Biostatistics Simplified

Sanjeev B Sarmukaddam



JAYPEE

Contents

1. **Introduction** **1**
2. **Data Description** **11**
3. **Sample to Population** **57**
4. **Study Designs** **95**
5. **Confidence Intervals and Logic of Tests of Significance** **169**
6. **Inference: Quantitative Variables** **202**
7. **Inference: Qualitative Variables (Categorical Data Analysis)** **237**
8. **Correlation and Regression** **261**
9. **Vital Statistics** **302**
10. **Miscellany** **335**

Appendix *397*

Index *407*

4

Study Designs

EPIDEMIOLOGY

Epidemiology is concerned with the distributions and determinations of disease frequency in human populations. The basic design strategies used in epidemiologic research can be broadly categorized according to whether such investigations focus on describing the distributions of disease or elucidating its determinants. Descriptive epidemiology is concerned with the distribution of disease, including consideration of what populations or subgroups do or do not develop a disease, in what geographic location it is most or least common, and how the frequency of occurrence varies over time. Information on each of these characteristics can provide clues leading to the formation of an epidemiologic hypothesis that is consistent with existing knowledge of disease occurrence. Analytic epidemiology focuses on the determinants of a disease by testing the hypothesis formulated from descriptive studies, with the ultimate goal of judging whether a particular exposure causes or prevents disease.

JAYPEE BROTHERS

Descriptive epidemiology is concerned with describing the general characteristics of the distribution if a disease, particularly in relation to person, place and time. Descriptive data provide valuable information to enable health care providers and administrators to allocate resources efficiently and to plan effective prevention or education programs. In addition, descriptive studies have often provided the first important clues about possible determinants of a disease. Due to limitations inherent in their design however, descriptive studies are primarily useful for the formulation of hypotheses that can be tested subsequently using an analytic design.

There are three main types of descriptive studies, which are listed in Table 4.1. The first type, *correlation study,* uses data from entire populations to compare disease frequencies between different groups during the same period of time or in the same population at different points in time.

Table 4.1: Overview of epidemiologic design strategies

- *Descriptive studies*
  - Populations (correlational studies)
  - Individuals
  - Case reports
  - Case series
  - Cross-sectional surveys
- *Analytic studies*
- *Observational studies*
  - Case-control studies
  - Cohorts-studies – retrospective and prospective
- *Intervention studies (clinical trials)*

JAYPEE BROTHERS

The *case report* is the most basic type of descriptive study of individuals, consisting of a careful, detailed report by one or more clinicians of the profile of a single patient. The individual case report can be expanded to a *case series,* which describes characteristics of a number of patients with a given disease. The third type of descriptive epidemiologic study design in individuals is the *cross-sectional* survey, in which the status of an individual with respect to the presence or absence of both exposure and disease is assessed at the same point in time.

All study designs involve some implicit (descriptive) or explicit (analytic) type of comparison of exposure and disease status. In a case report, for example, where a clinician observes a particular feature of a single case, a hypothesis is formulated based on an implicit comparison with the "expected" or usual experience. In analytic study designs the comparison is explicit, since the investigator assembles groups of individuals for specific purpose of systematically determining whether or not the risk of disease is different for individuals exposed or not exposed to a factor of interest. It is the use of an appropriate comparison group that allows testing of epidemiologic hypotheses in *analytic study designs.*

There are a number of specific analytic study design options that can be employed. These can be divided into two broad design strategies: observational and interventional. The major difference between the two lies in the role played by the investigator. In observational studies, the investigator simply observes

JAYPEE BROTHERS

the natural course of events, noting who is exposed and nonexposed and who has and has not developed the outcome of interest. In intervention studies, the investigators themselves allocate the exposure and then follow the subjects for the subsequent development of disease.

Observational Studies

There are two basic types of observational analytic investigations—the *cohort study* and the *case-control* study. In theory, it is possible to test a hypothesis using either design strategy. In practice, however, each design offers certain unique advantages and disadvantages. In general, the decision to use a particular design strategy is based on features of the exposure and disease, the current state of knowledge, and logistic considerations such as available time and resources.

In a cohort study, subjects are classified on the basis of the presence or absence of exposure to a particular factor and then followed for a specified period of time to determine the development of disease in each exposure group. In a case-control study, a case group or series of patients who have a disease of interest and a control, or comparison, group of individuals, without the disease are selected for investigation and the proportions with the exposure of interest in each group are compared.

Considerable confusion arises concerning the terms *retrospective* and *prospective* as applied to epidemiologic studies. Some investigators have used these terms synonymously with case-control and cohort,

JAYPEE BROTHERS

respectively, reasoning that the former looks backward from a disease to a possible cause, while the latter looks forward from an exposure to an outcome. It is more informative to use these terms to refer to the temporal relationship between initiation of the study by the investigator and the occurrence of the disease outcomes being studied. The feature that distinguishes a prospective from a retrospective cohort is simply and solely whether the outcome of interest has occurred at the time the investigator initiates the study. Thus, at the beginning of a prospective cohort study, the groups of exposed and unexposed subjects have been assembled, but the disease has not yet occurred, so that the investigator must conduct follow-up during an appropriate interval to ascertain the outcome of interest.

It is often possible to investigate a particular hypothesis using either a case-control or cohort study design. For example, the hypothesis that oral contraceptive use increases the risk of breast cancer has been evaluated in number of case-control studies that identified women with and without breast cancer and compared the proportions who were users of oral contraceptives. In addition, the question has been examined using a cohort design as well where women initially free from disease were classified according to their use of oral contraceptives and then followed forward over time to compare the development of breast cancer in the two groups.

The choice of which type of design to use to study a particular exposure-disease relationship depends on

JAYPEE BROTHERS

the nature of the disease under investigation, the type of exposure, and the available resources. For example, the case-control design is particularly efficient for investigation of a relatively rare disease since it selects a group of individuals who has already developed the outcome. In the example of artificial sweeteners and bladder cancer, it would have been much more costly and time-consuming to use a cohort approach involving identification of a necessarily very large group of individuals who used artificial sweeteners and a comparable group of nonusers and follow them to compare how many patients would subsequently develop bladder cancer over the next 20 years. Conversely, since cohort studies enroll individuals who are initially healthy and observe the subsequent development of disease of time, this design is best suited to investigations of relatively common outcomes that will accrue in sufficiently large numbers over a reasonable short period of follow-up. In general, therefore, cohort studies of rare diseases are far less feasible.

Intervention studies, also referred to as experimental studies or clinical trials, may be viewed as a type of prospective cohort study, because participants are identified on the basis of their exposure status and followed to determine whether they develop the disease. The distinguishing feature of the intervention design is that the exposure status of each participant is assigned by the investigator. When well-designed and conducted, intervention studies can indeed provide the most direct epidemiologic evidence on which to judge whether an exposure cause or prevents a disease.

JAYPEE BROTHERS

Other Operational Details: Cohort Study

Because participants are free from the disease at the time their exposure, the temporal sequence between exposure and disease can be more clearly established. They are particularly well-suited for assessing the effects of rare exposures and allow for the examination of multiple effects of a single exposure.

Cohort studies may be classified as either prospective or retrospective, depending on the temporal relationship between the initiation of the study and the occurrence of the disease. By definition, both prospective and retrospective cohort study designs classify subjects in the study on the basis of presence or absence of exposure. In retrospective cohort studies, however, all the relevant events (both the exposures and outcomes of interest) have already occurred when the study is initiated. In prospective studies, the relevant exposures may or may not have occurred at the time the study is begun, but the outcomes have certainly not yet occurred. Retrospective cohort studies usually evaluate exposures that occurred many years previously, they depend on the routine availability of relevant exposure data in adequate detail from pre-existing records.

In cohort studies in which a single, general cohort is entered and its members then classified into exposure categories, an internal comparison group can be utilized. That is, the experience of those cohort members classified as having a particular exposure is compared with that of members of the same cohort who are either nonexposed or exposed to a different

JAYPEE BROTHERS

degree. For cohort studies that involve the use of a special exposure group, such as in an occupational setting or for a particular environment, it is often not possible to identify a portion of the cohort that can safely be assumed to be nonexposed for comparison. In this instance, an external comparison group is used, such as the general population of the area in which the exposed individuals reside.

The major disadvantage of using the general population as a comparison group is that its members may not be directly comparable to those of the study cohort. People who are employed are on average healthier than those who are not. Since the general population includes people who are unable to work due to illness as well as those who are employed, rates of disease and death among the general population are almost always higher than they are for members of the work force. The effect of this phenomenon, termed the "healthy worker" effect is that any excess risk associated with a particular occupation will tend to be under estimated by a comparison with the general population.

In many cohort studies, a single classification of exposure is made for each individual at the time of his or her entry into the study. Frequently however, changes in exposure levels for the factors of interest will occur during the course of long-term follow-up. Individuals may cut down of their smoking habits, change jobs, or begin to eat a diet lower in saturated fats. Similarly, the introduction into the workplace of a new piece of standard equipment may affect the level

JAYPEE BROTHERS

of exposure experienced by all workers in a single plant or entire industry. Such changes will tend to result in an underestimate of the true strength of the association between an exposure and disease. Consequently, many cohort studies are designed to allow for periodic re-examination or resurvey of the members of the cohort to allow for revision of exposure categories according to the new information. The analysis can then take into account the reasons for these changes.

Whatever the procedures for identifying the outcomes of interest, it is crucial for the validity of the study that they are equally applied to all exposed and nonexposed individuals. As with any epidemiologic investigation, an evaluation of the validity of a cohort study requires consideration of the chance, bias, and confounding as alternative explanations for the study finding.

A principal advantage of cohort studies is that they are optimal for investigation of the effects of rare exposure. A second advantage of cohort studies is their ability to examine effects of a single exposure, thus providing a picture of the range of health outcomes that could be related to a factor or factors of interest. Third, since the participants are disease-free at the time exposure status is identified, the temporal sequence between exposure and disease can be more clearly elucidated. Moreover, since in a prospective cohort study the outcomes of interest have not yet occurred the time the study is begun, bias in the selection of subjects and ascertainment of exposure is minimized. For retrospective cohort studies where all the relevant

JAYPEE BROTHERS

events have occurred when the study begins, the potential for these biases is similar to that of a case-control study. Finally cohort studies allow the direct calculation of incidence rates of the outcomes under investigation in the exposed and nonexposed groups.

With respect to disadvantages, a cohort study is not an efficient design for the evaluation of a rare outcome unless the study population is extremely large or the outcome is common among those who are exposed. Moreover, if prospective, cohort study is very expensive and time-consuming compared with either a case-control or retrospective study. Finally, the unique potential for bias in the cohort design is the problem of losses to follow-up. If the proportion of those lost to follow-up is high or even if the proportion is low but is related to both, the exposure and outcome under investigation, the study findings may not be valid.

Other Operational Details: Case-control Study

The case-control design offers a solution to the difficulties of studying disease with very long latency periods, since investigators could identify affected and unaffected individuals and then look backward in time to assess their antecedent exposure rather than having to wait a number of years for the disease to develop. For a case-control study to provide sound evidence of whether there is a valid statistical association between an exposure and disease, comparability of cases and controls is essential. Important issue to be considered in the design of a case-control study is the definition

JAYPEE BROTHERS

of the disease or outcome of interest. It is important that this represent as homogeneous a disease entity as possible since often similar manifestational entities of disease have very different etiologies. To help ensure that cases selected for study represent a homogeneous entity, one of the first tasks in any study is to establish strict diagnostic criteria for the disease.

Depending on the certainty of the diagnosis, and the amount of information available, it is often useful to perform analyses separately for cases classified as definite, probable or possible. Once the diagnostic criteria and definition of the disease have been clearly established, the individuals with this condition can be selected from a number of sources. These include identifying persons with the disease who have been treated at a particular hospital or medical care facility during a specified period of time or selecting all persons with the disease in a defined general population at a single point or during a given period of time. The first approach, referred to as a hospital-based case-control study, is more common because it is relatively easy and inexpensive to conduct. The second, the population based case-control study, involves locating and obtaining data from all affected individuals or a random sample from a defined population. The scientific advantage of a population-based design is that it avoids bias arising from whatever selection factors lead an affected individual to utilize a particular health care facility or physician. Moreover, the population-based approach allows the description of the entire of the disease in that population and the

JAYPEE BROTHERS

direct computation of rates of disease in exposed and nonexposed individuals. Because the logistic and cost considerations are often prohibitively large, however, population based case-control studies are not routinely done.

Regardless of the source, the affected individuals can represent either incident (newly diagnosed) or prevalent (existing at a point in time) cases of the disease. Certainly, the inclusion of prevalent cases, especially of a rare condition, will greatly increase the sample size available for study in a given time period. Since prevalent cases reflect determinants of duration as well as the development of the disease, however the interpretation of the findings from such a study may be complicated. Clarification of the temporal sequence between exposure and disease is always an issue in case-control studies, it is a more serious problem when prevalent rather than incident cases are used.

The cases in a case control study should be selected to be representative of all persons with the disease. Rationale for selecting a random sample derives from the fact that findings from a study that utilized representative cases would be more readily generalizable to the broader population of all men. However, the primary concern in the design of any study must be validity, not generalizability. A case-control study can be restricted to a particular type of case on whom complete and reliable information on exposure and disease can be obtained.

The control subjects should be selected to be comparable to the cases, and as a consequence will

JAYPEE BROTHERS

Biostatistics Simplified

Salient Features

- An excellent text
- Covers all aspects of the subject in a concise format
- Explains all fundamental concepts and important terms in a simple and easy-to-understand language
- Presents abundant discussion on the topics like quality control, statistical fallacies, clinical significance, tests of significance, confidence interval, tolerance limits, etc.
- Useful for all medical/health fraternity including nursing and other paramedics.

Sanjeev B Sarmukaddam is a Consultant (Biostatistics), Maharashtra Institute of Mental Health, BJ Medical college and Sassoon Hospital Campus, Pune, Maharashtra (India). He has taught Biostatistics in many Medical Colleges/Institutions and Universities (including National Institute of Mental Health and Nuero Sciences, Bengaluru and Institute for Research in Medical Statistics, New Delhi, as permanent faculty member for 3 years each and at Pune University as guest faculty for 5 years). On special invitation, he also lectured at Lawrence Education Center of Borgess Medical Center (arranged by Department of Community Health) of University of Michigan, USA in their 'grand rounds'. His presentation during International Biometric Conference held at McGill University, Montreal, Quebec, Canada and at many places of academic interest in London, UK was appreciated. He has also worked in a tribal area for 2 years. He has more than 40 scientific publications in peer reviewed international journals. He co-authored a book titled "*Medical Biostatistics*" published in 2001 by a New York (USA) based publisher Marcel Dekker, Inc. He authored another book entitled "*Fundamentals of Biostatistics*" in 2006.

He has been awarded a gold medal and Prof BG Prasad award for best published paper of the year (1987) on Statistics Applied to Epidemiology. He was involved in planning, monitoring, analyzing and report writing activities of many national level studies/trials and sample surveys. For more information he can be contacted at his e-mail address ssanjeev2004@vsnl.net

Available at all medical book stores
or buy directly from Jaypee Brothers through online shopping at www.jaypeebrothers.com

or call + 91-11-32558559

JAYPEE BROTHERS
Medical Publishers (P) Ltd.
www.jaypeebrothers.com

ISBN 978-81-8448-748-0