Observational studies: How to go about them?

Maninder Singh Setia
Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada

Address for correspondence: Maninder Singh Setia, MD MPH Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montreal – QC Canada H3A 1A2. Email: maninder.setia@mail.mcgill.ca

INTRODUCTION

As dermatologists, we are often involved in clinical research, either during our residency or later in academics and clinical practice. Clinical research involves a few major steps: identifying the research question, review of the literature, designing a study, ethical conduct of the research, analysis of the collected data, and publication of results. Each and every step is equally important for conduct of good research. A badly designed and conducted study will provide biased estimates, and cannot be salvaged at the analysis stage. Hence, it is imperative to start thinking about all these issues when the research is being conceived and not wait to address them later in the process. The present article builds on previous articles published in the Residents' page on research methodology and discusses observational studies in clinical research.

Clinical studies can be classified into two main categories: experimental and observational. In experimental studies, the investigator intervenes to change the exposure status in the population under study and assesses the change in outcome. The main forms of experimental studies are randomized clinical trials and other large controlled trials, which may not be randomized (for example, field trials). In observational studies, the investigator does not actively intervene to change the exposure status, but merely observes the population under study. The current article focuses on three types of observational studies: cross-sectional studies, case-control studies, and cohort studies. It addresses the methods involved, advantages and potential limitations, and finally briefly mentions the types of analyses used in these methods.

CROSS-SECTIONAL STUDIES

Design

It is one of the most common types of study conducted in clinical settings. A cross-sectional study is a one-time snapshot study. The investigator usually selects a sample (random, systematic etc.) from the target population. The exposure of interest (X), the disease of interest (Y) and other covariates (*e.g.*, age, gender, socio-economic conditions and other potential risk behaviours) are measured in the selected population at one point of time. Finally, the association between X and Y is calculated.

Example of a cross-sectional study

Let us design a study to assess the association of cholesterol and psoriasis. We include about 250 subjects (assuming that this sample size is adequate for the research question) from a Dermatology clinic. We assess the presence of psoriasis and measure cholesterol levels in all of them. We also record other features (demographics, socio-economic conditions, alcohol use etc). The analyses of the data collected will tell us the prevalence of psoriasis and hypercholesterolemia in these subjects and their association.

Strengths, limitations, and biases

Cross-sectional studies are useful to study exposures that do not change over time (*e.g.,* blood group). They are often cost- and time-efficient. They can be used to generate new hypotheses for future research and are often conducted before initiating a clinical trial or a longitudinal study.

However, there are certain limitations of these studies - one usually cannot comment on the temporality of events

How to cite this article: Setia MS. Observational studies: How to go about them? Indian J Dermatol Venereol Leprol 2008;74:288-91.
Received: December, 2007. **Accepted:** December, 2007. **Source of Support:** See Acknowledgement. **Conflict of interest:** None declared.

in cross-sectional studies. In the above example, if we find that hypercholesterolemia is associated with psoriasis, we cannot comment on the sequence of events; whether hypercholesterolemia preceded psoriasis or vice versa. This is a potential limitation of cross-sectional studies. The assessment of exposure in this type of study does not necessarily distinguish between its etiological or prognostic roles. These studies are not very efficient to assess rare diseases. One can often only measure the prevalence of disease or exposure and not the incidence in these types of studies (e.g., HIV ELISA, VDRL, HSV 2 IgG, cholesterol levels). Hence, the term 'incidence' should be avoided while presenting results from these cross-sectional studies. Further, it is appropriate to report that X and Y are associated rather than X causes Y from a single cross-sectional analysis.

In a cross-sectional study, long duration cases will be over represented and short duration cases (mild or very severe leading to mortality) will not be adequately represented. This is called the length-time bias and one has to be aware of this while interpreting results from a cross-sectional study. The measure of association in cross-sectional studies is the odds ratio.

CASE-CONTROL STUDIES

Design

These studies are very popular with clinical researchers. The investigator selects two groups of individuals: diseased (cases) and non-diseased (controls). The exposure is then measured in both the groups and the association of exposure and disease is calculated. The cases and controls have to be selected from the same underlying population (*e.g.*, the same catchment area of the hospital).

Example of a case-control study

For example, let us design a study to assess the association of smoking and psoriasis. We select psoriasis patients who present to the clinic as cases (according to the calculated sample size). As controls, we select individuals with no history of psoriasis but who do present to the clinic with other complaints (such as dermatophyte infection). The next step is to collect data on the current and past smoking habits of these subjects and on other factors (age, gender, and other risk behaviours). Finally, the association between smoking and psoriasis is calculated.

Matching in a case-control study

The controls are usually matched to the cases for a few characteristics, *e.g.*, sex, age (±2 years of age). This process of matching can be efficient in case-control studies because a comparison group similar to the index group is selected. Usually, one control is selected per case; however, increasing the number of controls per case can increase the power of the study. Pragmatically, it is sufficient to select four controls per case.

Matching can have its disadvantages: it may be expensive or time-consuming to find a good matched control for the case. If there are no appropriate controls for the identified case, then the case is lost and cannot be included in the study. It may be difficult to account for the effects of the matching variable on the outcome of interest.

Selection of controls

Hospital controls

In the above example, we selected the cases and controls in a hospital setting. Hospital controls are easy to recruit, relatively easy to collect biological samples from if needed, and can provide comparable information. However, one has to be careful while selecting hospital controls. If our controls have medical conditions that are related to the exposure of interest, then it could lead to biased estimates. In the above example, if individuals with cardiovascular problems are chosen as controls (known association with smoking) then a bias is introduced in the selection process. The prevalence of exposure in the controls may not be representative of the population at large.

One should be aware of the population attending the hospital while selecting cases and controls. If we are interested in evaluating a rare disease and the hospital is a referral centre for this condition, it may be difficult to select the cases and controls from the same study base. The cases will be referred from all regions to this hospital, whereas the controls will be from the usual catchment area of the hospital. These issues have to be addressed while designing a case-control study in a hospital setting.

Other controls

There are other ways of selecting controls: friends controls, neighborhood controls, or controls from the general population in the same catchment area as the cases.

Although friends may be easy to access, it is possible that they may have similar exposure distribution (particularly for exposures such as alcohol, smoking, physical activity, socio-economic conditions etc). Hence, this might not be representative of the general population. Population controls from the same catchment area are good for these studies but very difficult to access and practically the most difficult group, particularly in our research settings.

Strengths, limitations, and biases

Case-control studies can be often completed in a short duration and are often less expensive than other forms of observational studies. The study population is sampled on the basis of disease status; hence, they can be used to study rare diseases and chronic diseases with variable latent periods. Thus, the case-control study design could be a useful tool in studying dermatological conditions.

As with other studies, there are certain biases one has to be aware of in these studies. Exposure information is collected after the occurrence of disease (in cases). Such individuals with the disease (cases) are more likely to recall past habits and exposures compared with those without the disease (controls). This may introduce a recall bias, leading to overestimation of the association between the exposure and the outcome.

There could be an interviewer bias or observer bias in ascertaining the exposure if the outcome is known (*i.e.*, the disease state is known). For example, many case-control studies were done to study the role of BCG in leprosy. Researchers identified cases of leprosy and controls without leprosy. They ascertained the exposure by looking at the BCG scar in both. It is known that sensitivity of reading a BCG can vary. Observers who are aware of the study hypothesis and disease status of the study subjects are more likely to report the presence of a scar in nonleprosy patients than in leprosy patients. Thus, the protective effect of BCG in leprosy will be overestimated (a biased estimate).

These biases can be minimized by adequately blinding the observers to the outcome status and standardizing the criteria for evaluation of exposure in both cases and controls. The measure of association in a case-control study is the Odds Ratio. The details of these ratios (calculation and interpretation) in these studies will be discussed in a subsequent article.

COHORT STUDIES

Design

Cohort: *a group of soldiers in Latin (cohors)* - these studies can be prospective or retrospective. In these studies, the subjects are selected on the basis of their exposure status.

Prospective studies

In a prospective cohort study, a group of individuals who do not have the disease (but may or may not have the exposure) are selected and followed over a period of time. In such a cohort study, the exposure in these individuals is assessed at baseline (entry into the study) and subsequently, at regular time intervals for the entire duration of the study. The follow-up period (*e.g.*, a year or five years) usually depends on the latency period of the disease of interest. The occurrence of the disease (new occurrence) in these individuals is assessed during the follow-up period. Finally, the association between the disease and the exposure is calculated.

It is very important that we define the exposure and outcome clearly. Often, it is easy to define the exposure; for example, blood group, some drug exposure, teratogen exposure at birth. It may be more complicated in other scenarios, such as smoking status (how much, how frequent, etc), alcoholism status or physical activity. Similarly, the outcome has to be clearly defined at the beginning of the study (*e.g.*, is depigmentation the end point or will hypopigmentation be sufficient as the end point?).

Retrospective studies

In a retrospective cohort study, the information on the exposure and disease is already collected (usually a part of another study or medical records). The investigator uses this existing information to evaluate the relation between exposure and disease over a period of time. A variant of this is when the initial part of the study involves analyses of the data already collected and the subsequent part involves follow-up of the same subjects over time to assess the occurrence of new outcomes.

Example of a cohort study

We are following a cohort of HIV-infected men presenting in the clinic regularly over the past five years. We have evaluated them at baseline, collected information on demographics and socio-economics, sexual behaviors, opportunistic infections and other diseases, CD4 counts, viral load, and other blood parameters. Initially, we analyzed the existing data to calculate the association between CD4 counts or viral load and the occurrence of various infections and diseases. Following our retrospective analyses, we continue to follow these individuals. Some of them receive a treatment regimen with non-nucleoside reverse transcriptase inhibitors; others receive a protease inhibitor in their therapy regimen. We follow them prospectively to assess which group shows other clinical conditions and changes in CD4 counts and viral loads as per our research hypothesis. This will be an example of the variant of the retrospective studies.

Strengths, limitations, and biases

In cohort studies, the temporal relation is clear as the exposure is measured before the disease (X causes Y). We can assess the role of multiple exposures in the outcome (*e.g.*, smoking, alcohol, contraceptive use). Often we can also assess the role of the same exposure for multiple outcomes. We can calculate the incidence of disease in these studies. However, they are expensive to conduct and time-consuming.

There could be losses to follow-up in prospective cohort studies; this may lead to bias if the loss to follow-up is due to the exposure. These studies are not very efficient if the outcome is a rare disease. In retrospective cohort studies, the investigator has to rely on the information already collected. There could be missing data or missing records. The measurement of exposure or outcome may not be by the most appropriate method. However, retrospective studies are less expensive and relatively quicker to conduct.

In prospective cohort studies, observers who are aware of the study hypothesis and exposure status may be more likely to confirm it in exposed individuals compared with nonexposed ones - the observer's bias. This may be more relevant for soft health outcomes (intensity of pain, improvement in a lesion, mental condition etc) or if the outcome is not clearly defined. We can calculate the incidence ratio (or incidence rate) in cohort studies as a measure of association.

There are multiple variants and exceptional situations for each of these studies; the information on these variants can be obtained from the references at the end of this article.

The purpose of this article was to provide an overview of these different types of observational studies. As mentioned earlier, one should choose an appropriate design to conduct research. The choice of this design will depend on the study hypothesis, and on the type of exposure and disease. It also depends on the time and funds available to conduct the study (particularly in case of postgraduate dissertations). Often, in academic settings time and funds are sought after designing the study.

The golden rule is to design the study well. All the potential limitations and biases in each study design and the methods to address them should be considered before initiating the study. It will be appropriate to reiterate here that one should not attempt to salvage a poorly designed and conducted study by using statistical methods at the end of the study. Finally, the type of study design, methods used to measure the exposure and outcome, and the limitations in the study should be clearly mentioned in the published report.

ACKNOWLEDGMENT

MSS is funded by CIHR-IHSPR Fellowship and CIHR-RRSPQ, Public Health Training Programme for his doctoral studies at McGill University.

REFERENCES

1. Singh G, Kaur V. Formulation of a research project. Indian J Dermatol Venereol Leprol 2007;73:273-6.
2. Zodpey S. Sample size and power analysis in medical research. Indian J Dermatol Venereol Leprol 2004;70:123-8.
3. Rothman KJ, Greenland S. Modern Epidemiology. 2nd ed. Philadelphia: Lippincott Williams and Wilkins; 1998.
4. Szklo M, Javier Nieto F. Epidemiology beyond the basics. 1st ed. Sudbury, US: Jones and Bartlett Publishers Inc.; 2004.
5. Kleinbaum D, Kupper L, Morgenstern H. Epidemiologic research. New York, US: John Wiley and Sons, Inc.; 1982.
6. Jewell N. Statistics for Epidemiology. Boca Raton, US: Chapman and Hall/CRC; 2004.