Chapter 4
Cohort Study

Li Liu

Key Points

- A cohort study is an observational study which begins with a group of people who are free of an outcome of interest and classified into subgroups according to the exposure to a potential cause of the outcome. Variables of interest are specified and measured, and the whole cohort is followed up in order to see how the subsequent development of new cases of the disease (or other outcomes) differs between the exposed and unexposed groups.
- There are three types of cohort studies according to the time when information on exposures and outcomes is collected, namely prospective cohort study, retrospective cohort study and ambispective cohort study.
- The measures of associations in cohort studies include relative risk, attributable risk or attributable fraction, population attributable risk or population attributable fraction, and dose-effect relationship.

4.1 Introduction

4.1.1 Definition

A cohort study is an observational study which begins with a group of people who are free of an outcome of interest and classified into subgroups according to the exposure to a potential cause of the outcome. Variables of interest are specified and measured, and the whole cohort is followed up in order to see how the subsequent development of new cases of the disease (or other outcomes) differs between the exposed and unexposed groups.

L. Liu (✉)
School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

C. Wang, F. Liu (eds.), *Textbook of Clinical Epidemiology*,
https://doi.org/10.1007/978-981-99-3622-9_4

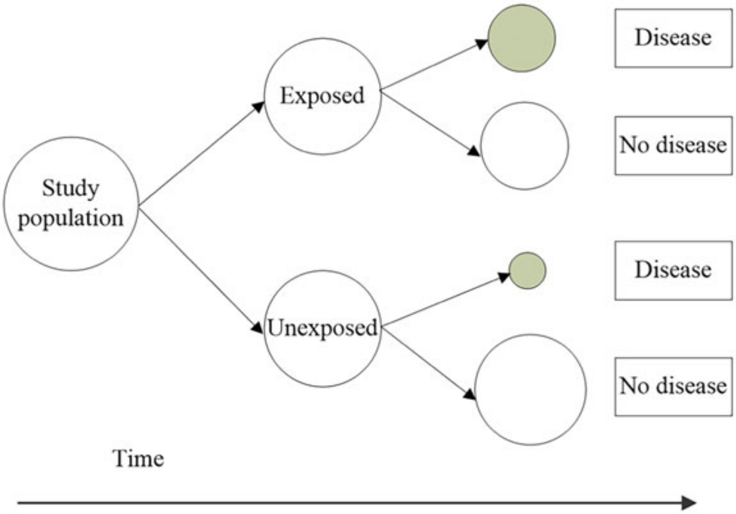


Fig. 4.1 The design of a cohort study

Exposure means that the subject has been exposed to some substances (e.g., heavy metals) or has some characteristics (e.g., being a carrier of a particular genotype) or behaviors (e.g., alcohol drinking).

The term "cohort" is derived from the Roman army, where it referred to a group of about 480 soldiers, or one-tenth of a legion. Soldiers remained in the same cohort throughout their whole military life, similar to members of epidemiologic cohorts.

According to the time of participants entering the study, the cohorts can be classified into two types: the fixed cohort and dynamic cohort. Fixed cohort means that all participants are enrolled in the cohort at a fixed time or in a short period of time, and followed up until the end of the observation period. The participants have not exited due to other reasons than the outcome, and no new members have joined it. During the whole period, the cohort remains relatively stable. Dynamic cohort, also known as an open cohort, refers to a cohort in which the original members continue to withdraw, and/or new members can join in during the follow-up.

The simplest situation of a cohort study is to recruit one group of population with a specific exposure and one group without that exposure and then follow up for a period of time to see if the participants develop the outcome of interest (Fig. 4.1). The participants must be free of the outcome of interest at the start of the follow-up, which makes it easier to be sure that the exposure precedes the outcome. After a period of time, the investigator compares the incidence rates of the outcome between the exposed and unexposed group. The unexposed group serves as the reference group, providing an estimate of the baseline amount of the outcome occurrence. If the incidence rates are substantively different between the exposed and unexposed groups, the exposure is said to be associated with the outcome. According to the basic principles of cohort studies, there are some basic characteristics:

4.1.1.1 Observational Study

The exposures in cohort studies are not given artificially, but objectively before the study, which is an important aspect of the difference between cohort studies and experimental studies.

4.1.1.2 Setting up a Comparison Group

Cohort studies usually set up an unexposed group for comparison during the research design phase. The control group may come from the same population as the exposed group or from different populations.

4.1.1.3 From "Cause" to "Outcome"

In the course of the cohort study, we usually know the "cause" (exposure factors) first, and then look into the "outcome" (disease or death) through longitudinal observation, which is consistent with experimental research.

4.1.2 Types of Cohort Study

There are three types of cohort studies according to the time when information on exposures and outcomes is collected (Fig. 4.2).

4.1.2.1 Prospective Cohort Study

In prospective studies, data on exposures are collected at baseline and updated during the follow-up. The outcomes are not available at the beginning of the cohort and should be collected during the follow-up. The investigators could use the most up-to-date measurements to address exposures of interest with minimized bias. However, the investigators need to wait for a relatively long time until a sufficiently large number of events occur. For rare outcomes, the follow-up period may span one, or even several decades.

4.1.2.2 Retrospective Cohort Study (Historical Cohort Study)

Data on the exposures and outcomes are collected from existing records and can immediately be analyzed. It relies on exposure measurements made before the study set up, which may be available from demographic, employment, medical, or other

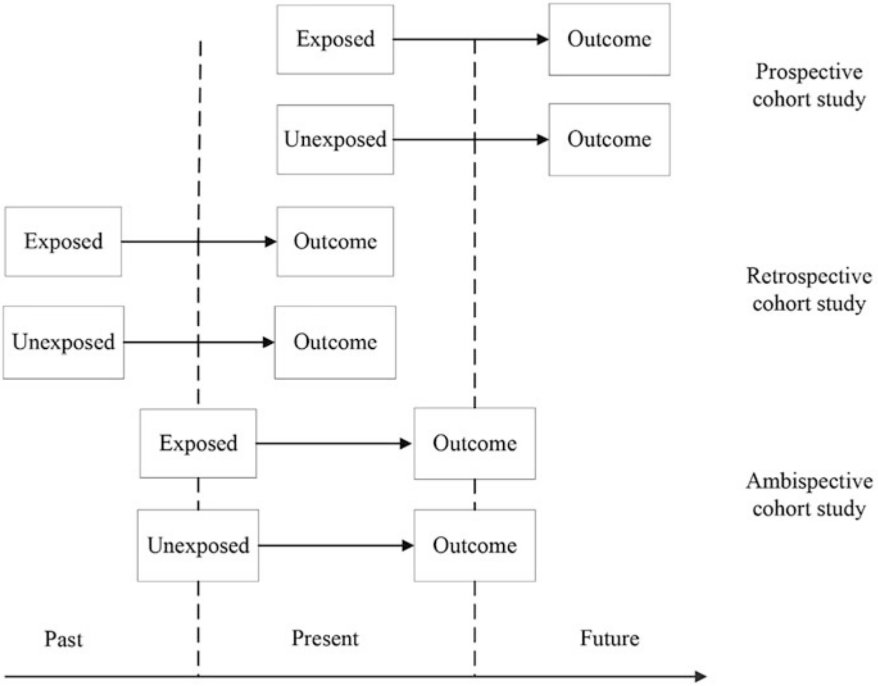


Fig. 4.2 Design of the three types of cohort studies

records. Compared with a prospective cohort study design, it is more useful for rare diseases with a long natural history. By using existing data, the wait time for the exposure to have any impacts on the risk of outcome could be largely reduced. A retrospective cohort study is particularly useful in occupational and environmental epidemiology because if there is a concern that a certain exposure may be a risk factor, it is not reasonable to wait for a long time to confirm in a prospective cohort study. The main disadvantage of a retrospective cohort study is that the exposure data available in records are usually less detailed and accurate than if they were collected prospectively. A retrospective cohort study can be particularly successful when biological specimens were collected in the past so that up-to-date laboratory techniques can be used to detect past exposures. This method could minimize inaccurate exposure measurements in the past, but the stability of the biomarkers during long periods of storage is largely unknown.

4.1.2.3 Ambispective Cohort Study

An ambispective cohort study is a design that combines prospective cohort study with retrospective cohort study. In an ambispective study, a large proportion of participants are still at risk of the outcome when the retrospective cohort are identified, and the follow-up period can be extended into the future to obtain the maximum amount of information from the cohort. So an ambispective cohort study combines the advantages of both retrospective and prospective cohort studies and to some extent, makes up for their respective deficiencies.

4.2 Design of a Cohort Study

4.2.1 Selection of the Cohort

The study population includes both exposure and control (unexposed) population. Depending on the purpose and the conditions of the study, the choice of study population varies.

4.2.1.1 Choice of the Exposure Population

The choice of the exposure population depends on practical considerations and the study hypotheses. There are usually four sources:

General Population

It refers to a well-defined region of the entire population or its samples. It is composed of individuals with different exposure factors and is suitable for simultaneous observation of multiple exposures and their relationships with multiple diseases. When the exposure group is chosen from the general population, there are two points to be considered: ① Do not intend to pay attention to the incidence of special population, but focus on the general population, so that the research results have universal significance. ② The exposure factors and outcomes of interest are very common in the general population, so there is no need to choose special populations or no special population to choose from. The Framingham Study is a well-known example of a general population cohort.

Occupational Population

If you want to study a suspicious occupational exposure factor and an outcome, you should select the relevant occupational population as the exposure group. In addition, records on occupational exposures and diseases are often more comprehensive, true, and reliable, so it is a very good source for retrospective cohort study.

Special Exposed Population

It refers to people with special exposure experiences, which is the only way to study some rare special exposures, such as selecting people who have undergone radiotherapy to study the relationship between radiation and leukemia. If an exposure factor has pathogenic effects, the incidence or mortality of certain disease in special exposed population should be higher than that that in the general population, which could facilitate the identification of the association between the exposure and the disease.

Organized Population

The organized population can be considered as a special form of the general population, such as school and army members. The selection of such a population mainly relies on the relevant organizational system to facilitate the efficient collection of follow-up information. Given the similar occupational experiences, the occupational people are more comparable.

4.2.1.2 Choice of Control Population

The unexposed group should be comparable to the exposed group in the distribution of factors that may be related to the outcome of interest except for the exposure. That is, in case of no association between the exposure and outcome, the outcome would have the same incidence in the exposed group and the unexposed group. The control groups mainly include:

An Internal Comparison Group

An internal comparison group includes unexposed members from the same cohort. Both exposed and unexpected groups are within the selected population. This is usually the best comparison group since the subjects are similar in a lot of aspects. For example, if we want to assess the association between yogurt consumption and the risk of conventional and serrated precursors of colorectal cancer, subjects from the cohort are categorized into groups according to the amount of yogurt consumption, and the group with the lowest intake is used as an internal comparison group, to which the other groups are compared [5]. Cohort studies should try to choose an internal control because it is comparable with the exposed population, easy to conduct, and able to understand the overall incidence of the study.

An External Comparison Group

When it is impossible to take a well-defined cohort and divide it into the exposed and unexposed groups, the comparison group should be selected outside the cohort, which refers to "external comparison group." A potential external comparison group is another cohort with similar characteristics but without the exposure of interest. The advantage of this approach is that the follow-up observation can be protected from the exposure group. The disadvantage is the effort to organize another population.

General Population as a Comparison Group

It can be considered as a kind of external control. The whole population of the geographic area where the exposed cohort is located might be selected as a comparison group (unexposed). Since it is highly impractical to follow up the entire population of a geographic region, incidence rates in the general population are typically derived from routine statistics, which represents an efficient approach compared to studying an additional, unexposed cohort. It takes advantage of existing incidence or mortality statistics across the region, as the morbidity or mortality of the general population is relatively stable and readily available and can save significant time and money, but the disadvantage is that information is often not very accurate. The quality of the information can hardly be checked because it is not collected directly by investigators. Information on potential confounders (other than age, sex, and other basic demographic characteristics) is typically not available in the general population, and confounding by factors such as smoking, cannot be controlled. It should be noted that the control group may contain some exposed subjects, so the total control population applies to a small proportion of the total exposure population. In practice, instead of using a direct comparison of the incidence of the exposed group and the general population, a standardized ratio is used. For example, the standardized mortality rate (SMR) is the ratio of the number of expected morbidity or mortality figures calculated from the incidence or death of exposed groups to the total population.

Multiple Comparison Groups

Multiple comparison groups refer to that more than one group of people listed above should be selected as control. It can reduce the bias caused by using only one kind of control and enhance the reliability of the results. However, multiple comparison groups undoubtedly increase the workload of the research.

4.2.2 Determine the Sample Size

4.2.2.1 Matters to Be Considered when Calculating Sample Size

1. In general, the sample size of the unexposed group should not be less than that of the exposed group, usually the same amount. Small sample size may cause increased standard deviation and unstable results.
2. Due to long-term follow-up of cohort studies, the loss of follow-up is inevitable. An estimated rate of loss to follow-up in advance helps to prevent the analysis from being affected by insufficient sample size in the later stage of the study.

4.2.2.2 Four Factors Affecting Sample Size

1. The incidence of the outcome of interest in the unexposed population p_0.
2. The incidence of the outcome of interest in the exposed population p_1. The greater the difference between the two incidences of the exposed and unexposed populations, the smaller the sample size requires. If the incidence of the exposed group is not easy to obtain, one can try to get the estimate of the relative risk (*RR*) and calculate p_1 by the formula $p_1 = RR \times p_0$.
3. Significance level α: That is the probability of the type I error when making a hypothesis test. The smaller the probability of false positives, the greater the sample size required. α is usually taken as 0.05 or 0.01.
4. Power $(1 - \beta)$: β is the probability of the type II error in the hypothesis test. Power of test refers to the ability to avoid false negatives when testing. The smaller the β, the greater the sample size required. Typically, β is 0.10 and sometimes 0.20.

4.2.2.3 Calculation of Sample Size

If the sample size for the exposed and unexposed groups is the same, the sample size required for each group can be calculated using the following formula:

$$n = \frac{\left(Z_\alpha\sqrt{2\overline{pq}} + Z_\beta\sqrt{p_0q_0 + p_1q_1}\right)^2}{(p_1 - p_0)^2} \tag{4.1}$$

p_0 and $p_1$in the formula represent the incidence of the unexposed and exposed groups, respectively;$\overline{p}$ is the average of the two incidences; $q = 1 - p$; Z_α and Z_β are standard normal distribution limits, which can be found from the Standard Normal Distribution Table.

4.2.3 Follow-Up

The follow-up of participants is a very arduous and important work in a cohort study. It should be planned and strictly implemented.

4.2.3.1 Purpose of Follow-Up

The purpose of follow-up includes three points: identifying whether a subject is still under observation; identifying various outcomes (e.g., disease incidence) in the study population; further collecting data on exposures and confounding factors.

For a variety of reasons, some participants are out of observation during follow-up, a phenomenon known as loss to follow-up, which would have an impact on the findings. When the loss rate is greater than 10%, measures should be taken to further estimate its possible impact. If the loss rate is very high, the authenticity of the study will be seriously questioned. Ensuring follow-up success is therefore one of the keys to successful cohort studies.

4.2.3.2 Follow-up Methods

Follow-up methods include direct face-to-face interviews, telephone interviews, self-administered questionnaires, periodic physical examinations, environmental and disease monitoring, etc. The follow-up methods should be based on follow-up contents, follow-up objects, and manpower, material, and financial resources. It should be emphasized that the same follow-up method should be used for the exposed and comparison groups, and the follow-up method should remain unchanged throughout the follow-up.

4.2.3.3 Follow-up Contents

The contents of follow-up are generally consistent with the baseline data, but the focus of follow-up is the outcome of interest. The specific items may be different depending on the purpose and design of the study. In general, one should mainly collect the following information: ① Study outcomes: whether the study population has some kinds of research outcomes. Suspected patients found for the initial examination should be further confirmed. ② Exposure data: what is the exposure of the study subjects? Is there any change? For example, if the study aims to detect the relationship between smoking and lung cancer, one should ask about the amount of cigarette smoking at baseline and during the follow-up. ③ Other relevant information of the study population: the same as the baseline items. ④ Changes in population information: information on lost or retired population, or new arrivals (dynamic cohorts).

4.2.3.4 Endpoint of Observation

The endpoint of observation means that the subjects develop the desired outcome. For example, when the etiological factor of the disease is studied, often the outcome is the occurrence of the disease or the death caused. When the study subjects develop the outcome of interest, they are no longer observed. In general, the endpoint of the observation is the disease or death, but may also be changes of certain indicators, such as the emergence of serum antibodies and elevated blood lipids, according to the study purpose.

4.2.3.5 Follow-Up Interval

In theory, follow-up should be carried out after the shortest induction period or incubation period of the disease. The follow-up interval depends on the intensity of exposure and the length of the incubation period of the disease. The weaker the exposure or the longer the incubation period is, the longer the follow-up interval needs. The induction or incubation period of chronic disease is not very clear. In general, the follow-up interval of chronic diseases can be set for several.

4.2.3.6 The Termination Time of Observation

The termination time of observation refers to the deadline of entire research work, and the expected time to get the result of interest. The termination time of observation is determined according to the length of the observation period, which depends on the incubation period of the disease. In addition, one should also take into account the amount of person-year. One should try to shorten the observation period on the basis of these principles so as to save manpower and material resources and reduce the number of loss to follow-up.

4.2.4 Quality Control

Cohort studies are by nature time-consuming and expensive. Therefore, the strict implementation process, especially the quality control during data collection, is of particular importance. Generally, the following quality control measures are taken:

4.2.4.1 Selection and Training of the Investigators

Investigators should maintain strict work ethic and scientific attitude. Honesty and reliability are the basic qualities that investigators should possess. Generally, investigators should possess the expertise and knowledge required for the investigation. The work ethic, scientific attitude, survey techniques of investigators, and the experience of clinical doctors and laboratory technicians will affect the reliability and authenticity of the survey. Therefore, before data collection, investigators should be trained for better performance during the investigation.

4.2.4.2 Preparation of an Investigator's Handbook

Due to the large number of investigators involved and the long duration of follow-up in cohort studies, an Investigator's Handbook, including operating procedures, precautions, and a complete description of the questionnaire is necessary.

4.2.4.3 Supervision during the Follow-Up

Common supervision measures include: repeating the survey among some participants by another investigator, checking numerical or logical errors, comparing the distribution of variables collected by different investigators, analyzing temporal trends of variables, and recording the interviews by using tape recorders, etc.

4.3 Data Collection and Analysis

4.3.1 Data Collection

The investigators should first collect the baseline information of every participant, mainly including information on exposure status (e.g., the type, duration, and dose of the exposure), personal characteristics (e.g., health status, age, gender, occupation, educational level, marital status), and other circumstances (e.g., home environment, lifestyle and family history of disease). Participants are followed over time, and baseline information is compared with later follow-ups. It also works as a basis to characterize baseline exposures (e.g., classify individuals into exposed or unexposed group, ascertain degrees of exposure and potential confounders), and to obtain tracking materials for follow-up and key information for inclusion or exclusion. The major methods to collect baseline information include data records (e.g., employment, medical examinations, insurance), questionnaires or interviews, physical examinations and tests of biological samples, as well as environmental measurements. Besides baseline information, data collection throughout the process of follow-up is also important (e.g., changes of exposures and measurements of outcomes over time). For more detailed information, please see the second section on follow-up of this chapter.

4.3.2 Measures of Outcome Frequency

4.3.2.1 The Basic 2 × 2 Tables Summarizing the Results of a Cohort Study

Disease incidence could be described by the cumulative incidence or incidence density. Cumulative incidence is generated by dividing the number of incident cases by the number of persons at risk in the cohort, as shown in Table 4.1:

The cumulative incidence in the exposed group $= d_1/n_1$.
The cumulative incidence in the unexposed group $= d_0/n_0$.

The cumulative incidence represents the individual risk of developing the disease of interest with no unit. It is a proportion, not accounting for possible different periods of follow-up time, thus mainly used in fixed cohorts. When studying acute outcomes within a short period of follow-up, such as outbreaks, cumulative incidence could be used to estimate the risk of the disease, given a fixed period of follow-up. However, in most circumstances, such as chronic disease research, the periods of follow-up are relatively long; thus, the cumulative incidence is no longer appropriate since the follow-up time usually differ across cohort members. In this situation, the outcome of interest is preferably described by rate, which is incidence density, the other index to reflect disease incidence, and it is widely utilized in dynamic cohorts. Incidence density is calculated by dividing the number of outcome events by the person-time at risk, as shown in Table 4.2:

Incidence density in the exposed group $= d_1/T_1$.
Incidence density in the unexposed group $= d_0/T_0$.

One should note that a person "at risk" refers to the fact that the outcome of interest can occur within the given time frame. Thus if subjects are immune, they are no longer at risk of getting this disease. If on the other hand, the event of interest is uterine cancer, a hysterectomized woman would not be "at risk." Measurements of risk and incidence of disease could provide valuable information related to the public health burden of the outcome of interest, which is important for disease prevention and public health management.

Table 4.1 Measures of cumulative incidence

Exposure status	Cases	Non-cases	Total	Cumulative incidence
Exposed	d_1	$n_1 - d_1$	n_1	d_1/n_1
Unexposed	d_0	$n_0 - d_0$	n_0	d_0/n_0

Table 4.2 Measures of incidence density

Exposure status	Cases	Person-time at risk	Incidence density
Exposed	d_1	T_1	d_1/T_1
Unexposed	d_0	T_0	d_0/T_0

4.3.2.2 Person-Time

In a dynamic cohort, study subjects have unequal periods of time from entry into the cohort to disease occurrence or end of follow-up, and this must be taken into account. Person-time is introduced to reflect the exposure experience of a subject in this circumstance. Total person-time is the summation of the time at risk of individual cohort members to develop the disease, which is often the denominator of the incidence density. The common unit of person-time is person-year. As shown in Table 4.3, people entered the cohort at different ages and experienced separate lengths of time. Before the end of the follow-up, four subjects were diagnosed with disease of interest. The person-years of each person are presented in the last column, and the total person-time in this example is 91 person-years.

This exact computation method is based on the duration of participation of each individual; however, for large cohorts, one may not obtain detailed information for each participant, then approximation method is an alternative though with less precision. The approximate person-years are considered as the average number of the population multiplied by the number of years of observation. The average number of the population refers to the average number of the population at the beginning of two contiguous years or the number of the population in the middle of a specific year. In a hypothetical cohort study which started on September 1, 2014, and finished on September 1, 2017, the numbers of subjects were 15,262 in the beginning, and 15,276 at the end, and more details are shown in Table 4.4. The average population in the 20–29 age group are 26,203 persons: (8724 + 8736) /

Table 4.3 Data from a fictitious cohort

Person ID	Age at entry	Years of follow-up	Age at end of follow-up	Age at diagnosis	Person-years at risk
1	34	14	48		14
2	37	20	57	52	15
3	30	12	42		12
4	33	17	50	41	8
5	37	9	46		9
6	38	16	54	49	11
7	43	11	54		11
8	32	20	52	43	11
Total		120			91

Table 4.4 Numbers of subjects in a hypothetical cohort study at different times stratified by age groups

Age groups	2014-09-01	2015-09-01	2016-09-01	2017-09-01
20–29	8724	8736	8740	8730
30–40	6538	6570	6554	6546
Total	15,262	15,306	15,294	15,276

Table 4.5 Data from a fictitious cohort study to calculate person-years with simple life table

Observing time (x)	No. of objects				Pearson-years (T_x)
	At the beginning (N_x)	Entering the cohort (E_x)	Occurring outcome events (D_x)	Lost to follow-up (L_x)	
2011	1898	76	4	22	1923
2012	1948	70	6	18	1971
2013	1994	52	7	15	2009
2014	2024	30	5	19	2027
Total					7930

2 + (8736 + 8740)/2 + (8740 + 8730)/2 = 26,203. The average population is then multiplied by the number of follow-up years to get the person-time.

Another method to calculate the person-time is to utilize simple life table. The basic equations are as follows:

$$T_x = N_x + \frac{1}{2}\ (E_x - D_x - L_x) \tag{4.2}$$

$$N_{x+1} = N_x + E_x - D_\mathrm{x} - L_x \tag{4.3}$$

x refers to a certain period of time, usually representing 1 year; T_x is the person-time during x time; N_x is the number of population at the beginning of x time; E_x is the number of subjects entering the cohort during x time; D_x is the number of occurring outcome events during x time; and L_x is the number of subjects who are lost to follow-up.

According to the equations above, one can get a simple life table, and the total person-years are the sum of every T_x.

For example, according to Table 4.5, the person-years in 2011 are

$$\begin{aligned} T_{2011} &= N_{2011} + \frac{1}{2}(E_{2011} - D_{2011} - L_{2011}) \\ &= 1898 + (76–4\ –22)\ \ /2 = 1923 \end{aligned}$$

The number of population at the beginning of 2012 is

$$\begin{aligned} N_{2012} &= N_{2011} + E_{2011} - D_{2011} - L_{2011} \\ &= 1898 + 76–4–22 = 1948 \end{aligned}$$

So the person-years in 2012 are

$$T_{2012} = 1948 + (70\text{–}6\ -\ 18)\ /2 = 1971$$

By that analogy, person-years in 2011, 2012, 2013, and 2014 are 1923, 1971, 2009, and 2027, respectively, and the total person-years are 7930.

4.3.2.3 Standardized Mortality Ratio (SMR)

For cohorts with a general population comparison group, one usually estimates the association between an exposure and an outcome by calculating standardized mortality (or incidence) ratios (SMRs). The SMR is the ratio of the observed number of deaths in the cohort and the expected number of deaths in the cohort, given the age-specific mortality rates of a reference population and the age structure of the cohort.

$$\mathrm{SMR} = \frac{\sum_{i=1}^{I} n_i}{\sum_{i=1}^{I} t_i \times a_i} \tag{4.4}$$

Where I stands for the age group, n_i denotes the number of observed deaths of the age group, t_i denotes the number of person-years in the age group, and a_i represents the age-specific mortality rate of the age group from the reference population.

The SMR is commonly adjusted for age, calendar period, and other characteristics like race. Example: There were 1000 workers aged between 40 and 50 in a factory, and four of them died of lung cancer in 2000. Assuming that the mortality of lung cancer among the total population aged between 40 and 50 is 2‰ in 2000, then the expected number of death is 2, and we have known that the practical number of deaths is 4; thus the SMR is 2 (4/2 = 2).

4.3.2.4 Statistical Tests

To test the statistical difference of incidence rate between the exposed and unexposed groups, U test is commonly used in practice. However, there are some noteworthy conditions to abide by relatively large sample size, not too small p (incidence rate) and $1 - p$ (e.g., $n \times p$ and $n \times (1 - p)$ are both over five), and approximately normal distribution of incidence rates.

$$u = \frac{p_1 - p_0}{\sqrt{p_c\,(1 - p_c)\,\left({}^{1}\!/_{n_1} + {}^{1}\!/_{n_0}\right)}} \tag{4.5}$$

p_1 and p_0 are incidence rates in the exposed group and the unexposed group, respectively; n_1 and n_0 are numbers of subjects in the exposed and unexposed groups, respectively; and p_c is incorporative sampling rate ($p_c = \frac{X_1+X_0}{n_1+n_0}$, X_1 and X_0 are the numbers of outcome events in the exposed and unexposed groups, respectively). One should subsequently compare the U value with the standard U table, then seek out the corresponding P value and make inference based on the significant level.

Other statistical tests include probabilistic methods based on binomial or Poisson distribution, Chi-Square test, or score test. Similarly, it is notable that each test has its conditions.

4.3.3 Measures of Association

4.3.3.1 Relative Risk (RR)

RR refers to the ratio of the probabilities of an outcome in the exposed and unexposed groups. Its value is a positive real number with a range from 0 to +∞, and could take the following form:

$$\mathrm{RR} = \frac{I_1}{I_0} \tag{4.6}$$

I_1 and I_0 refer to risk or rate of outcome in the exposed and unexposed groups, respectively.

There are two alternative and equivalent expressions: the risk ratio and the rate ratio.

Risk ratio is based on the cumulative incidence, with not accounting for person-time. In Table 4.1, risk ratio could be expressed as:

$$\mathrm{RR} = \frac{{}^{d_1}\!/_{n_1}}{{}^{d_0}\!/_{n_0}} \tag{4.7}$$

Rate ratio is the most natural way to express relative risk. It uses incidence density, which takes person-time into account. In Table 4.2, the rate ratio would then be:

$$\mathrm{RR} = \frac{{}^{d_1}\!/_{T_1}}{{}^{d_0}\!/_{T_0}} \tag{4.8}$$

One can also estimate the 95% confidence interval (CI) of the RR using the Woolf method based on the variance of RR. According to Table 4.1, the variance of *ln* RR is computed as follows:

Table 4.6 General criteria to estimate the strength of association of relative risk

Relative risk		Strength of association
1.0–1.1	0.9–1.0	None
1.2–1.4	0.7–0.8	Weak
1.5–2.9	0.4–0.6	Moderate
3.0–9.9	0.1–0.3	Strong
10-	<0.1	Infinite

Monson [6]

$$Var(ln\ \mathrm{RR}) = \frac{1}{d_0} + \frac{1}{d_1} + \frac{1}{n_0 - d_0} + \frac{1}{n_1 - d_1} \tag{4.9}$$

and

$$95\%\mathrm{CI\ of}\ ln\,\mathrm{RR} = ln\,\mathrm{RR} \pm 1.96\sqrt{Var(ln\ \mathrm{RR})} \tag{4.10}$$

One could obtain the 95% CI of RR by taking the antinatural logarithm of 95% CI of *ln*RR.

Risk ratio and rate ratio have the same epidemiological implication, but their values are usually different in the same study. The interpretation of the relative risk is as follows:

If RR $>$ 1, the risk of disease for the exposure is increased compared with the unexposed group;
If RR $<$ 1, the risk of disease for the exposure is decreased compared with the unexposed group;
If RR $=$ 1, there is no association.

The risk in the reference group multiplied by the corresponding RR approximates the risk in the exposed group. The value of RR reflects the level of association. Here are the general criteria to estimate the correlation intensity (Table 4.6):

4.3.3.2 Attributable Risk (AR) and Attributable Fraction (AF)

The RR mainly measures the level of risk associated with the exposure to a risk factor. It cannot reflect the impact of the factor in a population. To address this issue, AR and AF are introduced. RR mainly provides clues for etiology, while AR and AF are important for disease prevention and public health. AR, also known as the risk difference or excess risk, is the measure of the rate of disease related to the exposure to a risk factor. Attributable risk is applied to quantify risk in the exposed group which could be attributable to the exposure.

$$\mathrm{AR} = I_1 - I_0 = \frac{d_1}{n_1} - \frac{d_0}{n_0} \tag{4.11}$$

or

$$\mathrm{AR} = I_1 - I_0 = \mathrm{RR} \times I_0 - I_0 = I_0(\mathrm{RR} - 1) \tag{4.12}$$

AF is the proportion of the total number of cases related to the exposure to a risk factor. It allows to calculate the proportion of disease attributable to the exposure in the exposed group. This can also be viewed as the proportion of disease in the exposed group that can be avoided through the elimination of the risk factor. It is calculated by dividing the risk difference by the incidence of disease in the exposed group and then multiplying it by 100 to convert it into a percentage

$$\mathrm{AF} = \frac{I_1 - I_0}{I_1} \times 100\% \tag{4.13}$$

or

$$\mathrm{AF} = \frac{\mathrm{RR} - 1}{\mathrm{RR}} \times 100\% \tag{4.14}$$

AR and AF are both calculated from incidence rates. One should note that they only make sense for a causal association of a risk factor with an outcome occurrence. The underlying assumption is that no other potential confounders are involved in the occurrence of the outcome.

4.3.3.3 Population Attributable Risk (PAR) and Population Attributable Fraction (PAF)

PAR estimates the proportion of disease attributed to the exposure in the study population. PAR can be looked at as the proportion of a disease that could be prevented by eliminating a causal risk factor from the population. PAR tends to be a function of time because both the prevalence of a risk factor and its effect on the exposed population may change over time, as may the underlying risk of disease. Definitions for PAR and PAF are given by

$$\mathrm{PAR} = I_t - I_0 \tag{4.15}$$

$$\mathrm{PAF} = \frac{I_\mathrm{t} - I_0}{I_t} \times 100\% \tag{4.16}$$

Where I_t represents the incidence of disease in the total population, and I_0 indicates the incidence of disease in the absence of exposure.

PAF is also given as:

$$\text{PAF} = \frac{P_e(\text{RR} - 1)}{P_e(\text{RR} - 1) + 1} \times 100\% \tag{4.17}$$

Where the prevalence of exposure "P_e" is the proportion of individuals exposed to the risk factor.

4.3.3.4 Dose-Effect Relationship

In some circumstances, there may exist a dose-effect relationship between the exposure and the outcome. To address this, one could stratify the exposure into several levels, with defining the lowest level as a reference, and then calculate RRs of other groups compared to the referent group. Taking Table 4.7 as an example, along with the increase of serum cholesterol level, the relative risk of developing coronary heart disease also increases, which indicates that there may exist a dose-effect relationship between serum cholesterol levels and incidence of coronary heart disease. If necessary, one can further make a trend test.

4.4 Common Bias and Controlling

4.4.1 Selection Bias

Selection bias occurs when the selection of the exposed and unexposed individuals is related to the occurrence of the outcomes of interest. This is a major potential problem in retrospective cohort studies, since knowledge about the exposure and outcome is likely to differentially influence participants. However, it is generally not a problem in prospective cohort studies, since the outcome of interest has not occurred. A serious potential concern is loss to follow-up in prospective cohort studies [7], which arises when study subjects refuse to participate in or cannot be

Table 4.7 The occurrence of coronary heart disease stratified by serum cholesterol levels in a fictitious cohort study

Cholesterol level	No. of participants	No. of cases	Risk	Relative risk
Very low	200	2	0.01	1(reference)
Low	300	15	0.05	5
Intermediate	400	40	0.1	10
High	300	60	0.2	20
Very high	100	30	0.3	30

found for the data collection during follow-up. Retention of subjects might be differentially related to both exposure and outcome, and this brings a similar effect that can prejudice the results, causing either an underestimate or an overestimate of an association. For example, if an exposed individual will develop the outcome in the future, but she/he is more likely to be lost to follow-up, then the exposed incidence will be underestimated, along with the RR tending towards the null. Loss to follow-up can result in bias and reduce the statistical power. The primary way to reduce this bias is to improve compliance and response rate of participants.

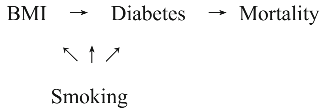
4.4.2 Information Bias

Similar to selection bias, information bias occurs in different ways under different study designs. Reporting bias is one of the potential information biases in cohort studies since the exposure status may influence the reporting of the outcome. For example, in an investigation about occupational hazard, workers are more likely to report having experienced various harmful exposures when this refers to labor guarantee or benefits; thus, some associations may be overestimated. If possible, it would be better to utilize some objective methods and sources of data, such as medical records and laboratory tests, to ascertain the exposure and outcome status. Another important form of information bias is detection bias. Detection bias occurs when knowledge of exposure status differentially increases the likelihood of detecting the outcome of interest among the exposed in cohort studies. A typical example is that a medically relevant exposure could bring about more medical visits and an increased possibility of a diagnostic evaluation, which increases the probability of detecting the outcome in the exposed group. An effective way to address this issue is to apply blinding method to collect information.

Besides, other factors may also contribute to information bias. For example, in the collection of laboratory data, the quality of instruments and reagents, selected measurement standard, measuring conditions and technical competence of the operator are all potential factors influencing the results. Additionally, scientific questionnaires and complete records are also imperative.

4.4.3 Confounding

Except for selection bias and information bias, confounding is also an important factor that can cause systematic bias in epidemiology, thus the investigators must consider it from study design to data analysis. Confounding distorts the underlying correlation of the exposure with the outcome of interest. The factors causing confounding are called confounders. The criteria for a factor to become a confounder are as follows: the factor must be related with both the exposure and the disease of interest, and at the same time it must not be an intermediate variable in the causal chain between the exposure and the disease of interest. Directed Acyclic Graph (DAG) is an effective method to distinguish a confounder and a collider. In the following example:



Smoking is a confounder when exploring the association between BMI and the prevalence of diabetes, or the association between the prevalence of diabetes and mortality. However, when exploring the association between smoking and BMI, diabetes acts as a collider (a variable directly affected by two or more other variables with arrows pointing to itself in the DAG, but not the other way around).

In cohort studies, confounding occurs when risk factors are unevenly distributed between the exposed group and the unexposed group. The major methods to control confounding are restriction on inclusion criteria, randomization, and matching. Besides, statistical procedures such as standardization, stratification analysis, and multivariate analysis are also available.

4.5 Advantages and Disadvantages of Cohort Studies

4.5.1 Advantages of Cohort Studies

1. Strong ability to identify cause-effect association because of the temporal relationship between the exposure and the outcome, reliable data personally observed by researchers and computable indicators reflecting relevance intensity such as RR, AR, etc.
2. Helpful in understanding the natural history of disease in the population.
3. Unexpected outcome data are obtained to analyze the relationship between multiple outcomes and a cause.
4. Able to study the effects of rare exposures.
5. Avoiding recall bias at enrollment.

4.5.2 Disadvantages of Cohort Studies

1. It is not suitable for disease with low morbidity because large sample size is needed.
2. In a long follow-up period, lost to follow-up of subjects would cause bias.
3. A large amount of manpower, material resources, and financial resources are required.

4. During the follow-up, the entry of unknown variables and the changes of known variables could influence the outcome, making the analysis complicated.

4.6 Example of a Cohort Study

To facilitate the understanding of cohort studies, the design, implementation and main results of a cohort study **“Fresh Fruit Consumption and Major Cardiovascular Disease in China** [8]” is cited. This study is from The China Kadoorie Biobank Study a nationwide, prospective cohort study involving 10 diverse localities (regions) in China. For more details, please see *Du H, Li L, Bennett D, Guo Y,* et al. *Fresh Fruit Consumption and Major Cardiovascular Disease in China [J]. N Engl J Med. 2016;374(14):1332-1343.*