catalogado

J Chron Dis Vol 38, No. 7, pp 543 548, 1985
Printed in Great Britain

0021-9681/85 $3.00 + 0.00
Pergamon Press Ltd

Variance and Dissent

THE "CASE-CONTROL" STUDY: VALID SELECTION OF SUBJECTS

OLLI S. MIETTINEN*

School of Public Health, Harvard University, Boston, Massachusetts, U.S.A.

INTRODUCTION

VALID selection of subjects in "case–control" studies remains problematic on the level of principles even [1], to say nothing about the difficulties of research practice itself.

As an illustration, consider the hypothetical example of testing the hypothesis that a cause of traveller's diarrhea is the consumption of tequila (a Mexican drink), with cases derived from a hospital in Acapulco, Mexico, over a defined period of time. What might be the proper "control" group?

Lilienfeld, in his textbook on epidemiology [2], expresses the commonly held view that the "control" group should be representative of "the general population" (as to the exposure rate). But the meaning of this is very obscure in the example at hand.

This general concept is made somewhat more restrictive in the recent textbook by Schlesselman, specific to "case–control" studies [3]. He refers to a need, in general, to sample "the target population," defined as "a subset of the general population that is both at risk of the study exposure(s) and the development of the study disease." While, as was noted, the meaning of even "the general population" is already quite unclear in the example at hand, totally mysterious is the concept of its subset that is "at risk" for both tequila use and the development of traveller's diarrhea.

Whatever may be the meanings of those concepts in this example, it is clear that the common practice of using neighbors (who might reside in Boston, Montreal, etc.) as "population controls" would be very inappropriate—grossly exaggerating the causal relation of the incidence of traveller's diarrhea to the consumption of tequila.

It is the purpose here to propose basic principles of valid selection of subjects in "case–control" studies. They derive from a reassessment of the presumed nature of the "case–control" study.

THE ESSENCE OF THE "CASE-CONTROL" STUDY

A commonly held concept of the "case–control" study is that it is the alternative to the cohort study. The distinction is taken to be one of "sampling" in the sense that in a cohort study one "samples" people free of the illness, representing different categories of the causal factor (potential or known) under study, and follows them forward in time to learn

regarding the causal factor. In other words, so goes the thinking, in a cohort study the investigative movement is "from cause to effect" and in a "case-control" study "from effect to cause" [3].

This concept of the essence "case–control" studies as representing the reverse of cohort studies I believe to be in error. I term it "the 'trohoc' fallacy," using Feinstein's [4] highly descriptive term for the commonly espoused reverse-of-cohort notion.

Almost as erroneous, and misleading, I consider the component notion that *in* "case–control" studies the concern is to compare cases with non-cases.

Any study—"case–control" or whatever—on the incidence of illness must be based on the incidence experience of a particular population as it moves over time. The study population may be defined by an *event* experienced by its members, with the membership lasting forever thereafter; or it may be defined by a *state*, lasting for the duration of that state. The two types of membership criterion define *cohorts* (i.e. adynamic, or closed populations) and *dynamic* (open) populations, respectively. They are exemplified, respectively, by the patients in a clinical trial—a cohort defined by the event of enrollment into it—and the catchment population (for a given illness) of a particular hospital—a dynamic population defined by the state that if the illness were to develop, one would go to the hospital. As another illustration of the duality, consider the Framingham Heart Study. The study population is a cohort, defined by the event of enrollment into it, in 1948, once and for all. An alternative to this would have been to follow the population of Framingham residents from 1948 onward—the catchment population for a case registry there—with people entering and exiting the resident status in the course of the study. Thus the term "cohort" refers to one of the two possible types of *dynamics* for the study population, and the alternative to a cohort study is a dynamic population study—rather than a "case–control" study.

Given a study population's experience over time, or a *study base*, it is necessary to ascertain the relevant facts about the occurrence of the illness in this experience. One approach to this is to employ a simple *census*, that is, to ascertain all of the relevant facts on all members of the study population, as is commonplace with populations (cohorts) involved in clinical trials. An alternative to this approach is one of combining census and sampling. First one uses a census of the base population as to outcome—to identify all cases. Then a second census is conducted on the cases to ascertain other facts (concerning the determinants, modifiers and confounders) on them. Finally, a sample of the base is used to obtain information of the latter type about it. This alternative to the census approach may be considered the *census–sample* or *case–base* approach [5]. It may also be termed the *case–referent* approach [6], since the study base, which the sample represents, is the direct referent of the empirical pattern of occurrence in the study. On the other hand, the term "case–control" approach is a misnomer, as the base sample is no more a control series than a census of the base (referent) is.

The "case–control" term is, I believe, a reflection of the "trohoc" fallacy of the essence of this type of study. It reflects the misguided notion and practice of comparing cases with noncases in "case–control" studies. If a census of the base were used, then, in a simple situation the concern would be to compare the index rate $r_1 = c_1/B_1$ with the reference rate $r_0 = c_0/B_0$, c_i and B_i denoting the number of cases and the size of base segment representing the ith category of the determinant. By no means would the natural comparison be between the case series and the base as to their distributions by the determinant. If a sample of size $b = b_1 + b_0$ is drawn from the base $B = B_1 + B_0$ so as to estimate the relative sizes of B_1 and B_0, then $r_1' = c_1/b_1$ and $r_0' = c_0/b_0$ are stochastically proportional to r_1 and r_0, so that the empirical rate ratio is estimable as $(c_1/b_1)/(c_0/b_0)$. The point is that the contrast is between the index and reference categories of the determinant regardless of whether a census or a sample of the base is used. And the utility of appreciating this lies in its accent on the study base—the population experience of which the reference series is to be a representative sample (as to the distribution of the determinant).

Summarizing, the "case–control" study is not the alternative, or even an alternative, to the cohort study. A cohort study involves a cohort as its base population, and its

alternative is the use of a dynamic population. A "case-control" study involves a census-sample—case-base or case-referent—strategy of ascertaining the facts about the study base, and its alternative is a simple census.

DEFINITION OF THE STUDY BASE

With the "case-control" study viewed as a matter of census-sample approach to fact-finding about the study base, valid selection of subjects (case census and base sample) presupposes understanding of the definition of the study base, notably a duality in it. This duality corresponds, roughly, to the common distinction between "population-based" case-referent studies on the side and "hospital-based" ones on the other [3], but it needs clarification. For, in terms of the conceptualization of the essence case-referent studies proposed here, all of them are population-based—having to do with the experience of the base population over the time-frame of the study.

The base may be demarcated *a priori*—as, for example, the population (dynamic) of a particular metropolitan area over a particular period of calendar time [7]. With such a definition—a *primary* definition—of the base, the cases of interest are defined, secondarily, as the entirety of cases arising from the base so defined. The challenges are to ascertain those cases on a census basis and to obtain a proper sample of the base itself.

Alternatively, the *cases* may be defined *a priori*—as, for example, those appearing in a particular hospital over a particular span of calendar time. Such a case series is, I propose, to be thought of as a census of cases in the corresponding base by definition—with the definition of the base thus *secondary* to the case selection. The justification for this proposition is the imperative that the case series and the base sample be representative of the same population experience, that is, that they be coherent. With a census ascertainment of cases achieved by definition, the challenge is proper definition of the secondary base inherent in the case enrollment—and, thereupon, its proper sampling.

Given that the case series is the totality of cases in a secondary base, such a base must be *the* population experience (the entirety of it) in which *each* potential case, had it occurred, *would* have been included in the case series. The introductory example serves as an illustration of this. For the cases of traveller's diarrhea identified in an Acapulco hospital over a particular span of time the corresponding base is the experience, over that time span, of the population in which each potential case of traveller's diarrhea, had it occurred, would have appeared in the hospital and would have been enrolled in the case series. In other words, it is the experience of the hospital's catchment population for traveller's diarrhea over the period of case accrual. This definition of the base makes it obvious that neighbors and siblings are unlikely to be even members of the study base (secondary), let alone representative of it (as to tequila use).

Among actual studies, exceptionally illustrative is the International Agranulocytosis and Aplastic Anemia Study [8]. Enrolled are two types of case—the one that is admitted to hospital because of the disease, and that which develops during hospitalization. For the cases of the former type the corresponding population experience (study base) is that of the catchment population (for agranulocytosis and aplastic anemia) of the participating hospitals over the period of case enrollment—an out-of-hospital population experience. By contrast, for the cases developing during hospitalization the base is the in-hospital experience of all patients in whom such development would have been diagnosed in the participating hospitals over the study period, regardless of the admission diagnosis. Thus, for the two types of case, the respective base samples must be representative of the hospitals' catchment populations and their monitored patients, respectively. These statements imply that in that study the primary commitment was to hospital-identified cases, and that the challenges were to properly define, and then to properly sample, the corresponding base experiences, with those definitions secondary to the means of case

primary definitions for them—the experience of the population of West Berlin over the study period, for example. Thus the challenge was to identify all relevant cases of the two illnesses arising from those population experiences, and monitoring of all hospitals in the areas involved was judged to afford a reasonable approximation to the desired census ascertainment of such cases. It deserves particular note that even if the base population is of the primary type, it by no means needs to be "the general population" (whatever may be the meaning of this phrase). This principle is well ingrained in the clinical trial paradigm. In these trials one studies the incidence of health outcomes in relation to treatment—not in the "general population" nor the "general patient population" but in the particular type of patient enrolled in the trial. It accords, as well, with the outlook in laboratory science: nobody demands that the study population be representative of "the general rat population" or its counterpart in some other species; what matters is that one is clear on what *particular* type of population it is.

It is worthy of further note that the definition of the study base, be it primary or secondary, does not restrict the population to people "at risk" for the "exposure" (index category of the determinant under study) nor to those "at risk" for the illness at issue. If everyone in the study population were predestined to their status of "exposure" or "nonexposure," this would be no worse than "self-selection" in the context of being "at risk," especially if predestination were based on randomization (by the Lord). As for the risk of the illness, there is no imperative to have the study population consist of individuals of nonzero risk. The point is, instead, that the inclusion of people of *known* zero risk is a matter of waste and/or obfuscation.

Finally, the concept and definition of the study base are critical to the understanding of whether the case series should be representative of all cases [2, 4], or whether instead this demand is "misplaced" [9]. As long as "all cases" means all cases arising from the study base, this quality is inherent in the census of cases—census ascertainment in the context of a primary base, and census-by-definition with a secondary one.

VALID SELECTION OF SUBJECTS

As was noted above, the overriding principle of subject selection in case-referent studies is that the case series and referent sample be representative of the same base experience.

When the base is defined in *primary* terms, the challenges are, as was noted, to obtain a census of cases in it together with a sample of the base itself, with the latter representative of the base as to its distribution by the determinant(s) of interest. The particulars of complete case ascertainment are matters of procedural technics etc. and the attainment of valid sample of such a base is a matter of sampling theory in general. Both of these topics are outside the scope of the presentation here.

When the base definition is secondary to case selection, and the case series thus valid by definition, the challenge is valid sampling of the base. This sampling is guided by the definition of the secondary base. Thus, as a first example, for the population experience (population time) formed by people who *would* have come to the Acapulco hospital had traveller's diarrhea occurred in the time period of case accrual, an appropriate sample is constituted by people who *did* come there due to a condition which is known to be interchangeable with traveller's diarrhea as a reason for ending up in the Acapulco hospital, and whose occurrence is known to be unrelated to the use of tequila. Similarly, as a second example, for cases of agranulocytosis appearing in the study hospitals because of this disease, a suitable sample of the base (in respect to recent drug use) is patients who did come to those hospitals for some other acute condition which is known to be referred to the hospitals in the same circumstances (of geographic location etc.) as agranulocytosis is, and whose occurrence is known to be unrelated to the use of the drugs at issue. Neighbors, when not travelling, are likely to be members of this base population during the study period, but their histories of recent drug use are not representative of those in the study base if the neighbors' histories are taken under circumstances that tend to be atypical in terms of recent drug use. In particular, if the history of recent exposure is taken

as of the time of home interview, arranged at the neighbor's convenience, the history of the use of analgesics hours or days earlier is unlikely to be typical of such time segments in the total period of the neighbor's membership in the study population: he/she may be available at a particular time because a condition leading to analgesic use is present at that time, or because *it* is absent. As a third and final example, for cases of agranulocytosis developing in the participating hospitals (after admission), the reference series must be selected from the in-hospital population, without regard for the reason for hospitalization —in sharp contrast to the reference series for the cases who were hospitalized with, and because of, agranulocytosis.

As has been implied, in the selection of a hospital reference series for cases incident outside of the registry, the need is to be able to defend the diagnostic inclusions, not exclusions. The criteria for admissible diagnostic entities, given above, have to do with diagnostic entities as they represent the reason for entry into the hospital rather than incidental conditions [9]. In order that such a condition (primary diagnosis) be, as a reason for hospitalization, similar to the illness under study, it is commonly important to choose diagnostic entities for which hospitalization is equally elective, or obligatory, as for the illness under study. In respect to incidental (secondary) diagnoses, the admissibility (non-exclusion) criteria for the reference subjects and the cases must be the same, as these series represent, respectively, the study base itself and the cases that develop in it.

If the cases are identified from a registry of deaths, then the selection of reference subjects must be guided by principles completely analogous to those guiding the selection of hospital reference subjects for cases identified from a hospital. The main principle is that for the cases (of death) so identified the corresponding secondary base is the catchment population of the registry for those deaths—the experience of an out-of-registry, living population. Thus, a proper sample of it is not, generally, deaths from all other causes indiscriminately but, insofar as a registry series is to be used at all, deaths from causes whose occurrence is unrelated to the determinant under study [11].

Thus far, in the context of a secondary base, the focus has been on the selection of the reference series—on the premise that the case series is valid by definition. It is important to note, however, that the feasibility of finding a proper base sample depends on how the case series is defined, because the definition of the secondary base is inherent in case selection. Consequently, the attainment of the cardinal condition of validity—coherence of the numerator (case) and denominator (reference) series in terms of the population experiences represented—may be enhanced by care in the definition of the case series itself. In particular, it is commonly helpful to restrict case (and, secondarily, reference subject) admissibility according to area of residence. The more the admissibility is restricted to the vicinity of the source of cases (hospital, say), the more likely it is that the corresponding reference subjects, had they developed the illness under study, would have presented themselves to the source, thus ensuring, at least, that they are members of the base population. This, in turn, enhances the likelihood that the reference series is representative of the base.

Whatever has been said here about representativeness refers to the distribution of the determinant(s) conditional on subject characteristics controlled in the analysis of the study. Thus, if the reference series is unrepresentative of the age—and thereby determinant—distribution of the base at large (merely by virtue of being a hospital series or as a result of matching by age), the imperatives of representativeness (concerning distribution of the determinant in the base) may still be satisfied conditionally on age; and with age controlled in the analysis, validity is maintained. The requirement of conditional validity here is analogous to that of conditional simple random sampling in the context of stratified sampling.

The issue of validity in the selection of subjects into a case-referent study is not simply a matter of coherence between the numerator (case) and denominator (reference) series as to what they represent—as census and sample, respectively. To be coupled with the imperative of representativeness of the same base is that of *comparability*, unrelated to