and more items. Thus one has to temper knowledge with restraint. Attempting to match on too many variables will generally lead to unmatchable cases. One-to-one matching has several built-in diseconomies, and it may be better to reduce categories and match by groups, allowing the exploitation of more data.

There is no situation in epidemiological research which compares strictly with the controlled, randomized clinical trial. In diseases with a long time span from exposure to onset, prospective studies are almost prohibitively difficult. It may be possible to do retrospective case-control studies that will yield valid answers. Prospective of itself is not a universal virtue. Good clinical observation often leads to good follow up ideas. Clinical observation of itself, however, almost always needs to be tested, verified, and confirmed.

We have several times mentioned “representativeness” of cases and controls. By this we mean to imply representativeness for items under study. Thus, if we show a relationship between exposure to UV in skin cancer in Australian men and attempt to generalize this finding to all men and women, we are immediately assailed by problems of representativeness. Are we safe in saying something about men and women? We are if we think men are representative of both men and women in the matter of how skin reacts to UV exposures. And we are safe in generalizing from fair-skinned Australians to dark-skinned Africans? Only if we are prepared to show evidence that reaction to UV is independent of skin pigment.

Thus we come full circle to an equivocal answer. Controls are satisfactory if they are “representative.” Of what? Only our subject-matter knowledge tells us what is important to control for (to seek representativeness for) and what is not important. The decisions can then never be completely beyond challenge. For what one man’s subject-matter knowledge may tell him is of no consequence, another man’s knowledge may tell him is important.
References

Parallels, Convergences, and Departures in Case-Control Studies and Clinical Trials

Marvin A. Schneiderman and David L. Levin