Choosing Cases and Controls: the Clinical Epidemiology of "Clinical Investigation"

Alvan R. Feinstein and Ralph I. Horwitz
Department of Internal Medicine and the Clinical Epidemiology Unit, Yale University School of Medicine, New Haven, Connecticut 06510

Choosing a suitable "control" has never been an easy job.

One immediate difficulty is in the concept itself (1). The word "control" has been applied for many different things beyond the standard idea of *comparison.* They include the idea of regulation—such as a controlled environment, a well-controlled blood sugar, or quality control in a measurement process. The ideas also include an antecedent baseline—such as the control value or control period that precedes the subsequent values in a before-and-after set of observations. Statistically, the word control can denote an analytic activity, such as the partition of control strata or the variables that are controlled in a multivariable analysis.

In most types of medical scientific thought, however, control refers to something chosen as the comparison for the principal agent that presumably leads to the outcome of a cause-effect relationship. For the comparisons of either therapeutic or etiologic agents, a placebo may be the control for active treatment, medical therapy for surgery, unexposed nonsmokers for exposed smokers. The controlled comparison may extend into a "bioassay" or "dose-response curve" if the effects of higher doses or durations of the active agent—smoking, alcohol, pharmaceutical substances—are contrasted against the effects of lower doses or none.

This straightforward scientific concept of control is easily converted into a research design if the investigator can govern (i.e., control) the assignment of the compared agents. This type of governance has always occurred in laboratory experiments with animals or inanimate substances. The governance has also occurred in experiments with people during the past four decades when randomized clinical trials have been used to test therapeutic agents. In these governing circumstances, the investigator decides which active and control agents will be compared, arranges to assign them appropriately, and then carries out the follow-up observations.

In many research studies of people, however, the investigator does not have the opportunity to construct an experimental test of the active agent. If it is suspected of being a noxious cause of disease, randomized trials will usually be unfeasible and often unethical. Few people will volunteer to participate in a randomized trial of smoking vs. nonsmoking, or to be exposed to possibly hazardous chemical agents. If the research is concerned with pathophysiologic effects rather than etiologic causes of disease, the main causal agent is the disease itself, and the control agent is the absence of the disease. In a true pathophysiologic experiment—testing whether diabetes causes retinopathy or whether myocardial infarction elevates certain enzymes—the investigator would see the effects that occur after the disease is induced as the active agent in some members of a group of healthy people, while keeping others as nondiseased controls. These pathodynamic relationships, in which the disease is studied as the cause of a pathophysiologic effect, can seldom be investigated by creating the disease experimentally. Even if healthy people were willing to volunteer, and even if an institutional review board approved the ethics of the research, the investigator might not know exactly how to produce the desired natural form of the disease.

Consequently, the scientific advantages of a truly experimental research design can seldom be applied to investigate etiology or pathophysiology. Instead, the research is often done by selecting a group of "cases," who already have the target disease, and a group of controls, who do not. The investigator then examines and compares the level of the focal variable in the diseased cases and nondiseased controls. In both etiologic and pathodynamic research, the investigator works with a cross-sectional collection of cases and controls, and with concurrently obtained data for the focal variable. In an etiologic interpretation, however, the focal variable represents an antecedent event: the active agent that allegedly caused the disease. In a pathodynamic interpretation, the focal variable represents a subsequent event: the pathophysiologic effect for which the disease was the active cause.

Regardless of whether the directional interpretation of the data is backward towards etiology or forward towards pathophysiology, this type of research is particularly difficult to design. A cross-sectional snapshot of disease is being substituted for a longitudinal moving picture; and the causal pathway has been completed when the research begins. In case-control studies, the causal agent, the comparative agent, and their corresponding outcomes have already occurred before any research data are collected. In a study of etiology, each person has already been exposed or nonexposed to the suspected causal agent, and the outcome disease did or did not develop. In a study of pathophysiology, each person has either developed or not developed the "causal" disease, and the appropriate pathodynamic derangement has or has not appeared.

Having been chosen at the end of the causal pathway, the cases and controls have a particularly difficult scientific role. To substitute for the governed plans and forward observations of an experiment, the cross-sectional attributes of the cases and controls must somehow allow the investigator to recapitulate the entire causal pathway. Without an effective recapitulation, the available data can lead to substantial errors of interpretation. If substance X is initially elevated but is later lowered in

Received for publication 14 September 1987 and in revised form 6 October 1987.

J. Clin. Invest.

0021-9738/88/01/0001/05 $2.00
Volume 81, January 1988, 1-5

the course of disease D, the distinction will be missed if the chosen cases have had disease D for too long a time. This type of problem occurs when an elevated systolic blood pressure is lowered by a completed myocardial infarction, or when a previously positive skin test turns negative in disseminated tuberculosis. If people who develop an elevated substance Y with disease E are likely to die early, the elevation will be absent if the chosen cases consist of long-term survivors. An example of this problem is the statistical attrition that occurs when people with particularly high levels of cholesterol die at younger ages than those with lower levels. In a longitudinal set of cross-sections of people assembled at ages 40, 50, 60, and 70, the level of cholesterol will seem to be lowered by advancing age, rather than by the deaths that removed high values from the older-age groups.

In addition to these problems in the serial timing of events, a case-control study of pathophysiology has two other major difficulties. One of them is to disentangle the temporal sequence of cause and effect. Did the abnormality noted in the focal variable occur before or after the onset of the target disease? Thus, if enzyme zeta is found to be substantially higher in cases of myocardial infarction than in the controls, is a high enzyme zeta an etiologic risk factor that may lead to myocardial infarction, or a pathodynamic consequence of the established disease? In patients with cancer, does a low cholesterol precede the cancer as a cause, or follow it as an effect?

A second additional problem is to provide distinctive accountability in separating the primary effects of a disease itself either from its secondary complications, or from intervening extraneous factors, such as therapy or associated ailments. When substance Z is found to be elevated in patients with a particular cancer, is the elevation due to the primary cancer, to a secondary metastasis, to the associated malnutrition or debility, to the chemotherapeutic agents, or to the co-morbidity of concomitant chronic lung disease?

These issues in seriality, sequence, and accountability create distinctive clinical epidemiologic challenges for the type of research that often appears in *The Journal of Clinical Investigation.* The challenges are seldom regarded as epidemiologic, however, because the clinical investigators usually concentrate on their own "experimental" activities in designing the particular arrangements and procedures that eventually yield the measured results of enzyme zeta or whatever principal focal variable will be compared in the diseased cases and nondiseased controls. With an emphasis on the main "experiment" that is governed during the clinical research, the investigators often overlook the previous ungoverned experiment of nature that produced the clinical conditions of the cases and controls who are under investigation. In that previous natural experiment, the active agent is either an etiologic entity or a disease; it was assigned without a deliberate research plan; and the investigator's experiment is done to determine what happened afterward. The total challenge thus has two components: the epidemiologic selection of suitable cases and controls, and the clinical investigation of the chosen focal variables.

Even when adequately considered, however, the epidemiologic challenges of this form of clinical investigation are substantially different from what is conventionally encountered in epidemiologic case-control studies. The main source of the difference is that conventional epidemiologic studies are usually aimed at the etiologic roles of *externally* imposed maneuvers—public health "risk" factors (smoking, alcohol, high fat diets) or the adverse effects of therapeutic agents that allegedly promote one disease while treating another (reserpine/breast cancer, estrogens/endometrial cancer, non-steroidal antiinflammatory agents/agranulocytosis). Because the external exposures can usually be clearly identified and dated, conventional epidemiologists seldom have the problems of sequence that arise when the focal variable is an internal chemical or biologic entity that is paraclinically measured in tissue, blood, or other body substances. This paraclinical entity is noted at the same time as the target disease; and clinical investigators cannot easily answer the sequence question about which came first: the disease or the focal entity. Uncertain about whether the focal entity is an etiologic predecessor or a pathodynamic consequence of the disease, the investigators can often draw only the "pathoconsortive" conclusion that a cause-effect relationship exists, that its sequential direction is unknown, and that "this interesting finding warrants further research."

Another distinctive feature of conventional epidemiologic studies is that information about the external exposures is readily acquired by direct interview, telephone calls, mailed questionnaires, or review of previous records. Because this type of information is easy to get, the epidemiologic investigators can readily assemble large numbers of cases and controls. In clinical investigations, however, data for the focal variable are internal. They may be obtained with a special research protocol, using experimental procedures that often involve diverse ingestions, injections, and tests, as well as admission to a clinical research center. Even if large numbers of people were willing to volunteer for the research process, the time and costs of investigating large groups would be prohibitive. For this reason, clinical investigators often work with groups that most epidemiologists would regard as tiny.

In quantitative statistical appraisals, the small "sample sizes" of the clinical investigations immediately raise questions about representativeness. Is the universe of the disease under study suitably represented by the few people who are recruited for the research—the three patients with cancer of the colon or the four patients with hyperparathyroidism? In qualitative scientific thinking, however, the research is intended to explain biologic mechanisms, not to offer a mathematically accurate portrait of the disease. The investigator wants to determine not what the disease *is,* but what it *does* or how it *develops.* For this purpose, the investigator is usually concerned less with quantitative representation than with qualitative validity. The qualitative characteristics of the individual members of the case and control groups should allow a suitable resolution of the problems of sequence, seriality, and accountability that may otherwise cloud the cause-effect interpretation of biologic mechanisms.

The choice of suitable qualitative characteristics for cases and controls has received relatively little attention in conventional epidemiologic research, as well as in clinical investigation. Despite the formidable scientific challenges of using arbitrarily chosen cross-sectional groups as substitutes for the governed design and forward observations of a planned experiment, the conventional epidemiologic studies have usually been regarded more as an exercise in statistical associations than in scientific architecture (2, 3). In recent years, however, the scientific challenges have become increasingly discussed, and the proposed new approaches have evoked lively controversies (2–6).

Most of the controversies deal with the methods to be used for preventing, reducing, or adjusting the biases that can arise when external exposures are the focal variables under study. Although easy to investigate, external etiologic agents are chosen by personal decisions of the recipients or by the clinical recommendations of physicians. These selective choices can lead to substantial biases in results for the compared groups that receive or do not receive the main agent under study. The biases arise if the exposed and nonexposed groups are substantially different in baseline susceptibility to the outcome disease, in subsequent detection of the outcome events, and in retrospective ascertainment of data regarding exposure (2). Although a prime threat to the validity of many epidemiologic case-control studies, these biases are less cogent problems in clinical investigations where the focal variables are internal "paraclinical" entities, which are not openly manifested or selected, and which are identified with objective tests.

Consequently, the case-control studies of clinical investigation have not been (and need not be) embroiled in scientific conflicts about avoiding bias. The clinical investigations have also not been subjected to the statistical theories, multivariate models, and other mathematical strategies that are often proposed as guidelines for conventional epidemiologic research. Instead, some simple scientific principles can be used for dealing with the problems of accountability, scope, seriality, and sequence that are the main epidemiologic challenges in the paraclinical case-control studies of pathophysiology.

A key concept in these principles is to recognize the diverse clinical conditions that can occur in the spectrum of a disease. Regardless of whether the disease itself is an abnormality in structure (such as carcinoma of the colon) or function (such as hyperparathyroidism), the scope of the spectrum will contain a large diversity of clinical conditions. For example, a cancer can be morphologically localized, regionally spread, or distantly disseminated. Clinically, the cancer may have produced no symptoms, primary symptoms (such as bleeding), systemic symptoms (such as weight loss), or metastatic symptoms (such as bone pain). Paraclinically, the cancer may have led to diverse abnormalities in blood, urine, or feces. From the current morphologic, clinical, and paraclinical manifestations and from any previous information about them, the investigator can classify the cancer as having been present for a short, long, or unknown length of time. Beyond the cancer itself, the patient may have one or more co-morbid diseases, each with its own spectrum of manifestations, and the patient's condition may have been affected by antecedent or concomitant therapy. In addition to these ailment-oriented attributes, the patient's general functional status may range from bedridden, requiring total care, to unimpaired performance in all tasks of daily life. Finally, the people who are hosts to the clinical condition may be demographically old or young, male or female, in high or low socioeconomic status.

All of these variations produce a complex array of different patterns for clinical conditions in the spectrum of a disease. The complexity of the spectrum creates both perils and helpful opportunities in the research. The peril is that an investigator who thinks only about the disease itself may erroneously hold it responsible for phenomena that are really caused by other events in its total spectrum. The helpful opportunity is that many of the cited scientific problems can be resolved by choosing cases and controls from suitable parts of the spectrum. The rest of this discussion contains some suggestions about how to make those choices.

Accountability

To make cause-effect decisions about pathodynamic effects, the disease must be held accountable for the abnormalities in the focal variable. Healthy controls will suffice to show that the abnormalities are absent in health, but will not allow a clear separation of what is caused by the disease itself from what is caused by its clinical complications, such as debility or depression, or what arises from extraneous concomitant phenomena, such as therapy or co-morbidity. Consequently, the cases should preferably be "pure" instances of the disease, without any clinical complications or extraneous phenomena.

These pure cases may be hard to find, particularly since such patients are often in relatively excellent functional condition and are seldom eager to volunteer for research procedures. On the other hand, the "sick" cases, who are more amenable to participating in clinical investigation, will often have the undesirable complications and extraneous phenomena. If the cases contain these complexities, the control group will have to account for them by including people who do not have the target disease, but who have other sources of debility, depression, therapy, co-morbidity, or whatever complexity exists in the cases.

Even with pure cases of disease, however, controls may be needed to distinguish what is caused by local derangements or general pathologic abnormalities from what is due to the distinctive pathophysiology of the disease itself. For example, in patients with cancer of the colon, abnormalities may arise from the local derangement of the colon or from a general effect of neoplasia. Appropriate controls for these possibilities might have nonneoplastic diseases of the colon, or neoplastic disease at other sites. Analogously, in patients with hyperparathyroidism, some other parathyroid or endocrine diseases might be appropriate controls.

Scope

Although clinical complications and extraneous phenomena should either be avoided in the cases or accounted for in the controls, the scope of the spectrum of pure disease will often include patients with distinctive secondary phenomena, such as metastases of cancer, lupus nephritis, posthepatitic cirrhosis, or hyperparathyroid bone disease. An investigator who wants to consider only the *primary* features of the main disease might want to exclude such patients from the group of pure cases. Their exclusion, however, may lead to limited or erroneous conclusions because patients in the secondary parts of the spectrum may reveal distinctions that are not otherwise apparent for the scope of biologic mechanisms with which the disease can evolve. The metastases of a cancer may arise from different pathodynamic (or etiologic) mechanisms than the local lesion. Lupus nephritis, lupus cerebritis, and lupus alone (with or without arthritis) may have different causes or consequences.

Perhaps the best way to preserve a suitable scope for the disease spectrum, while avoiding the confusion of mixing primary and secondary instances of disease, is to have two separate groups of cases. One group would consist of patients with primary manifestations only. The second group would contain cases from the distinctive secondary parts of the spectrum.

The scope of the spectrum is particularly important in the special types of case-control research used to assess the efficacy of diagnostic marker tests, such as enzymes for myocardial infarction or carcinoembryonic antigens (CEA)[1] for cancer of the colon. The marker test usually depends on a pathodynamic abnormality produced by the disease, but the research is aimed at identification rather than mechanism. The goal is to identify the disease whenever it is present, regardless of what form it may take, and to exclude the disease when it is absent.

A suitable scope of the full spectrum of disease and controls is needed to avoid both false negative and false positive results for the marker test (7). The false negative results can occur if the proposed marker has not yet become elevated in "early" or primary cases, or if positive tests in the cases are obscured by concomitant debility, therapy, co-morbidity, or other extraneous events. The false positive results may arise if a similar pathodynamic effect is produced by other diseases or by associated phenomena, such as debility or therapy. The neglect of a suitable spectrum of cases and controls has been a common source of defects in the initial evaluations of diagnostic marker tests; and the fallacious results have often not been recognized until long after the tests had become popular and widely used (7, 8).

For example, elevations in CEA were originally noted as a pathodynamic consequence of carcinoma of the colon. An elevation in CEA then became widely used as a diagnostic marker test until its many false positive and false negative results indicated that it had little diagnostic value in discriminating colon cancer from other diseases. The CEA test is now applied mainly in patients with a resected colon cancer as a prognostic marker to suggest recurrences that have not yet become clinically evident. An abnormal dexamethasone suppression test, which was originally regarded as a pathodynamic consequence of psychiatric depression, also had a period of popularity as a diagnostic marker test until false positive results were often noted in many other clinical conditions.

In pure pathophysiologic research, however, the investigator need not be concerned about the surrogate accuracy of diagnostic marker tests. The cases and controls therefore need not cover a large, broad spectrum of candidate possibilities. The individual members of each group can be chosen with deliberate specifications for the challenges to be met. Although the exact guidelines will differ for each disease under study, the main point to bear in mind is that the cases for pathophysiologic research should preferably cover a suitable scope of the pure spectrum of the disease, and that members of the control group should account for whatever "impurities" (debility, therapy, co-morbidity) could not be avoided in the cases.

The problem of representativeness may sometimes create a problem in future clinical practice, rather than in mechanisms of etiology or pathophysiology. Because rare or unusual cases of disease are often referred to medical centers for diagnosis or therapy, the frequency of these cases may be overemphasized in the ensuing spectrum of the reported clinical investigations of the disease. The misrepresented spectrum will not distort studies of the disease's mechanisms, but may lead to excessive future "workups" searching for unusual cases at less specialized medical settings. For example, many clinical investigations have been published for patients having renal, adrenal, or other ailments as secondary causes of hypertension. The frequency of the reports led to many routine but nonproductive searches for these ailments in hypertensive patients until clinicians recognized that the secondary causes occurred too rarely for the routine workups to be justified.

Seriality

With suitable use of data from history taking and review of previous tests, the pure cases of disease can often be divided into those who have had the disease for short or longer lengths of time. By comparing the levels of the focal abnormality in the early and later cases, the investigator may be able to discern the longer-term pathodynamic consequences of the disease. This type of "longitudinal cross-section" comparison will be effective, however, only if reassurance is available to rule out the changes, discussed earlier, that are produced by the statistical attrition of early deaths, rather than by the pathophysiology of the disease.

Alternatively, the investigator may observe serial effects by expanding the research from its cross-sectional status to include a longitudinal component. After the early cases have been suitably examined, they can be followed thereafter and reexamined at a later date to see what has happened as the disease evolves. With either the cross-sectional or longitudinal approach, however, the observed effects must be clearly attributable to the disease itself, rather than to clinical complications or to extraneous phenomena, such as therapy or co-morbid ailments.

Sequence

Deciding whether the focal abnormality preceded or followed the onset of the disease is the most difficult problem in cross-sectional paraclinical research; and it cannot be easily solved. The previously cited difficulties in accountability, scope, and seriality can be eliminated (or reduced) by choosing pure cases, in a primary part of the spectrum, with relatively early disease. If a focal abnormality is found in such patients (and is absent in suitable controls), it can clearly be ascribed to the disease. There is no certain way, however, to decide routinely whether the abnormality is an etiologic cause or a pathodynamic effect of the disease.

A distinction can sometimes be achieved if the abnormality was previously tested, and found to be present or absent, before the overt appearance of the disease. Such tests are seldom likely to have been done, however. In most instances, the decisions will depend on what happens to the focal abnormality if early patients are followed or if a group of appropriate later patients are examined cross-sectionally. In general, a serial decrease in magnitude of the abnormality will suggest that it is etiologic, and an increase will suggest that it is pathophysiologic, but sometimes pathophysiologic effects may go down instead of up.

Sometimes the distinction can be clarified if suitable therapy is available. For example, Orlowski (9) noted elevated levels of cerebrospinal β-endorphin in the infant apnea syndrome, but the case-control study did not permit a decision about whether the elevations "were the result or the possible cause of the apneas." He concluded that the elevation was

[1] *Abbreviation used in this paper:* CEA, carcinoembryonic antigen.

probably causal, however, when further research showed that naloxone infusion reduced the apnea episodes in three patients whose raised endorphin levels were lowered. The naloxone had no clinical effect in one patient whose endorphin level was normal.

The accurate differentiation of directional sequence in "pathoconsortive" relationships requires intimate knowledge of the disease process, immaculate care in the choice of suitable cases and controls, and a large dose of good luck. In certain situations, however, an accurate differentiation may not be necessary. The demonstration of a close causal relationship between the disease and the focal abnormality, even if uncertain in directional sequence, may yield enough "insight" to lead to subsequent studies that can clarify the issue, or convert it into a more solvable problem. For example, a substantial pathoconsortive problem occurred several years ago when HIV (formerly HTLV-III) antibodies were noted to be elevated in patients with AIDS. The HIV elevation could have been an etiologic antecedent of AIDS or the pathodynamic consequence of opportunistic co-morbid infection. The close relationship between HIV and AIDS was the main cause-effect discovery, however, and it led to significant progress in subsequent research. Eventually, the question of sequence was resolved with studies of pure early cases of AIDS, and with longitudinal follow-up of patients with HIV elevations who had not yet developed AIDS.

Conclusion

Although clinical epidemiologists often study topics in diagnosis, prognosis, and therapy that have been omitted from the inventory of conventional epidemiology, another distinction of clinical epidemiology (2) is the sophisticated clinical attention given to the choice of case and control groups for studies of etiology. This type of sophistication becomes an important epidemiologic challenge in selecting patients for clinical investigations that are aimed at understanding the pathophysiologic consequences of disease and at differentiating etiologic or extraneous causes from inherent pathophysiologic effects.

The challenge requires a conjunction of scientific activities in taxonomic categorization and in biologic explication. The current era of molecular biology began with such a conjunction, when electrophoresis was applied to the hemoglobin of patients with sickle cell anemia. If the electrophoretic material had come from a nonspecific collection of general patients, or from a group of patients with nonspecific anemias, the results might have been too inconsistent to be valuable. Before the electrophoresis was done, however, clinical hematologists had engaged in an act of clinical epidemiology. They had explored the spectrum of anemia, and had culled out the cases with sickle cell anemia. When this well-delineated group was examined, dramatic results were found.

Similar challenges exist today in the conjunctive interplay between clearly specified clinical groups and clearly identified biologic mechanisms. The challenges occur for current research in "biologic psychiatry" and in the oncogenes of cancer, as well as in the pathophysiology or molecular biology of many other "organic" diseases. The scope of the challenges allows a fertile meeting ground for the epidemiologist's concern with groups, for the pathophysiologist's concern with biologic mechanisms, and for the concern of all scientists with research that is "basic" because the results are important, accurate, and enduring.

Acknowledgments

This work was supported in part by grants from the A. W. Mellon Foundation and The Council for Tobacco Research.

References

1. Feinstein, A. R. 1973. Clinical biostatistics. XIX. Ambiguity and abuse in the twelve different concepts of 'control'. *Clin. Pharmacol. Ther.* 14:112–122.

2. Feinstein, A. R. 1985. Clinical Epidemiology. The Architecture of Clinical Research. W.B. Saunders Co., Philadelphia.

3. Horwitz, R. I. 1987. The experimental paradigm and observational studies of cause-effect relationships in clinical medicine. *J. Chronic Dis.* 40:91–99.

4. Ibrahim, M. A., and W. O. Spitzer, editors. 1979. The case-control study: consensus and controversy. *J. Chronic Dis.* 32:1–144.

5. Miettinen, O. S. 1985. The "case-control" study: valid selection of subjects. *J. Chronic Dis.* 38:543–548.

6. Schlesselman, J. J. 1982. Case-Control Studies: Design, Conduct, Analysis. Oxford University Press, New York.

7. Ransohoff, D. F., and A. R. Feinstein. 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N. Engl. J. Med.* 299:926–930.

8. Nierenberg, A. A., and A. R. Feinstein. 1988. The rise and fall of the dexamethasone suppression test. Lessons for the evaluation of diagnostic markers. *JAMA (J. Am. Med. Assoc.).* In press.

9. Orlowski, J. P. 1986. Cerebrospinal fluid endorphins and the infant apnea syndrome. *Pediatrics.* 78:233–236.