Design and data analysis 1 study design

Karthik Suresh, Geetha Suresh[1], Sanjeev V. Thomas[2]

Departments of Pulmonary and Critical Care Medicine, Johns Hopkins University School of Medicine, Baltimore, MD,
[1]Justice Administration, University of Louisville, Louiseville, KY, USA, [2]Neurology, Sree Chitra Tirunal Institute for Medical
Sciences and Technology, Trivandrum, Kerala, India

Abstract

This article is intended to give the reader guidance in evaluating different study designs used in medical research for better scientific quality, reliability and validity of their research. This article explains three main types of study designs with understanding examples. Care and caution with skills and experience needed to design suitable studies and appropriate design coupled with reliable sampling techniques and appropriate statistical analysis, supported by clear objectives with decent communication of the findings, are essential for good research.

Key Words

Data analysis, study design, trials

For correspondence:
Dr. Geetha Suresh, Departments of Pulmonary and Critical Care Medicine, Johns Hopkins University School of Medicine,
Baltimore, MD, USA. E-mail: uoflgita@gmail.com

Ann Indian Acad Neurol 2012;15:76-80

Introduction

The term "study design" is not used consistently in the scientific literature. The term is often restricted to the use of a suitable type of study. However, the term can also mean the overall plan for all procedures involved in the study. If a study is properly planned, the factors that distort or bias the result of a test procedure can be minimized. Three broad classifications of medical research study designs are observational studies, experimental studies and metaanalyses.

Observational studies

Observational studies are categorized into four different types: case series studies, case–control studies, cross-sectional studies and cohort studies.

A case series is a study on a group of patients based on an observation of a specific disease. Lack of a control group in this type of study is a major disadvantage. Case series are primarily a descriptive report observed in a group under study.

Despite limitations, case series can often have a significant impact on the current practice of medicine. Consider the report of Kaposi's sarcoma and pneumocystis pneumonia among homosexual men in Los Angeles and New York, first appearing in the Morbidity and Mortality Weekly Report (MMWR) from the Centers for Disease Control in 1981, before the isolation of the human immunodeficiency virus. Of course, more such case series emerged subsequently, leading to the search for the cause of immunodeficiency in these patients. Case series are often used to put together case definitions of new diseases and to define future areas of clinical study.

Access this article online

Quick Response Code:

Website:
www.annalsofian.org

DOI:
10.4103/0972-2327.94987

Case–control studies

In case–control studies, cases (disease present) are compared with controls (disease not present). The controls can be matched to cases on variables only so far as these variables are not actively studied (i.e., one cannot match cases and controls for age, say, if age is included as a variable in subsequent analysis). Figure 1 explains to what extent persons in the case and control groups were exposed to infection (case–control study sampling design).

Researchers using a case–control design normally try to match cases with control groups based on age, gender or medical records. The researcher should make sure that both groups are similar with respect to important characteristics that may otherwise confound the conclusions .

In case–control studies, the most important statistical parameter is the Odds Ratio (OR).[1] Case–control studies usually require less time and fewer resources than cohort studies. The disadvantage of case–control studies is that the incidence rate[2] (rate of new cases) cannot be calculated. There is also a great

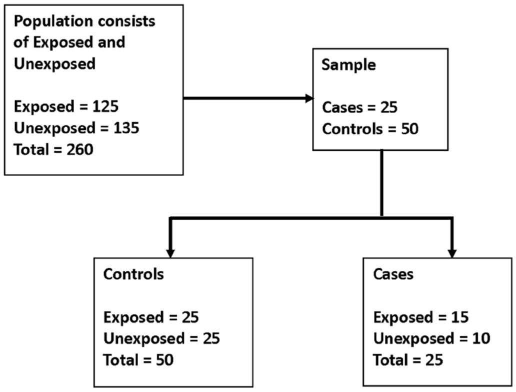


Figure 1: Case–control study sampling design

risk of bias from the selection of the study population ("selection bias[3]") and from faulty recall ("recall bias[4]").

Case–control is an effective strategy when the cases have already been "discovered," leaving the researcher only to establish matched controls. Chalmers *et al*. looked to study the role of past medical and environmental risk factors in the development of various neurologic symptoms in Leber's Hereditary Optic Neuropathy (LHON), a relatively rare disease. Given a group of 50 patients with known LHON, they established 50 control cases for comparison. This allowed the investigators to compare effects of certain environmental factors in their target populations (patients affected with LHON) and use the general, unaffected public as a control.

Like randomized controlled studies and other studies, the number of cases and case–controls is not chosen at random. Consider the study by van der Mei *et al*. entitled "Past exposure to sun, skin phenotype and risk of Multiple Sclerosis." In this study, the authors determined that 200 controls and 100 cases needed to be enrolled so that their previously chosen OR could be achieved. They enrolled 136 cases and 272 controls. It is important to note that the authors had to collect background data on baseline exposure rates (i.e., the percentage of the population that is exposed to the variable that is being studied) before they could determine the number of cases needed and controls needed. Sample size calculation in case–control studies typically requires some knowledge of prevalence of the rates of exposure to risk factors being studied. The number of cases and controls needed also depends on matching status (matched vs. unmatched). In matched studies, increasing the ratio of matched controls to matched cases improves the precision of the OR. When in doubt, using 2:1 or even 3:1 controls to cases ratio is useful, provided appropriate matched controls are available.

Case–controls cannot be used to look for "causality." This is partly due to their retrospective nature, which precludes the investigator from assessing incidence. That is, because the cases in a case–control study have already been diagnosed with the disease under study, it is not possible to establish the rate at which the disease develops between exposure-positive and exposure-negative individuals. Additionally, case–control studies must be adjusted for confounders.[5] There are three major ways to account for confounders: exclusion, matching and statistical adjustment.

By either excluding or matching cases and controls across various confounders, the investigators ensure that either (i) nobody in the experiment (cases or controls) is positive for the confounding variable or (ii) the proportion of patients with confounding variable positivity is uniform across cases and controls.

Matching is typically done to ensure that the case population is similar to the control populations across various variables. In matching, each case is assigned a certain number (usually one, less than three to four) of controls that are individually matched to that case across a preset number of matching variables. None of the variables used to match cases and controls can subsequently be analyzed for relationship to disease. Therefore, it is crucial that factors chosen for matching satisfy the following requirements:

- The matching variable should be associated with both disease and exposure.
- The matching variable should not be on the causal chain. For instance, in studying sunlight exposure relationship to multiple sclerosis (MS), cases and controls cannot be matched for Vitamin D levels as this variable is involved in the causal chain under investigation.
- The matching variable's impact on disease should not be of interest to the investigator. Additionally, the matching variable should not be strongly associated with a variable that is being studied. If this occurs, the cases and controls will be inadvertently matched on two variables (the matching variable and the investigational variable that happens to be closely associated with the matched variable).

One disadvantage of matching is the ability to find proper controls. As the number of matching variables increases, it becomes harder to find matched controls. In this case, adjustment for variables left unmatched can be done with statistical analyses. Another disadvantage is the loss of efficiency; in the effort to match, cases and matched controls may become too similar, biasing the results toward the null hypothesis (no difference between studied groups).

In summary, case–control studies have several advantages including feasibility and relative ease of case identification and control selection, but must be carefully designed with regards to sample size, matching and identification of potential confounders.

Cross-sectional studies

The third type of observational studies consists of cross-sectional studies, surveys, epidemiologic studies and prevalence studies. Cross-sectional studies analyze the data at one particular point of time. Figure 2 explains the cross-sectional sampling design.

Figure 2 explains prevalence of disease and random selection method among the study population. Ten randomly selected members from the study population have disease and the remaining 40 selected have no disease. From this sample, we can make a point estimate of the prevalence of the disease.

The prevalence in our sample is 20% and, therefore, our point estimate of the prevalence in the population is 20%.

Environmental and cross-sectional studies are often used to test popular hypotheses about certain risk factors and their relationships to disease. For instance, in order to investigate the theory that the incidence of MS has a latitudinal gradient (with higher incidences farther away from the equator), Alonso *et al.* conducted a metaanalysis (a different type of study) of a group of cross-sectional studies looking at risk factors for the development of MS. This is therefore a study of a group of cross-sectional studies. Each cross-sectional study collected large populations of MS patients and compared their risk factors, demographics and disease type at one point in time. Unlike other studies, there is no "follow-up" in cross-sectional studies. Patient outcomes over time cannot be measured in these trials.

Cohort studies

A cohort is a group of people who have something in common. In medical research, the cases in cohort studies are selected by some unique characteristic or risk factor. Cohort studies ask the question "what will happen," and thus the direction of cohort studies is forward in time, referred to as prospective studies. The schematic design of a cohort study is shown in Figure 3, where arrows indicate prospective or retrospective cohort study design.

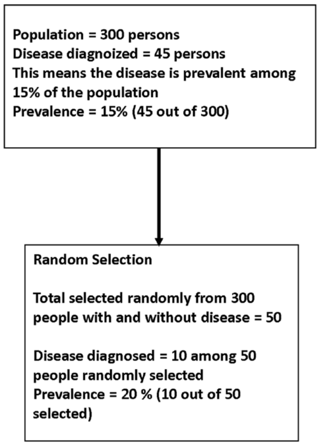
Difference between case–control and cohort study designs

Meirik provides the following differences between case–control and cohort studies. The starting point of a cohort study is the recording of healthy subjects with and without exposure to the putative agent or the characteristic being studied. Individuals exposed to the agent under study (index subjects) are followed over time and their health status is observed and recorded during the course of the study. In order to compare the occurrence of disease in exposed subjects with its occurrence in nonexposed subjects, the health status of a group of individuals not exposed to the agent under study (control subjects) is followed in the same way as that of the group of index subjects.

The starting point of a case–control study is subjects with the disease or condition under study (cases). The cases' history of exposure or other characteristics, or both, prior to onset of the disease, is recorded through interview and sometimes by means of records and other sources. A comparison group consisting of individuals without the disease under study (controls) is assembled and their past history is recorded in the same way as for the cases. The purpose of the control group is to provide an estimate of the frequency and amount of exposure in subjects in the population without the disease being studied. Whereas the cohort study is concerned with frequency of disease in exposed and nonexposed individuals, the case–control study is concerned with the frequency and amount of exposure in subjects with a specific disease (cases) and people without the disease (controls).

In the simplest case, in cohort studies, the incidence for the occurrence of the disease can be determined for both groups. Moreover, the relative risk[6] is a very important statistical parameter that can be calculated in cohort studies. For rare types of exposure, the general population can be used as controls. All evaluations naturally consider the age and gender distributions in the corresponding cohorts. One well known cohort study is the British Doctors Study, which prospectively examined the effect of smoking on mortality among British doctors over a period of decades. Cohort studies are well suited for detecting causal connections between exposure and the development of disease. On the other hand, cohort studies often demand a great deal of time, organization and money. So-called historical cohort studies represent a special case. In this case, all data on exposure and effect (illness) are already available at the start of the study and are analyzed retrospectively. For example, studies of this sort are used to investigate occupational forms of cancer. They are usually cheaper.

Like case–control studies, cohort studies generally stem from an already established group, or cohort, of patients. Consider the study by Zephir *et al*[7] In this study, the authors looked at treating MS patients (an established cohort) with cyclophosphamide as a disease-modifying therapy. This is an example of the "what will happen" style of clinical experiment design. In these types of studies, a group of patients who share a risk factor are either (a) followed without intervention or (b) all exposed to the same intervention and the results are observed. In this study, the authors note stabilization of



Population = 300 persons
Disease diagnoized = 45 persons
This means the disease is prevalent among 15% of the population
Prevalence = 15% (45 out of 300)

Random Selection

Total selected randomly from 300 people with and without disease = 50

Disease diagnosed = 10 among 50 people randomly selected
Prevalence = 20 % (10 out of 50 selected)

Figure 2: Prevalence of disease and random selection

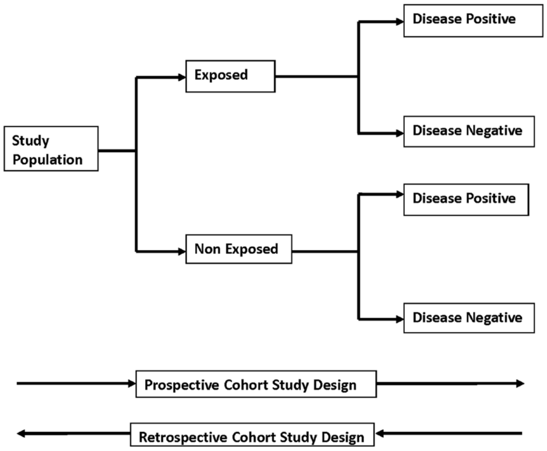


Figure 3: Case–control study design

MS disease progression scores following cyclophosphamide treatment. Naturally, the lack of a placebo control makes cohort studies that utilize an intervention as part of the study less efficacious than the gold standard for such trials, the randomized controlled trial.

Experimental studies or randomized clinical trials

Experimental studies or clinical trials are of two categories: those with and without controls. Because the purpose of an experiment is to determine whether the intervention (treatment) makes a difference, controlled studies are viewed as having greater validity. Controlled trials are of three categories: trials with concurrent controls, self-controls and external controls.

Under clinical trials, subjects are selected at random for both treatment and control groups. "The randomized clinical trial is the epitome of all research designs because it provides the strongest evidence for concluding causation: It provides the best insurance that the result was due to the intervention". Studies that do not use randomized assignment are referred to as nonrandomized trials. The RCT design is shown in Figure 4.

The intervention group receives treatment whereas the control group gets placebo as treatment. Trials with concurrent controls have two groups of subjects: one that receives the treatment and another that receives the placebo. Both the groups are treated similarly; interventions (treatment and placebo) are planned for the same time period. To reduce bias, researchers design double-blind trials, where neither the subjects nor the investigators know whether a subject is in the treatment or control group. As Stanley explains, if any of the outcome measures of an RCT (Randomized Control Trials) are responsive to treatment, then it is important that the trial be designed as a double-blind, placebo-controlled trial to control potential bias influencing outcome.

Studies with one group of subjects in which subjects (patients) are assessed before and after the intervention are called self-controlled studies. Clinical trials that use patients as their own controls with no control group are subject to the Hawthorne effect.[6]

Another form of clinical trial study is referred to as a crossover study. This design uses two groups of patients, where one group is assigned to treatment and the other to the placebo. After a period of time, both the groups are temporarily withdrawn ("washout") from the study, with no treatment. Then, the groups are altered for treatment. The treatment group receives the placebo and the control group receives the treatment.

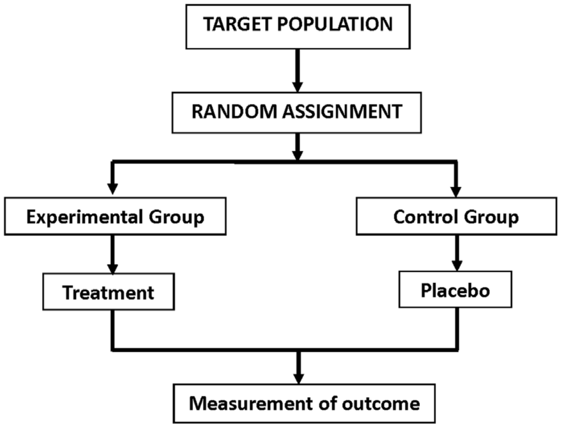


Figure 4: Randomized control study

Clinical trials with external controls use controls external to the study. The researchers in such situations use another investigator's research or patients the investigator has treated in another research (called historical controls) as a comparison. There are many trials that use such historical controls. There are preexisting data sets (such as data from large already existing trials, data from federal databases such as National Health And Nutrition Examination Survey in the US) that lend themselves toward use as historical controls. This saves data collection time and resources and, also, data can simply be mined from these preexisting databases to use as controls. Consider for example a trial evaluating a new intervention in a relatively common condition such as strokes or migraines. In this case, historical controls could be used, assuming that such data exists.

Metaanalysis

Metaanalysis uses published information and the data from other studies to address a set of related research hypotheses. A common use of metaanalysis is to study the overall effect of a very commonly used drug or procedure on a very prevalent disease. Consider antithrombotic therapy in stroke. There are many RCTs comparing Aspirin (ASA) to placebo in secondary stroke prevention. Sze *et al.* performed a metaanalysis of these RCTs and showed significant benefits to ASA administration in stroke.[8] Metaanalysis is particularly useful when a statistically significant relationship is suspected, but none of the trials to date happen to be powered sufficiently to unearth this relationship. In this case, the increased patient number afforded to metaanalyses (due to pooling of patients from different trials) may lend itself to the discovery of new relationships between variables. Of course, this comes at a cost of increased bias and comparing across research methodologies (and, often, in the case of drug trials, drug doses).

Conclusion

This article is intended to give the reader guidance in evaluating the design of studies in medical research, which will enable the reader to design medical studies better and to assess their scientific quality more accurately. Care and caution with skills and experience is needed to design suitable studies. Appropriate design coupled with reliable sampling techniques and appropriate statistical analysis, supported by clear objectives with decent communication of the findings of the result, are not easy to acquire. In our next article, we will discuss the case–control study design in more detail.

Notes

Odds Ratio

An estimate of the relative risk calculated in case–control studies. It is the odds that a patient was exposed to a given risk factor divided by the odds that a control was exposed to the risk factor.

Incidence rate

A rate giving the proportion of people who develop a given disease or condition within a specified period of time.

Selection bias

Selection bias exists due to a flaw in the sample selection process.

Recall bias

Recall bias occurs when the way a survey respondent answers a question is affected not just by the correct answer but also by the respondents memory.

Confounding variable

A confounding variable, also known as a third variable or a mediator variable, can adversely affect the relation between the independent variable and the dependent variable. This may cause the researcher to analyze the study results incorrectly.

Relative risk

The ratio of the incidence of a given disease in exposed or at-risk persons to the incidence of the disease in unexposed persons. It is calculated in cohort studies.

Hawthorne effect

This is referred to the tendency of some subjects to work harder and perform better when they participate in an experiment. Individuals may change their behavior due to the attention they are receiving from researchers rather than because of any manipulation of independent variables.

References

1. Alonso A, Hernan MA. Temporal trends in the incidence of multiple sclerosis: A systematic review. Neurology 2008;71:129-35.
2. Charlmers RM, Harding AE. A case-control study of Leber's hereditary optic neuropathy. Brain 1996;119:1481-6.
3. Dawson B, Trapp RG. “Basic and Clinical Biostatistics”. New York: Lange Medical Books, McGraw-Hill; 2001.
4. Doll R, Peto R. Mortality in relation to smoking: 20 years' observations on male British doctors. Br Med J 1976;2:1525-36.
5. Meirik O. ‘Cohort and Case Control Studies’. 2008 Available from: http://www.gfmer.ch/Books/Reproductive_health/Cohort_and_case_control_studies.html [Last accessed on 2011 Nov 24].
6. Stanley K. ‘Statistical Primer for Cardiovascular Research’, Circulation, American Heart Association Journal. Available from: http://circ.ahajournals.org/content/115/9/1164.full [Last accessed on 2011 Nov 24].
7. Zephir H, de Seze J, Duhamel A, Debouverie M, Hautecoeur P, Lebrun C, *et al.* Treatment of progressive forms of multiple sclerosis by cyclophosphamide: A cohort study of 490 patients. J Neurol Sci 2004;218:73-7.
8. Sze PC, Reitman D, Pincus MM, Sacks HS, Chalmers TC. Antiplatelet agents in the secondary prevention of stroke: Meta-analysis of the randomized controlled trials. Stroke 1988; 19:436-42.

How to cite this article: Suresh K, Suresh G, Thomas SV. Design and data analysis 1 study design. Ann Indian Acad Neurol 2012;15:76-80.

Received: 09-12-11, **Revised:** 26-01-12, **Accepted:** 28-02-12

Source of Support: Nil, **Conflict of Interest:** Nil