

METHODOLOGIC PROBLEMS AND STANDARDS IN CASE-CONTROL RESEARCH

ALVAN R. FEINSTEIN

IN SUCH scientific domains as physics, chemistry, and biology, rigorous standards have been developed for proving the claim that a particular effect occurs as a consequence of a particular cause. The investigator is asked to perform an experiment, imposing the alleged causal agent, and contrasting its effects with those that follow the imposition of a comparative agent, or 'control'.

In the domain of clinical medicine, these standards have been applied during diverse experiments conducted to clarify cause-effect relationships in human physiologic function during health or disease. Other kinds of rigorously designed experiments, called therapeutic trials, have been performed to evaluate the efficacy of different modes of clinical treatment. Accordingly, when a clinical investigator wants to extol the benefits of a particular treatment, the existing scientific standards require that the evidence come from an experimental therapeutic trial, with the treatment assigned by randomization; with the baseline state of subjects clearly identified; with the target or outcome variables clearly delineated; and with the observations performed under double-blind circumstances to prevent the occurrence of bias either when the outcome events are examined or when treatments are assigned.

In epidemiology, however, the particular kind of cause-effect investigation that is called a 'case-control study' is not conducted as an experiment and has not had any analogous scientific standards established for its performance or its interpretation. As long as the investigator is a 'licensed' epidemiologist, he can choose cases and controls arbitrarily; and he can obtain and manipulate the data in diverse ways that are sanctioned not by the delineated standards of science, but by the traditional practice of epidemiologists.

The purpose of this essay is to demonstrate some of the many hazards that occur in the absence of rigorous scientific principles for case-control research; and to suggest some specific standards that can be used to prevent or eliminate the problems.

Let us begin with a simple issue in nomenclature. The word *case-control* describes the structure of a type of investigation in which the prevalence of a focal entity, as found in a group of cases, is compared against the prevalence of that entity in a group of controls. The cases are usually chosen because they have a particular disease or clinical condition in which the investigator is interested. The controls are people who do not have that condition. At least three distinctively different kinds of research [1], however, all have this same case-control structure. In a *diagnostic marker case-control study*, the focal entity is a positive diagnostic test for the disease. Thus, the sensitivity of the CEA test might be determined for a group of cases with colon cancer, and its specificity is determined for a group of controls, without that cancer. In a *longitudinal prevalence case-control study*, the focal entity is a particular outcome effect or complication of a disease. For example, the prevalence of hypercholesterolemia can be contrasted in a group of cases, such as patients who have had diabetes mellitus for 20 yr, and in a group of controls, who may have had diabetes for 10 yr or less, or who may not be diabetic at all. In a *retrospective or etiologic case-control study*, the focal entity is antecedent exposure to a suspected etiologic agent. Thus, the prevalence of a previous usage of reserpine or estrogen can be compared in cases of women with breast cancer and in a control group.

Each of these three kinds of studies has the same case-control structure, but the studies differ in their purposes, in the temporal direction of the reasoning, and in their interpretation. In diagnostic-marker research, the time direction is concurrent; and the entity found in the diagnostic test coexists with the presence or the absence of the disease. In a longitudinal prevalence study, the time direction is forward and the disease is believed to cause the focal entity. In the third kind of study, the time direction is backward and the focal entity is suspected of having previously caused the disease.

The use of the same name, *case-control study*, for three different kinds of research, with three different purposes and temporal directions, creates major scientific ambiguity. To say that the controls are not really 'controls', and to give them a different name—such as the *referent* or *compeer* group—does not alter the ambiguity, since all three kinds of studies would then have to be called *case-referent* studies or *case-compeer* studies. The word 'retrospective' is often prefixed to indicate the backward direction of the etiologic case-control study, but the word *retrospective* is also ambiguous. It can be used for at least two different ideas: the first idea refers to the direction of populational pursuit as going backward from effect toward cause; the second idea refers to the timing of the collection of the research data as occurring after the date when the investigated events actually occurred. In a retrospective collection, the research data are obtained from routine medical records or from other sources of information that was previously recorded for purposes other than research.

In all three types of case-control studies, the data are usually collected in a retrospective manner. Long before the investigator ever reaches the research scene, the disease has already appeared or not appeared in the cases and controls; and much of the primary information has already been recorded. Since the direction of populational pursuit is not distinguished from the timing of data collection, the prefix *retrospective* does not alter (and may increase) the ambiguity of the phrase *case-control study*.

The word *etiologic* is also not a desirable prefix. The causal agent under investigation need not be etiologic, and may exert a protective rather than noxious effect. Because the key distinction of a retrospective or etiologic type of case-control study is that the examined groups are pursued backwards, toward an 'outcome' that is actually an antecedent entity, I recently suggested [2] the name *trohoc*, which is *cohort* spelled backward, for such research. Since a cohort study goes forward from cause to effect, the word *trohoc* seemed reasonable for studies that go backward from effect toward cause. I wanted something short and clear; and neither Latin and Greek dictionaries nor a Professor of Classics had a better term to suggest. I intended the word to illuminate, not to denigrate. Anyone who dislikes *trohoc* is invited to supply a suitable substitute that will replace the ambiguity and confusion of the words *retrospective* and *case-control*. Until such a substitute is provided, *trohoc* can continue to serve as a straightforward non-pejorative description of the research structure and purpose.

Turning now to the more fundamental scientific issues, let us begin by considering the basic scientific model of cause-effect reasoning. When we compare a causal agent against a comparative agent, the agents are imposed upon groups of people in a baseline condition. The people are then followed forward to observe the subsequent occurrence of the outcome event that is the alleged effect of the causal agent. The proportional occurrence or rate of that event is then compared statistically in the two groups. To attribute this statistical difference to the effects of the causal agent requires much more than merely a comparison of the rates in two groups. We must be sure that no major biases have occurred as the true or alternative causes of the observed differences.

The term *confounding factors* (or *confounding variables*) is sometimes used as a general name for problems that can distort or bias the result of an analytic study. The exact sources of the confounding variables are seldom specified, however, and their discovery often seems to occur via intuition rather than by direct attention to their hiding places. In the rest of this discussion, I shall specify three of the most important hiding places and I shall suggest ways of removing the associated biases.

For the subsequent statistical differences to be attributed to the causal agent, we must be assured that the two groups at baseline had equal susceptibility to the development of the outcome event; that the causal and comparative agents were imposed with an equal performance; and that the outcome event was sought with an intensity of identification procedures that would provide equal detection in the two groups. When substantial disparities occur in either susceptibility, performance, or detection, the associated biases can distort or invalidate the statistical comparisons.

Investigators who are contented with the current scientific status of trohoc research may argue that the type of longitudinal model I have shown here is irrelevant. In a trohoc study, we do not follow people longitudinally from baseline exposure or non-exposure to a causal agent. Instead, the groups are chosen from people with the effects or non-effects noted at the outcome end of the causal pathway. As long as the cases and controls seem reasonably chosen, and as long as the previous exposure to a causal agent is properly ascertained, a defender of the status quo in trohoc research may argue that we need not worry about the problems of susceptibility bias, performance bias, and detection bias.

This argument is scientifically untenable if the investigator wants to conclude that exposure to the alleged cause has indeed led to the suspected effect. Since the research receives a longitudinal interpretation, the evidence must be satisfactory for its longitudinal direction and for avoiding the biases that can distort the forward pathway that begins with exposure or non-exposure and that ends with the subsequent effect or non-effect. For sound scientific evidence, we need satisfactory methodologic standards to cope with these three sources of bias. What are those sources and what are the current standards?

Susceptibility bias can arise in at least three different ways. The first is demographic. Such features as age and gender can be risk factors for cancer and for many other diseases. The potential for this type of bias is well known and it is usually managed by methodologic standards in which the cases and controls are matched or stratified by demographically. The second source of susceptibility bias is clinical. For example, if Type A personality predisposes to cardiovascular disease and if women with Type A personalities decide preferentially to take oral contraceptive pills, a higher rate of cardiovascular disease will be found in association with the oral contraceptives. Similarly, if development of the post-menopausal syndrome is a prognostic harbinger of endometrial cancer and if estrogens are commonly prescribed for women with the post-menopausal syndrome, endometrial cancer will commonly be found in association with estrogens. As another example, if women with threatened abortion in early pregnancy are likely to produce deformed children, and if hormone therapy is prescribed for the threatened abortion, an association will be encountered between the treatment and the deformities. In all three of the cited examples, however, the decision to use the therapeutic agent may have denoted a prognostic distinction or risk for the subsequent disease. The agent may not have caused the disease.

A third source of susceptibility bias can be called *protopathic* [3]. In this situation, the agent that is later found to be present or absent in association with the disease was actually started or stopped because of an early manifestation of that disease. For example, oral contraceptives may be discontinued in patients in whom a lump in the breast is noted. If such women are later included as cases in a case-control study, they will have a spurious deficit of oral contraceptive takers, as Janerich *et al.* [4] have noted. Similarly, estrogens may be prescribed for an irregular bleeding that is the first manifestation of an endometrial cancer that has not yet been detected. When the cancer is finally identified at a later date, it becomes associated with the use of estrogens.

In contemporary trohoc research, no routine methodologic standards exist to deal with either clinical or protopathic sources of bias. The best way to look for such bias is to find out why the agent was prescribed and then to match or stratify the analyses appropriately. Such inquiries are seldom made and the data are seldom, if ever, analyzed

according to the clinical, pharmaceutical, or other reasons why the suspected agent was chosen or prescribed.

Performance bias arises during the procedures used for definition and ascertainment of the causal agent. One source of bias relates to the specification of exposure. The dosage and duration of whatever is meant by exposure should obviously be established before the data are analyzed and should be clearly reported for readers of the results. No general methodologic standards seem to exist for this activity. For example, in many studies of breast cancer or of endometrial cancer, the investigators have not stipulated what is meant by 'use of reserpine' or 'use of estrogens'. The second source of bias is the interviewer's preconceptions. A methodologic standard is used for this problem only sometimes. In many studies, the interviewer who inquires about whether or not the patient was exposed to the causal agent is aware of the research hypothesis or of the subject's identity as a case or a control. If the interviewer cannot be 'blinded' to either the research hypothesis or the subject's identity, the interview should be conducted under highly structured circumstances that arrange for the questions to be asked and recorded in exactly the same way for all of the people who are interviewed.

A third source of bias is the anamnestic recall of the interviewed subject. If the disease itself acts as a stimulus that makes the subject carefully review the possibility or degree of antecedent exposure, the control group should be stimulated to perform a similar review. This methodologic standard is observed only sporadically. Efforts are seldom made to stimulate the memory of the control group or to check the statements made by either the cases or the controls.

The problem of detection bias, despite its importance, has been remarkably ignored in epidemiologic research. Most clinicians are well aware of the fact that gallstones, cancers, thromboembolic phenomena, and many other chronic conditions can often exist in a sub-clinical or asymptomatic form, escaping detection during the patient's life and often remaining undetected unless necropsy is performed. For example, about a third of the gallstones, half of the coronary disease, and many of the cancers identified at necropsy were unsuspected or undiagnosed while the patient was alive. Since necropsy is performed in only about 20% of all deaths in the United States, a great many of these diseases are never diagnosed. If the appropriate diagnostic test happens to be ordered during the patient's lifetime, such diseases will be detected, but otherwise they will be missed.

To understand the sources of detection bias and to plan strategies for removing it, we must consider the entire pathway, shown in Fig. 1, that lies between a person's

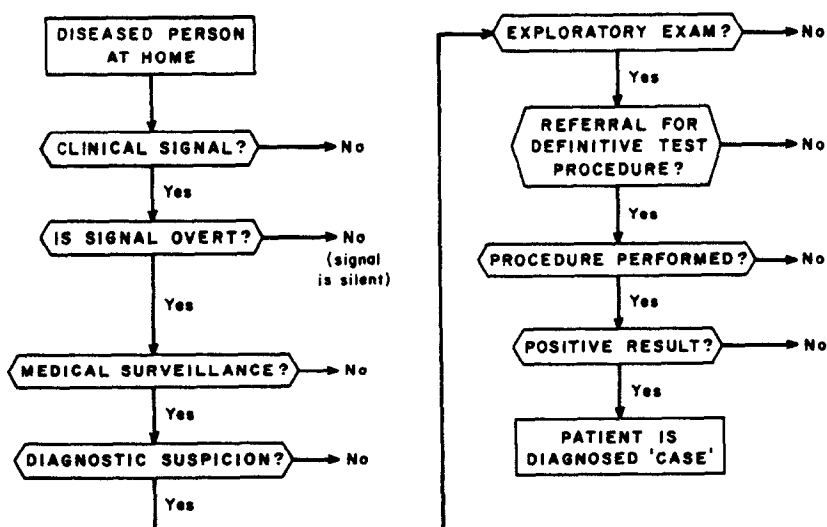


FIG. 1. Pathway between a disease and a diagnosis.

anonymity at home as a possessor of a disease and his ultimate identity as a diagnosed case. The diseased person at home may or may not develop a clinical signal event. The signal event can be an overt manifestation, such as uterine bleeding, or the signal event can be silent, such as a lump in the breast. The person may then receive medical surveillance (provoked by the signal event or by other reasons), or the person may not receive surveillance. If surveillance occurs, diagnostic suspicions may or may not develop in the doctor, who may or may not perform an appropriate exploratory examination, such as palpation of the breast or pelvic examination. The doctor then may or may not refer the patient for a definitive test procedure, such as a biopsy of the breast or a uterine dilatation and curettage. Although planned, this procedure may or may not be performed. If it is performed, its result must be interpreted, and if the interpretation is positive, the patient has finally become a diagnosed case.

Exposure to an alleged etiologic agent can create detection bias at several places along this complex pathway. The first location of bias is in community surveillance. Exposure to an ongoing treatment, such as reserpine or estrogen, may increase a person's medical surveillance, without affecting a silent signal event, such as a lump in the breast, but may thereby allow the lump a greater chance of being medically detected. Alternatively, exposure to an agent such as estrogen may convert a 'no signal' to an overt signal event, such as uterine bleeding, without affecting the true occurrence of a disease such as endometrial cancer. This overt event may then lead to the diagnostic testing and identification of a previously asymptomatic cancer. In both of these two instances, the exposed group would receive a greater community surveillance than the people who were non-exposed. Thus, if estrogens provoke the bleeding that leads to a referral for dilatation and curettage, asymptomatic uterine cancer is much more likely to be detected in women receiving estrogens than in those who do not receive it.

A second source of detection bias occurs if the exposure leads to an increased ordering of exploratory or definitive diagnostic procedures not just for people in the community, but for patients hospitalized for other reasons.

A third source of detection bias can occur if a knowledge of exposure or non-exposure alters the objectivity of the person who interprets the results of the test. For example, if a pathologist examining endometrial tissue knows whether or not the patient is receiving estrogen, the pathologist's diagnostic decision may be affected by that previous knowledge. This third source of detection bias is sometimes considered in trohoc research, and separate reviews are sought to eliminate the false positive results of histologic or other diagnostic evidence in the cases. Unfortunately, such reviews do not deal with the problem of false negative diagnoses in the controls, or with the problem of diagnostic referral bias in both the cases and controls.

This term, *diagnostic referral bias*, includes the impact of both community surveillance bias and diagnostic examination bias. It is substantially different from the type of hospitalization bias that was first described by Berkson [5] as a purely passive mathematical phenomenon, in which people with two ailments have a higher probability of being hospitalized than people who have only one ailment. Diagnostic referral bias is an active clinical entity, in which physicians create different rates of hospitalization and/or diagnostic testing according to the clinical events that occur in the four groups of people who are exposed or non-exposed, diseased or non-diseased.

TABLE 1. EXPOSURE, DISEASE AND HOSPITAL REFERRAL

Presence or absence of exposure	Proportion	Presence or absence of disease	Proportion	Proportion referred to hospital	Proportions in hospital
Exposed	e	{ Case	p_2	h_4	ep_2h_4
		{ Control	$1-p_2$	h_3	$e(1-p_2)h_3$
Non-exposed	$1-e$	{ Case	p_1	h_2	$(1-e)p_1h_2$
		{ Control	$1-p_1$	h_1	$(1-e)(1-p_1)h_1$

To see the effects of diagnostic referral bias, let us consider the algebra shown in Table 1. In the community, the proportion of exposed people can be indicated as e , and the non-exposed as $1-e$. The target disease develops at rate p_1 to create the cases of the non-exposed group, and rate p_2 in the exposed group. The rates in the corresponding controls are $1-p_1$ and $1-p_2$. These four groups will have hospital referral rates ranging from h_1 in the non-exposed controls up to h_4 in the exposed cases. What is found for these four groups in the hospital will then have the proportions shown in the column on the far right of Table 1.

TABLE 2. CALCULATION OF ODDS RATIO

$$\begin{aligned}
 O &= \frac{\text{Exposed cases} \times \text{non-exposed controls}}{\text{Exposed controls} \times \text{non-exposed cases}} \\
 &= \frac{ep_2h_4 \times (1-e)(1-p_1)h_1}{e(1-p_2)h_3 \times (1-e)p_1h_2} \\
 &= \frac{p_2}{p_1} \times \frac{1-p_1}{1-p_2} \times \frac{h_4}{h_2} \times \frac{h_1}{h_3} \approx \frac{p_2}{p_1} \times \frac{h_4}{h_2} \times \frac{h_1}{h_3} \\
 &\approx \frac{p_2}{p_1} \times \frac{k_d^*}{k_c^\dagger}, \text{ if } k_d = \frac{h_4}{h_2} \text{ and } k_c = \frac{h_3}{h_1}
 \end{aligned}$$

* k_d = bias due to exposure in cases.

† k_c = bias due to exposure in controls.

As shown in Table 2, when the odds ratio is calculated for the hospitalized cases and controls, the terms containing e and $1-e$ cancel out. The terms $1-p_1$ and $1-p_2$ are very close to 1 and can be ignored, and so the ratio becomes $(p_2/p_1) \times (h_4/h_2) \times (h_1/h_3)$. If we let k_d indicate h_4/h_2 as the bias due to exposure on the case group's referral, and if we let k_c indicate h_3/h_1 as the analogous bias of exposure on the control group's referral, the odds ratio is the true risk ratio p_2/p_1 , multiplied by k_d/k_c , which is the ratio of the two referral biases. If those two biases are equal, they will cancel one another. If not, they can convert the odds ratio into a substantial distortion of the true risk ratio. The information shown in Table 2 also indicates why it is often a futile exercise to select control groups from neighbors or other community sources. Such a selection can remove bias from the control group ratio of h_3/h_1 , but as long as the case groups are chosen from hospitalized diagnosed patients, the h_4/h_2 bias will remain.

These problems in detection bias are generally ignored in contemporary trohoc research and they are particularly difficult to deal with. Working at the case-control end of the pathway, the investigator cannot determine the true forces of surveillance and referral that may have biased his collection of exposed and non-exposed cases and controls. To try to equalize those external forces of surveillance and referral, we can choose both the cases and controls from a single sampling frame, consisting of people who received the particular diagnostic test used to identify the target disease. For example, in a study of breast cancer, the sampling frame might be a registry of all patients who have received a biopsy of breast lesions. For endometrial cancer, the sampling frame might be all patients who have received dilatation and curettage, or hysterectomy. This kind of sampling frame also has the advantage of avoiding the investigator's arbitrary decisions about the choice of diagnoses for the case and control groups. The people who are potentially eligible to be cases and controls are defined by the results of the diagnostic test.

With further algebra [6], it can be shown that this kind of sampling frame will substantially reduce diagnostic referral bias, but will not eliminate it. To get closer to elimination, we need yet another step, which consists of stratifying the patients according to presence or absence of the cogent clinical reason(s) that can lead to referral to the hospital for the diagnostic test. For example, in studies of endometrial cancer,

the patients sampled from the diagnostic-test registry should be stratified into two groups: those who were or were not referred to the hospital for uterine bleeding. If a similar odds ratio is found in both groups, the two ratios are quite likely to reflect the true risk ratio.

Regardless of whether or not you like the proposed solutions, my main point is that we can no longer remain complacent about the problems. The many contradictions and conflicts that have occurred in case-control studies of the same topic [3], most recently illustrated by reserpine and breast cancer, provide prominent, compelling evidence that the problems are real, important, and distressing. The main value of trohoc research is that it can be used, when experiments are either impossible or unfeasible, as a convenient, quick, and easy way to explore cause-effect relationships. But the scientific standards for performing and interpreting trohoc research should not also be convenient, quick, and easy. Those standards should reflect the same rigorous criteria that pertain when cause-effect reasoning is applied in other scientific activities; and the standards should contain the same careful attention to the prevention or removal of susceptibility bias, performance bias, and detection bias.

Every scientific domain must go through a stage of early growth, often with many errors and fallacies, before it reaches scientific maturity. Before modern chemistry reached its current intellectual stature, the formerly plausible beliefs in phlogiston and in other concepts of alchemy had to be replaced or altered by newer scientific approaches and standards. There is nothing shameful about the currently underdeveloped state of science in case-control research. The domain is quite young, with many problems to be solved and many challenges to be mastered. This symposium gives us an excellent opportunity to recognize those problems and challenges, and to begin planning effective scientific methods for improving the future.

REFERENCES

1. Feinstein AR: Clinical biostatistics: XLII. Architecture of cross-sectional research (Part I). *Clin Pharmacol Ther* 23: 81-96, 1978
2. Feinstein AR: Clinical biostatistics: XX. The epidemiologic trohoc, the ablative risk ratio, and 'retrospective' research. *Clin Pharmacol Ther* 14: 291-307, 1973
3. Horwitz RI, Feinstein AR: Post trohoc ergo propter trohoc: problems, conflicting results, and criteria for scientific standards in retrospective 'case control' research. *Clin Res* 24: 248A, 1976
4. Janerich DT, Glebatis DM, Dugan JM: Benign breast disease and oral contraceptive use. *J Amer Med Assn* 237: 2199-2201, 1977
5. Berkson J: Limitations of the application of fourfold tables to hospital data. *Biometrics Bull* 2: 47-53, 1946
6. Feinstein AR, Horwitz RI: On the production and removal of bias in 'retrospective case-control' research: an algebraic analysis. (Submitted for publication.)