Study Design: Observational Studies

SIDDARTH RAMJI
From Department of Neonatology, Maulana Azad Medical College, New Delhi.
Correspondence to: Dr Siddarth Ramji, Director-Professor, Department of Neonatology, Maulana Azad Medical College, Delhi.
siddarthramji@gmail.com

Observational study designs are those where the investigator/researcher just observes and does not carry out any intervention(s)/actions to alter the outcome. The three most common types of observational studies are cross-sectional, case control and cohort (or longitudinal). In cross-sectional studies, both the exposure/risk factor(s) and the outcome(s) are determined at a single time point. They can provide information on prevalence of a condition and snapshot of probable associations that can be used to generate hypothesis. Case-control studies are where subjects are selected based on presence/absence of outcome and the risk factors are determined during the study after enrolment of study subjects. The association between exposure and outcome is reported as odds ratio. These studies; however, have high risk of bias, which must be taken care of during study design. Cohort studies are prospective in nature, where subjects are selected based on presence/absence of exposure, and the outcome(s) is determined at the end of study. These studies can provide incidence of disease/outcome and the association between exposure and outcome is reported as relative risk. They are useful to ascertain causality. High dropouts of study participants and confounding can be problems encountered in these studies.

Keywords: *Case-control, Cohort, Cross-sectional, Odds ratio, Relative risk, Survey*

Published online: April 26, 2022; ***PII:*** S097475591600418

Observational studies are research study designs where the investigator/researcher just observes and does not influence either the exposure or the outcome. In contrast, in experimental studies, the intervention/exposure is under the control of the investigator to bring a change in outcome [1]. These types of studies are important as they address many research questions which cannot be answered by experimental studies/clinical trials or where the latter study designs are not possible (e.g., health outcome after natural disasters such as the gas leakage from an industrial unit in Bhopal in 1984 or after a nuclear accident as occurred in Chernobyl Nuclear power plant in 1986), wherein it would be unethical for such events to be purposefully induced for purposes of experimentation. These may be relatively easier and faster to conduct than experimental designs.

Observational studies can either be descriptive or analytical. **Table I** summarizes the various ways in which observational studies can be categorized.

DESCRIPTIVE STUDIES

Descriptive studies generally describe the magnitude of a problem and characteristics of the population/individuals. The various types of such studies include case reports,

Table I Categorization of Observational studies

Classification category	*Qualifier/Explanation*
Relation to the population	• Subjects are selected based on presence of risk factors as in Cohort studies
	• Subjects are selected based on presence/absence of outcome as in case-control studies
Period of observation	• Single time point as in cross sectional and case-control studies
	• Followed longitudinally over time as in cohort studies
Timing of measurements	• Concurrent as in cohort studies
	• Non-concurrently as in case-control studies
	• Both concurrent and non-concurrent as in cross-sectional studies (depends on type of data being recorded)
Direction of investigation	• Prospective when investigation moves from risk factors/exposure to outcome (as in cohort studies)
	• Retrospective when the investigation moves from outcome to determine the risk factors (as in case-control studies)

case series or surveys. A case report generally describes a patient presenting with an unusual disease, or simultaneous occurrence of more than one condition, or uncommon clinical features in a known disease. A case series is a collection of similar cases. Such studies, other than providing some advancement to knowledge of a disease, are of limited value. Another method often used in epidemiological health care research is conducting surveys. Surveys are done during a defined time-period and information on several variables of interest is collected from the target population. They provide estimates of prevalence of the various variables of interest, and their distribution. Such studies could also provide insight into individual opinions and practices. Advantages include ease of conduct and cost efficiency. The disadvantages include low response rates and a variety of biases.

ANALYTICAL STUDIES

An analytical study tests a hypothesis to determine an association between two or more variables, like causation, risk, or effect. Such studies have two or more study groups for comparison. The primary focus of this article will be the three most common types of analytical observational studies - cross-sectional, case control (also known as retrospective) and cohort (or longitudinal, also known as prospective) studies. It may be pertinent to note that the primary objective of most clinical studies is to determine one of the following - burden of disease (prevalence or incidence), cause of disease, prognosis, or effect of treatment/intervention. Each of the study designs should mention the details of participant, exposure and comparison/control group (as applicable), outcome to be analyzed and time as per the PICOT format clearly in the protocol.

Incidence (new events) can be determined from cohort studies and prevalence (old plus new events) from cross sectional studies. Cause/etiology can be ascertained from cohort, case control or cross-sectional studies where there are two or more comparison groups (in their decreasing order of reliability). Prognosis can be provided by cohort studies that prospectively measure outcome. However, effect of treatment cannot be reliably obtained from observational studies and would need a controlled clinical trial (experimental design).

Cross-Sectional Studies

Cross-sectional studies provide a snapshot of both the exposure (or risk factor) and the outcome in the sample. Here the information on exposure and outcome are noted at the same time. They are generally descriptive and provide information on prevalence (the number of cases in a population at a given point/period). However, cross-sectional studies can also be analytical where the strength of association between two variables can be estimated and is reported as an odds ratio (OR), which can be useful for hypothesis generation.

Conducting Cross-Sectional Studies

The first step would be to formulate a research question. The next step is to identify the target population to whom the results would be generalized and then select a study sample as per the logistic considerations. The sampling strategy could be simple random, stratified (e.g., age and gender where outcomes are dependent on these variables), systematic (especially in hospital-based studies), multi-stage (e.g., districts, villages, households) or cluster sampling (the last two sampling methods are useful when large populations are to be included in the study) [2]. Lastly, the variables relevant to the research question must be identified during the study planning stage.

Advantages and Disadvantages

Cross-sectional studies are useful when one wishes to gather information rapidly and in an inexpensive way. It can provide descriptive data and can also determine an association between two or more groups. Another advantage is that multiple outcomes can be studied.

However, cross-sectional studies cannot assess risk but can provide information on association. They would also not be an appropriate choice where one or more of the variables of interest are rare. They do not provide reliable information on causality or the sequence of events. An example is provided in **Box I**.

Case-Control Studies

Case-control studies are retrospective in nature. Subjects are enrolled based on "presence (cases)" or "absence (controls)" of outcome (or disease). Information related to

Box I An Example of a Cross-Sectional Study

Example: A study in Ethiopia aimed to determine the prevalence and associated factors of malaria in under-five children. This was a facility-based cross-sectional study conducted among 585 under-five children who attended public health facilities. Health facilities were selected by stratified cluster sampling, and systematic random sampling was done to select study participants from the selected facilities. Malaria was defined as a positive thin or thick blood film for the *Plasmodium* parasite. It was observed that 51 (8.7%) children had malaria. [3]

Comment: In this example, the study was conducted across multiple hospitals (hence cluster sampling was used) and systematic random sampling was used to select participants within each health facility. The research question required estimation of point prevalence of malaria as per a predefined diagnostic criterion.

potential risk factors is collected by the investigator after the outcome has occurred. As the direction of enquiry is from outcome to exposure, hence is termed as a retrospective study.

Conducting Case-Control Studies

As for all studies a research question must be formulated. Unlike cross sectional studies, case-control studies would require an *apriori* hypothesis and hence one must decide what must be measured and how. The next step is to choose the case and control group with valid and precise definitions. The definition of a case should be objective (e.g., diagnostic criteria of a disease or event) such that there is no ambiguity in type of cases and/or their severity. Cases can be hospital-based or population-based. Defining control is equally important and critical to the outcome of case-control studies. Every effort must be made to ensure absence of outcome in controls. The controls should represent the population from where the cases have been selected and need not always be healthy. However, they should be equally predisposed to develop the outcome under study and are selected from the same source population as the cases. They can be selected from friends, relatives or from the neighbourhood as this can be done with minimum effort. Such controls, however, have the risk of being very similar with respect to exposures and other characteristics (can be overmatched). On the other hand, selecting an appropriate hospital control can be challenging and care must be taken to prevent selection bias. Use of two or more controls increases reliability (e.g., outpatients, inpatients, from general practice, etc.). Selecting an appropriate control is one of the most important steps in a case-control study [4].

The exposure of interest is considered 'a risk factor' and its association with the outcome is reported as odds ratio (OR) (or as adjusted OR after confounder control). The odds ratio from case control studies informs one as to how much higher the odds of exposure are among case-patients than among controls (or if it is associated with reduced risk, it would inform about how much lower is the odds of exposure among case-patients than among control). For rare diseases, such as cancer, the odds ratio is likely to approximate relative risk values obtained from prospective (cohort) studies. An example is provided in **Box II**.

Nested case-control study

This is a variant of a case-control study in which the cases and controls are drawn from a pre-existing cohort study. For every defined case in the cohort a matched control from the cohort is selected. They are particularly useful for studying the biological precursors of disease. The advantage of this design is that it minimizes selection, recall bias and measurement bias (as the variables have been pre-defined and collected concurrently in a standardized manner having been part of a cohort study) in comparison to case-controls studies, and is faster and less expensive than a cohort study. An example is provided in **Box III**.

Advantages and Disadvantages

Case control studies are relatively simple to perform and provide results over a relatively shorter time than cohort studies and require lesser resources. They provide information about predictors of an outcome. The problem of confounding can be overcome using matched controls for a few of the important selected confounding characteristics.

The disadvantages are that they cannot provide information on incidence or prevalence and can look at only one outcome. They are also more prone to bias, particularly selection, observation, recall and measurement bias. Selection bias could be minimized by use of matched or population-based controls. Recall bias could be minimized by using recorded data before the occurrence of the outcome being studied. Because of the risks of confounding and bias, case control studies are less reliable for ascertaining causality but can help in hypothesis generation.

Box II Example of a Case-Control Study

Example: This study attempted to identify the characteristics of malnourished children below five years of age and the risk factors of childhood malnutrition. It hypothesized that risk of childhood malnutrition (outcome) was increased in poor households (exposure). Case was a child with moderate to severe malnutrition (z-scores <-2SD from the median of WHO reference of any anthropometry - weight, length/height, weight/length). Control was a child with z-scores between -2SD and +2SD and was age matched with the cases. The participants were identified from those attending the maternal-child health clinics. The study identified the variables/exposures that could affect nutrition from three domains-socio-economic characteristics, household food security and child's dietary intake, and caregiver practices and resources. The study clearly defined how the above measures were to be collected. There were 137 cases and 137 controls. The study identified the following as significant risk factors for childhood malnutrition – Household poverty (OR 3.15,95% CI: 1.65-6.04),…. [5]

Comment: In the above example there is a clearly defined research question and hypothesis. The cases and controls are clearly defined. Variables to be measured and process of measurement were also clearly defined. The study has reported the association of exposure and outcome as OR. Though not depicted in the example, the study had observed several significant risk factors and hence it also reported adjusted ORs after adjusting for confounding.

Box III Example of a Nested Case-Control Study

Example: This study was designed to determine predictors of tuberculosis recurrence in the UK. Leicester TB service has a cohort of all TB cases reported, managed, and kept under follow up and surveillance in the region in a prospectively maintained database. For this nested case–control study the participants were identified from the entries in the database between 1994 and 2014. All participants who had been cured after their first episode of TB infection were eligible for inclusion. Those who had been reported as new and first recurrence episode of TB in the database were defined as cases. Controls were those with no recurrence and were matched to cases in a ratio of two to one by the date of notification. Clinical, sociodemographic and TB specific risk factors for recurrence were identified and extracted from the database of the cohort. From the original cohort, there were 4764 eligible patients included in the current study. In these 82 TB recurrences were noted (1.8%) (cases). Smoking (OR 3.8; p=0.04), grade 3/4 adverse drug reactions (OR 5.6; p=0.02), ethnicity 'Indian subcontinent' (OR 8.5; *P*=<0.01), ……were predictors of TB recurrence [6].

Comment: Since the cases and controls were selected from an existing cohort of TB patients based on a defined outcome that had already occurred i.e., TB recurrence, it qualifies as a nested case-control study.

Cohort Studies

Cohort studies (or longitudinal studies) are generally prospective but can also be retrospective. However, unless qualified, the term cohort studies imply prospective cohort studies. Cohort studies are the designs of choice for determining incidence and natural course of a disease. The association of risk factors with outcome is generally reported as relative risk (RR) or attributable risk.

Prospective Cohort Studies

In prospective cohort studies, the participants are a group of individuals in whom the outcome of interest (e.g., lung cancer) has not occurred at the time of selection into the study. The investigator identifies all possible relevant variables that may contribute to the development of the outcome and measures them accurately in the participants during follow-up. The participants during follow-up are carefully followed up and observed to see if they develop the outcome of interest. The steps in a prospective cohort are as follows: *i*) identify a research hypothesis, *ii*) define objectively the exposure (risk factors) and outcomes, *iii*) develop a standard data collection tool (to minimize information bias), *iv*) identify steps during follow-up to minimize the drop outs (selection bias), and *v*) define the analysis plan to measure the association between one/ many factors and the outcome.

a. *Descriptive Cohort*: A descriptive cohort study describes outcome over time for a specific group of individuals, without any comparison of groups. Examples are patients with a defined type of cancer(s) who are followed up to describe the epidemiology of the disease.

b. *Single analytic cohort*: In such a study, those who develop the outcome of interest are treated as 'cases' and those who do not develop the outcome of interest are treated as 'controls' (also known as internal controls). The investigator may also opt for an external control when internal controls are not available (when exposure cannot not be verified/ was not measured).

c. *Two group analytic cohort*. When there are two cohorts being followed, one of the groups would have been exposed to the variable of interest while the other would not. Both groups would be followed to observe for the occurrence of the outcome of interest. Here those without exposure to the variable of interest would serve as the comparison/control group (external controls).

Retrospective Cohort Study

This is a type of cohort study where the investigator looks back retrospectively at already collected data (prospectively) after the outcome of interest has occurred i.e., post-hoc. The important distinguishing feature of retrospective from prospective cohort study is that the investigator comes up with the idea of the study and begins to identify variables and subjects after the outcome of interest has occurred. An *ambispective cohort* study design allows the researcher to retrospectively measure the exposure in a cohort and follow them prospectively for a disease outcome. An ambispective design saves resources and time.

Advantages and Disadvantages

Cohort studies are the choice where randomized control trials would be considered unethical e.g., effect of smoking on lung cancer. As the design affords a temporal sequence, it is a very good method to establish cause and effect (hence allows one to measure incidence) and analyze risk factors or predictors (allows to measure relative and attributable risk). They also have the advantage of being able to measure multiple outcomes from a single study e.g., effect of breast-feeding duration on child growth, obesity, diarrhea, and acute illnesses [10] or even multiple exposures e.g., effect of multiple environmental exposures to child health outcomes [11].

The disadvantages could include loss of subjects (loss of cases becomes more important as it can alter incidence rates) during the follow-up, confounding and non-representative nature of the cohort sample (selection bias).

Examples of various cohort studies are provided in **Box IV**.

Comparison of Cohort, Case-control, and Cross-sectional studies

Table II summarizes the comparison of the three types of studies. Let us see an illustration of how each of the study designs can be used to address the same research question. Let us say, if one wanted to assess the association of maternal anemia in pregnancy and low birth weight (LBW), one could address the research question by all three study designs.

Cohort study design: Here, mothers with a hemoglobin below and above a pre-defined level (to define anemia) at a defined gestational period of pregnancy would be enrolled (where the hemoglobin estimation is standardized) and followed till birth of the baby and outcome determined by the weighing the baby and classifying as low or not low birth weight. The study would need each mother to be followed up till outcome. To enrol the required number of women could take a long time. There could be drops outs, there could be other factors that could appear during follow up that could contribute to LBW (e.g., hypertension). But the advantage is that data would be collected concurrently, would be reliable and not only association but causality could also be established.

Case-control design: For this, infants who are born with LBW and normal birth weight would be sampled as cases and controls. Information about the hemoglobin status of the mothers of these infants would be sought from records. However, there could be concerns about the accuracy and reliability of the hemoglobin estimation. In addition, details of other risk factors may be missing/ incomplete. The study can establish an association but causality would have lower reliability than a cohort study.

Cross-sectional study: For this study design, we would sample women who have delivered recently. The hemoglobin could be either estimated at the time of participant selection or the information could be obtained from maternal antenatal records. The weight of baby could be measured at time of mother-baby dyad selection or noted from hospital records. While this method would be rapid, the methods used to obtain information/measure anemia or birth weight could affect the interpretation of the results. The issues related to reliability of data and confounding will remain. Hence, this study design would at best help generate a hypothesis.

CONCLUSION

Observational studies are useful because the associations that these studies provide can be used to generate hypothesis. They are also useful to study rare events. The

Box IV Examples of Cohort Studies

Single group prospective cohort study

Example: A prospective birth cohort focused on atopy and asthma development in children that hypothesized low physical activity as a risk factor of asthma. Asthma was identified between 6-10 years using ISAAC criteria. Physical activity was assessed using questionnaire at 4-5 years age. The children were followed at regular 1-2 year interval till 10 years of age. There were 1957 children who met the inclusion criteria at the age of 4 to 5 years. Of these, 1838 children (94%) were evaluated for asthma symptoms between 6 and 10 years. A total of 186 children (10.1%) met the ISAAC definition of asthma between the age of 6 and 10. No association was found between physical activity and asthma (RR 1.13, 95% CI:0.95-1.34) [7].

Comment: In this birth cohort study the cases were those who developed asthma and the internal controls were those who did not develop asthma. The definitions of outcomes are clearly defined. The exposure variables that needed to be studied and measured at each visit had been identified. The study reported the incidence of asthma and the association between exposure and outcome as RR.

Two-group prospective cohort study

Example: This study evaluated longitudinal changes in cardiac structure and function of patients with Rheumatoid arthritis (RA) compared with persons in the general population. A cohort of 160 patients with RA and 1,391 persons without RA (non-RA cohort), each underwent 2-dimensional, pulsed-wave tissue Doppler echocardiography at baseline and after 4 to 5 years of follow-up. The mean mitral inflow E/A ratio decreased faster in the RA cohort than the non-RA cohort ($p<0.001$), the left atrial volume index increased at a higher rate in the RA cohort than the non-RA cohort ($p<0.001$) [8].

Comment: In this example the exposure was RA/no RA, and outcome was cardiac structure and function. As the outcomes were continuous variables, instead of RRs, the investigators compared the means and their variances between the two groups. It is important to note that the comparison of outcomes will depend on the characteristic of the variable being measured.

Retrospective cohort study

Example: In this retrospective cohort study (from a prospectively followed British birth cohort), the association of life course events with adult irritable bowel syndrome (IBS) was evaluated. Adult subjects were enrolled after the outcome of interest – irritable bowel syndrome - had occurred. The investigators extracted data on life course events data from the birth cohort that was being followed. The outcome was self-reported IBS by the age of 42 years. The prevalence of self-reported IBS in this cohort was 8.4% (95% CI=8.2-8.6). Being female (OR=2.00, 95% CI=1.67-2.3), and having psychopathology at 23 years (OR=1.25, 95% CI=1.01-1.54) were associated with increased odds for IBS [9].

Key Messages

- Observational studies are the option when randomized clinical trials are not feasible or unethical.
- Observational studies are useful for generating hypothesis and studying rare events.
- Main problem with observational studies is confounders and biases; but advanced statistical methods may help in controlling for many confounders.

Table II Comparison of Cohort, Case-control, and Cross-Sectional Studies

Parameters	*Cohort study*	*Case-control study*	*Cross-sectional study*
Direction of investigation	From Exposure to Outcome (forward)	From outcome to exposure (backward)	As it exists
Recruitment of subjects	Based on presence/ absence of exposure	Based on presence/ absence of outcome	Neither exposure nor outcome
What does it measure	Incidence, Relative risk	Odds ratio	Prevalence
Temporal relationship	Good	Difficult	Not possible
Causality assessment	Good	Fair	Poor
Suitability for rare exposures	Good	Poor	poor
Suitability for rare outcomes	Poor	Good	Poor
Biases	Low	High	Low
Confounding	Major problem	Major problem	Major problem
Loss of subjects	High	Low	None
Completeness of data	High	Low	Complete
Range of exposures that can be assessed	Small	Large	Large
Duration and cost	High	Low	Low
Range of outcomes that can be assessed	Large	Small	Large

choice of study design is not only determined by the research question, but also the pros and cons of each study and the feasibility for its implementation (**Table II**).

Acknowledgement: Dr. Aashima Gupta for finalization of the manuscript.
Funding: None; *Competing interests*: None stated.

REFERENCES

1. Khan AM, Gupta P, Mishra D. The 3-question approach: a simplified framework for selecting study designs. Indian Pediatr. 2019;56:669-72.
2. Indrayan A. Medical Statistics,3rd edn. Boca Raton, CRC Press, Taylor and Francis Group, 2013; 83-110.
3. Tsegaye AT, Ayele A, Birhanu S. Prevalence and associated factors of malaria in children under the age of five years in Wogera district, northwest Ethiopia: A cross-sectional study. PloS One. 2021;16:e0257944.
4. Case control and cross-sectional studies. *In*: Coggon D, Rose G, Barker DJP, editors. Epidemiology for the Uninitiated, 5th ed. BMJ publishing group, 2003.
5. Wong HJ, Moy FM, Nair S. Risk factors of malnutrition among preschool children in Terengganu, Malaysia: A case control study. BMC Public Health. 2014;14:785.
6. Rosser A, Richardson M, Wiselka, M.J. et al. A nested case–control study of predictors for tuberculosis recurrence in a large UK Centre. BMC Infect Dis. 2018;18:94.
7. Eijkemans M, Mommers M, Remmers T, et al. Physical activity and asthma development in childhood: Prospective birth cohort study. Pediatr Pulmonol. 2020;55:76-82.
8. Davis JM, Lin G, Oh JK, et al. Five-year changes in cardiac structure and function in patients with rheumatoid arthritis compared with the general population. Int J Cardiol. 2017;240:379-85.
9. Goodwin L, White PD, Hotopf M, et al. Life course study of the etiology of self-reported irritable bowel syndrome in the 1958 British birth cohort. Psychosom Med. 2013;75:202-10.
10. Pattison KL, Kraschnewski JL, Lehman E, et al. Breastfeeding initiation and duration and child health outcomes in the first baby study. Prev Med. 2019;118:1-6.
11. Aasvang GM, Agier L, Andruðaitytë S, et al. Human early life exposome (HELIX) study: A European population-based exposome cohort. BMJ Open. 2018;8:e021311.