

Methodology Series Module 1: Cohort Studies

Maninder Singh Setia

Abstract

Cohort design is a type of nonexperimental or observational study design. In a cohort study, the participants do not have the outcome of interest to begin with. They are selected based on the exposure status of the individual. They are then followed over time to evaluate for the occurrence of the outcome of interest. Some examples of cohort studies are (1) Framingham Cohort study, (2) Swiss HIV Cohort study, and (3) The Danish Cohort study of psoriasis and depression. These studies may be prospective, retrospective, or a combination of both of these types. Since at the time of entry into the cohort study, the individuals do not have outcome, the temporality between exposure and outcome is well defined in a cohort design. If the exposure is rare, then a cohort design is an efficient method to study the relation between exposure and outcomes. A retrospective cohort study can be completed fast and is relatively inexpensive compared with a prospective cohort study. Follow-up of the study participants is very important in a cohort study, and losses are an important source of bias in these types of studies. These studies are used to estimate the cumulative incidence and incidence rate. One of the main strengths of a cohort study is the longitudinal nature of the data. Some of the variables in the data will be time-varying and some may be time independent. Thus, advanced modeling techniques (such as fixed and random effects models) are useful in analysis of these studies.

Key Words: Cohort studies, limitations, strengths

From the Department of
Epidemiology, MGM Institute of
Health Sciences, Navi Mumbai,
Maharashtra, India

Address for correspondence:

Dr. Maninder Singh Setia,
MGM Institute of Health Sciences,
Navi Mumbai, Maharashtra, India.
E-mail: maninder.setia@
karanamconsultancy.in

Introduction

Cohort studies are important in research design. The term "cohort" is derived from the *Latin* word "Cohors" – "a group of soldiers." It is a type of nonexperimental or observational study design. The term "cohort" refers to a group of people who have been included in a study by an event that is based on the definition decided by the researcher. For example, a cohort of people born in Mumbai in the year 1980. This will be called a "birth cohort." Another example of the cohort will be people who smoke. Some other terms which may be used for these studies are "prospective studies" or "longitudinal studies."

Design

In a cohort study, the participants do not have the outcome of interest to begin with. They are selected based on the exposure status of the individual. Thus, some of the participants may have the exposure and others do not have the exposure at the time of initiation

of the study. They are then followed over time to evaluate for the occurrence of the outcome of interest.

As seen in Figure 1, at baseline, some of the study participants have exposure (defined as exposed) and others do not have the exposure (defined as unexposed). Over the period of follow-up, some of the exposed individuals will develop the outcome and some unexposed individuals will develop the outcome of interest. We will compare the outcomes in these two groups.

Examples of Cohort Studies

Framingham cohort study (<https://www.framinghamheartstudy.org/index.php>)

This cohort study was initiated in 1948 in Framingham. Framingham, at the time of initiation of the cohort, was an industrial town 21 miles west of Boston with a population of 28,000. This Framingham Heart Study recruited 5209 men and women (30–62-year-old) in the study to assess

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as the author is credited and the new creations are licensed under the identical terms.

For reprints contact: reprints@medknow.com

How to cite this article: Setia MS. Methodology series module 1: Cohort studies. *Indian J Dermatol* 2016;61:21-5.

Received: December, 2015. **Accepted:** December, 2015.

Access this article online	
Quick Response Code: 	Website: www.e-ijd.org
	DOI: 10.4103/0019-5154.174011

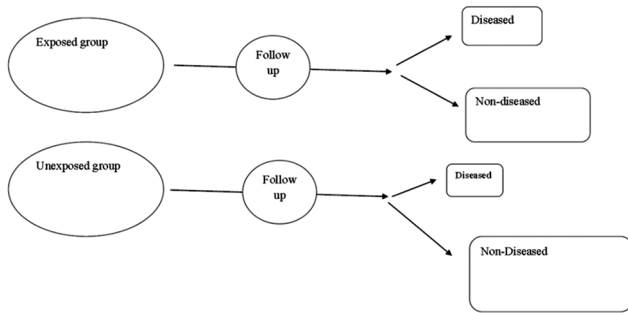


Figure 1: Example of a cohort study

the factors associated with cardiovascular disease (CVD). The researchers also recruited second generation participants (children of original participants) in 1971 and the third general participants in 2002. This has been one of the landmark cohort studies and has contributed immensely to our knowledge of some of the important risk factors for CVD. The investigators have published 3064 publications using the Framingham Heart Study data.

Swiss HIV cohort study (<http://www.shcs.ch/>)

This cohort study was initiated in 1988. It was a longitudinal study of HIV-infected individuals to conduct research on HIV pathogenesis, treatment, immunology, and coinfections. They also work on the social aspects of the disease and management of HIV-infected pregnant women. The study started with a recruitment of individuals ≥ 16 years. The cohort was gradually expanded to include the Swiss Mother and Child HIV Cohort Study. The cohort has provided useful information on various aspects of HIV and published 542 manuscripts on these aspects.

The Danish cohort study of psoriasis and depression (Jensen, 2015)

This is another large cohort study that evaluated the association between psoriasis and onset of depression. The participants in the cohort were enrolled from national registries in Denmark. None of the included participants had psoriasis or depression at baseline. The outcome of interest was the initiation of antidepressants or hospitalization for depression. The authors compared the incidence rates of hospitalization for depression in psoriasis and reference population. The psoriasis group was further classified as mild and moderate psoriasis. The authors found that psoriasis was an independent risk factor for new-onset depression in young people. However, in the elderly, it was mediated through comorbid conditions.

We have presented examples of some large cohort studies. It will be worthwhile to read the design and conduct of these studies, and it will help the readers understand the practical aspects of conducting and analyzing cohort studies.

Types of Cohort Studies

Prospective cohort study

In this type of cohort study, all the data are collected prospectively. The investigator defines the population that will be included in the cohort. They then measure the potential exposure of interest. The participants are then classified as exposed or unexposed by the investigator. The investigator then follows these participants. At baseline and during follow-up, the investigator also collects information on other variables that are important for the study (such as confounding variables). The investigator then assesses the outcome of interest in these individuals. Some of these outcomes may only occur once (for example, death), and some may occur multiple times (for example, conditions which may recur in the same individual – diarrhea, wheezing episodes, etc.).

Retrospective cohort study

In this type of cohort study, the data are collected from records. Thus, the outcomes have occurred in the past. Even though the outcomes have occurred in the past, the basic study design is essentially the same. Thus, the investigator starts with the exposure and other variables at baseline and at follow-up and then measures the outcome during the follow-up period.

Sometimes, the direction may not be as well defined as prospective and retrospective. One may analyze retrospective data on a group of people well as collect prospective data from the same individuals.

Examples of prospective and retrospective cohort studies

Example 1

Our objective is to estimate the incidence of cardiovascular events in patients with psoriasis. We have decided to conduct a 10-year study. All the individuals who are diagnosed with psoriasis are eligible for being included in this cohort study. However, one has to ensure that none of them have cardiovascular events at baseline. Thus, they should be thoroughly investigated for the presence of these events at baseline before including them in the study. For this, we have to define all the events we are interested in the study (such as angina or myocardial infarction). The criteria for identifying psoriasis and cardiovascular outcomes should be decided before initiating the study. All those who do not have cardiovascular outcomes should be followed at regular intervals (predecided by the researcher and as required for clinical management). This will be a prospective cohort study.

Example 2

Our objective is to assess the survival in HIV-infected individuals and the factors associated with survival.

We have clinical data from about 430 HIV-infected individuals in the center. The follow-up period ranges from 3 months to 4 years, and we know that 33 individuals have died in this group. We decide to perform the survival analysis in this group of individuals. We prepare a clinical record form and abstract data from these clinical forms. This design will be a retrospective cohort study.

Outcomes in a Cohort Study

A cohort study may have different types of outcomes. Some of the outcomes may occur only once. In the above mentioned retrospective study, if we assess the mortality in these individuals, then the outcome will occur only once. Other outcomes in the cohort study may be measured more than once. For instance, if we assess CD4 counts in the same retrospective study, then the values of CD4 counts may change at every visit. Thus, the outcome will be measured at every visit.

Strengths of a Cohort Study

- **Temporality:** Since at the time of entry into the cohort study, the individuals do not have outcome, the temporality between exposure and outcome is well defined
- A cohort study helps us to study multiple outcomes in the same exposure. For example, if we follow patients of hypercholesterolemia, we can study the incidence of melasma or psoriasis in them. Thus, there is one exposure (hypercholesterolemia) and multiple outcomes (melasma and psoriasis). However, we have to ensure that none of the individuals have any of the outcomes at the baseline
- If the exposure is rare, then a cohort design is an efficient method to study the relation between exposure and outcomes
- It is generally said that a cohort design may not be efficient for rare outcomes (a case-control design is preferred). However, if the rare outcome is common in some exposures, then it may be useful to follow a cohort design. For example, melanoma is not a common condition in India. Hence, if we follow individuals to study the incidence of melanoma, then it may not be efficient. However, if we know that, theoretically, a particular chemical may be associated with melanoma, then we should follow a cohort of individuals exposed to this chemical (in occupational settings or otherwise) and study the incidence of melanoma in this group
- In a prospective cohort study, the exposure variable, other variables, and outcomes may be measured more accurately. This is important to maintain uniformity in the measurement of exposures and outcomes. This is also useful for exposures that may require

subjective assessment or recall by the patient. For example, dietary history, smoking history, or alcoholic history, etc. This may help in reducing the bias in measurement of exposure

- A retrospective cohort study can be completed fast and is relatively inexpensive compared with a prospective cohort study. However, it also has other strengths of the prospective cohort study.

Limitations of a Cohort Study

- One major limitation of a prospective cohort design is that is time consuming and costly. For example, if we have to study the incidence of cardiovascular patients in patients of psoriasis, we may have to follow them up for many years before the outcome occurs
- In a retrospective cohort study, the exposure and the outcome variables are collected before the study has been initiated. Thus, the measurements may not be very accurate or according to our requirements. In addition, the some of the exposures may have been assessed differently for various members of the cohort
- As discussed earlier, cohort studies may not be very efficient for rare outcomes except in some conditions.

Additional Points in Cohort Studies

Multiple cohort study

Sometimes, we may be interested to compare the outcomes in two or more groups of individuals. Thus, we may have a multiple cohort study. It is important the exposure, outcome, and other variables should be measured similarly in both the study and the comparison group.

Measurement of exposure and outcome

Since the individuals are included in the study based on the exposure status, this has to be well defined and accurate. The outcomes also have to be well defined and measured similarly in all the participants. If you have more than one group in the cohort (as in multiple cohorts or reference population), you should ensure that the follow-up protocols are similar in all the groups.

Question: What if there is an error in measuring the exposure or the outcome?

It is quite possible that individuals participating in a cohort study may not be correctly classified – some exposed individuals may be classified as unexposed and the other way round. If the misclassification of the exposure or the outcome is random or nondifferential, then the two groups will be similar and the estimates from the study will be biased towards the null. Thus, we will underestimate the association between the exposure and the outcome. If, however, the misclassification is differential or nonrandom, then the estimates may be

biased toward the null, away from the null, or may be an appropriate estimate.

Follow-up

Follow-up of the study participants is very important in a cohort study and losses are an important source of bias in these types of studies. Some patients are lost to follow-up in large cohorts; however, if the proportion is very high (>30%), then the validity of the results from this study are doubtful. This loss to follow-up becomes all the more important if it is related to the exposure or outcome of interest. For example, in our prospective study, majority of the patients who were lost to follow-up had severe psoriasis at the baseline, then we will get biased estimates from the study. Thus, managing follow-ups and minimizing losses are an important component of the design of a cohort study.

Nested case-control study

This is a specific type of study design nested within a cohort study. In this, the investigator will match the controls to the cases within a specific cohort. The exposure of interest will be assessed in these selected cases and controls. For example, our hypothesis is that there is a biological marker that is present/elevated (to begin with) in individuals who develop cardiovascular events in psoriatic patients. It is expensive to assess this marker in all patients. Thus, we select all those who develop the outcomes (cases) in our cohort and a sample of individuals who do not develop the outcomes (controls). An important aspect, however, is that we should have stored the biological material that we have collected at baseline, and the biological marker should be assessed in this sample. This procedure maintains the temporal strength of the cohort study.

Analysis

Cohort studies will help us to estimate the cumulative incidence and incidence rate.

Cumulative incidence

Example

We follow 10,000 psoriatic patients for 10 years. Of these, 50 have a cardiovascular event. Thus, the cumulative incidence will be $50/10,000$ or 0.005. This measure is a proportion. Thus, the cumulative incidence will be 0.5% or $5/1000$.

Incidence rate

Example

We follow-up 10,000 psoriatic patients for 10 years. Of these, 50 have a cardiovascular event.

How do we calculate the incidence rate?

Let us assume that all the cardiovascular events occurred at the end of the 2nd year. Our outcome of interest was

the first cardiovascular event. Thus, at the end of the 2nd year, 50 individuals have the outcome.

The total time contributed by these 50 individuals is 50×2 years = 100 person years (PY) - (A).

The total time contributed by the rest of the cohort is $(10,000 - 50) \times 10 = 99,500$ PY - (B).

Thus, the total person time is $A + B = 99,600$.

The incidence rate is $50/99,600$ or 0.000502. As it is obvious from the term, this measure is a rate (compared with cumulative incidence which was a proportion). Thus, the incidence rate of first cardiovascular event in psoriatic patients is 0.502/1000 PY or 5.02/10,000 PY.

Other analysis

Other methods such as logistic regression, Kaplan-Meier curves, cox-regression, Poisson regression, lognormal regression may be useful in cohort studies. These are relatively advanced analyses and should be discussed with a statistician.

Fixed and random effects models

One of the main strengths of a cohort study is the longitudinal nature of the data. Some of the variables are time varying (such as blood pressure), and some may be time independent (such as sex). The fixed and random effects models are useful to handle longitudinal data. The random effects model provides both between- and within-individual variance and is useful for time-dependent and time-independent variables. These models are used in linear outcomes (such as body mass index) or categorical outcomes (such as presence/absence of psoriasis). These are advanced modeling techniques and should be discussed with a statistician.

Some Practical Points

Project management

The investigator should remember that conducting a large-scale prospective cohort study requires proper project management.

Follow-up of participants

The investigator should devise strategies to ensure proper follow-up of individuals at the designated time intervals. A computer program should be put in place at the start of the prospective study. The program should indicate the number of participants due for a visit every day. If the individual does not visit for the next week, a reminder should be sent to the individual. This can be performed through texting or a phone call to the individual. Some investigators hire field workers or outreach workers to ensure follow-up of study participants.

It is important that we include only patients with permanent addresses in the area for long-term cohort studies. Details about the stay (permanent address,

temporary address, and duration of residence in the current address) should be a part of the inclusion criteria.

Data management

The investigator should prioritize data management in these studies. The data entry program should be installed at the start of the project. In addition, data entry and cleaning should be done as soon as data are collected. This will help us to identify the lacunae in the existing data, loss of follow-ups, and missing data points.

Missing data

It is very important to address missing data in cohort studies. There are statistical methods to handle missing data in studies – such as complete case analysis, available case analysis, single imputation, or multiple imputations. The investigator should work with a statistician to address missing data in the dataset. These methods should also be described in the statistical analysis section of the manuscript.

Summary

In a cohort study, participants who do not have the outcome at baseline are followed over time to estimate the incidence of the outcome. In this type of design, the temporality between the exposure and outcome is well defined. The studies may be prospective, retrospective, or a mixture of both. Prospective cohort studies may be time consuming and expensive. Losses during follow-up are an important source of bias in cohort studies; thus, measures to ensure follow-up of participants should be included in the design of a prospective cohort study. Advanced modeling techniques are useful to analyze longitudinal data and are preferred in cohort studies.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

Bibliography

1. Hennekens CH, Buring JE. *Epidemiology in Medicine*. 1st ed. Philadelphia, USA: Lippincott Williams & Wilkins; 1987.
2. Egeberg A, Khalid U, Gislason GH, Mallbris L, Skov L, Hansen PR. Impact of depression on risk of myocardial infarction, stroke and cardiovascular death in patients with psoriasis: A Danish Nationwide Study. *Acta Derm Venereol* 2015. DOI: 10.2340/00015555-2218.
3. Framingham Heart Study. Available from: <https://www.framinghamheartstudy.org/index.php>. [Last accessed on 2015 Nov 14].
4. Jensen P, Ahlehoff O, Egeberg A, Gislason G, Hansen PR, Skov L. Psoriasis and new-onset depression: A Danish Nationwide Cohort Study. *Acta Derm Venereol* 2015. DOI: 10.2340/00015555-2183
5. Jewell N. *Statistics for Epidemiology*. Boca Raton, US: Chapman and Hall/CRC; 2004.
6. Twisk JW. *Applied Longitudinal Data Analysis for Epidemiology*. 2nd ed. Cambridge, UK: Cambridge University Press; 2013.
7. Keiser O, Taffé P, Zwahlen M, Battagay M, Bernasconi E, Weber R, *et al*. All cause mortality in the Swiss HIV Cohort Study from 1990 to 2001 in comparison with the Swiss population. *AIDS* 2004;18:1835-43.
8. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, USA: Lippincott Williams and Wilkins; 2008.
10. Kleinbaum D, Kupper L, Morgenstern H. *Epidemiologic Research*. New York, US: John Wiley and Sons, Inc.; 1982.
11. Pigott TD. A review of methods for missing data. *Educ Res Eval* 2001;7:353-83.
12. Samet JM, Munoz A. Cohort Studies. *Epidemiol Rev* 1998;20:1-136.
13. Swiss HIV Cohort Study, Schoeni-Affolter F, Ledergerber B, Rickenbach M, Rudin C, Günthard HF, *et al*. Cohort profile: The Swiss HIV Cohort study. *Int J Epidemiol* 2010;39:1179-89.
14. Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB. *Designing Clinical Research*. 2nd ed. Philadelphia, USA: Lippincott Williams and Wilkins; 2001.
15. Swiss HIV Cohort Study. Available from: <http://www.shcs.ch/>. [Last accessed on 2015 Nov 14].
16. Szklo M, Nieto FJ. *Epidemiology: Beyond the Basics*. Sudbury, MA: Jones and Bartlett Publishers, Inc.; 2004.
17. Snijders TA, Bosker RJ. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. London, UK: Sage Publications; 2012.

Copyright of Indian Journal of Dermatology is the property of Medknow Publications & Media Pvt. Ltd. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.