Chapter 3
Descriptive Study

Zhenxing Mao and Wenqian Huo

Key Points

- Descriptive studies are mainly used by observing, collecting, and analyzing relevant data to describe the distribution of disease, health status, and exposure and generate hypotheses for further investigations.
- Descriptive studies mainly include case and case series report, cross-sectional studies, and ecological studies.
- The ability of descriptive studies to prove whether it is a causal association or coincidental phenomenon between exposure and outcome is limited.
- Selection bias, information bias, confounding bias are three major sources of bias in cross-sectional study. Ecological fallacy and confounding factors are the main limitations in ecological study.

Descriptive study, also known as descriptive epidemiology, is the most basic type of epidemiological research method. Descriptive studies are mainly used for describing the distribution of disease, health status, and exposure and generating hypotheses for further investigations but cannot tell causal relations between disease and exposure. Descriptive studies are also mainly used for ascertaining high-risk individuals and evaluating the effects of public health measures, etc. Descriptive studies mainly include case and case series reports, cross-sectional studies, and ecological studies.

Z. Mao (✉) · W. Huo (✉)
College of Public Health, Zhengzhou University, Zhengzhou, China
e-mail: huowenqian@zzu.edu.cn

C. Wang, F. Liu (eds.), *Textbook of Clinical Epidemiology*,
https://doi.org/10.1007/978-981-99-3622-9_3

3.1 Introduction

3.1.1 Concept

Descriptive study is a research method that describes the distribution of diseases or health status and their influencing factors at different times, regions, and populations without changing the current disease status and exposure characteristics of the subjects.

3.1.2 Characteristics of Descriptive Studies

1. Descriptive studies take observation as the main research method and do not impose any intervention measures on research subjects. Only by observing, collecting, and analyzing relevant data do descriptive studies analyze and summarize the distribution of diseases, health conditions, relevant characteristics, and exposure factors.
2. Descriptive studies generally do not set up a control group. And the ability to prove whether it is a causal association or coincidental phenomenon between exposure and outcome is limited. However, it could provide a preliminary contribution to subsequent studies.
3. Descriptive studies have a shorter duration. The distribution of disease and health status in a population is typically analyzed for transient or temporal characteristics. However, it is easy to implement. The distribution of disease and risk factor distribution can be obtained in a relatively short time.

3.1.3 Application

1. To describe the prevalence of disease in different regions and different population characteristics. Continuous descriptive studies at different intervals can also provide time trend data of disease.
2. To describe the regional, population, and temporal distribution of risk factors.
3. To provide etiological clues and form a preliminary etiological hypothesis.
4. Through descriptive research, patients at early or different stages can be found and accept early treatment. At the same time, patients with different disease stages and different infection patterns in the population can also be found. So it can be used to study the natural history of diseases.
5. To provide baseline data as the basis for the longitudinal study.
6. Descriptive studies can evaluate the effectiveness of preventive and control measures in the same population before and after the implementation of interventions.

3.2 Case and Case Series Report

3.2.1 Concept

Case reports usually study a newly discovered or specific disease and its characteristics. A complete case report includes the patient's epidemiological data, such as pre-onset lifestyle characteristics and history of exposure to suspected risk factors. In the case of infectious diseases, attention should also be paid to investigating and reporting the possible exposure to patients, animals, and the environment before and after the onset of illness.

Case series reports are conducted on the basis of case reports used for describing a series of clinical features or cases with similar diagnoses. The content of the case series report is exactly the same as that of the case report mentioned above. However, it should be emphasized that the case series reports should pay more attention to the demographic characteristics of each case, especially the similarity of risk factor exposures and clinical characteristics. Focusing on the chronological sequence of cases and their interconnections is more conducive to forming etiological hypotheses. Generally, case series reports often provide evidence better than case reports.

Hospitals are important places to detect the potential new and special cases; case reports and case series reports are usually carried out by clinicians. Only those with systematic epidemiological training and keen insight can catch abnormal cases in daily diagnosis and treatment activities and report to the local CDC in time. Measures are also taken to prevent further spread of the disease. Both approaches are applicable to infectious and chronic non-communicable diseases.

3.2.2 Application

3.2.2.1 Identifying New Diseases

When a new disease occurs, it is necessary to describe the clinical, demographic, and lifestyle characteristics of the patients, behavioral risk factors, the characteristics of working and living environment in detail. Then, we explore the possible causes for diagnosis and make prevention. On the basis of the above research contents, clinicians can also evaluate treatment measures and effects and expand the research contents.

3.2.2.2 Establishing the Diagnosis

Based on the clinical symptoms, signs, and laboratory examination results of the patients provided by the case reports and case series reports, by combining with the patient's demographic characteristics and epidemiological data (lifestyle characteristics, targeted risk factor exposure history, time and place of onset, etc.), summarizing the common clinical and epidemiological characteristics of patients, diagnostic criteria can be established for the identification and diagnosis of subsequent similar diseases.

3.2.2.3 Forming an Etiological Hypothesis

From the characteristics of individual cases, it can be preliminarily speculated that some characteristics may be associated with the onset of disease. The characteristics of multiple patients can be obtained from the case series reports. Analyzing the characteristics of these patients can provide more information about the relationship between exposure and disease. On these bases, it can form a preliminary hypothesis that a certain characteristic may be the cause of the disease. However, the power to provide evidence is very weak because of the limitations of this approach. It is a very preliminary etiology suggestion and further research using other epidemiological methods and causal demonstration is needed to validate this etiological hypothesis.

3.2.2.4 Identifying Early Disease Outbreaks and Epidemics

The early manifestations of disease outbreaks and epidemics usually occur in one case and then in several cases, followed by more cases of the same characteristics in susceptible contacts. If the outbreak is not identified and controlled early, the disease can continue spreading through a population, leading to outbreaks and epidemics. Therefore, clinicians should have a keen epidemiological thought and vision. When encountering unusual disease or disease symptom and sign, they should be very vigilant. If this may be a sign of an early outbreak and epidemic of a certain infectious disease, clinicians should report to the local CDC timely and take corresponding preventive and control measures.

3.2.3 Case

3.2.3.1 Estrogen Chemical Bisphenol a and Breast Cancer

A case series report described 15 cases of breast cancer in young women. Nine of the women reported consuming food packaged with estrogenic chemical bisphenol A (BPA) at least once a week, and urine samples of nine patients demonstrated the presence of BPA.

3.2.3.2 Occupational Exposure to Vinyl Chloride and Hepatic Hemangioma

In 1974, Creech and Johnson reported that three of the workers in the vinyl chloride plant were found to have hepatic hemangioma. Three of these patients are clearly unusual in such a small population, and it is easy to form the cause hypothesis that "the occupational exposure to vinyl chloride caused the occurrence of hepatic hemangioma." In the following year, this hypothesis was confirmed by data from two analytical studies. If there is only one patient, it is not enough to form the cause hypothesis.

3.2.3.3 AIDS Discovery Case

From October 1980 to May 1981, a report of pneumocystis pneumonia was found among young, healthy gay men and women in Los Angeles, United States. This series of reports was unusual because pneumocystis pneumonia previously only occurred in elderly cancer patients with inhibition of the immune system due to chemotherapy. At the beginning of 1981, many cases of Kaposi's sarcoma were found in young gay men, which is also a noteworthy new discovery. Because this malignant tumor always occurs in the elderly, and the chances of men and women are equal. As a result of these extraordinary discoveries, the US Centers for Disease Control and Prevention immediately implemented monitoring to determine the severity of the problem and developed diagnostic criteria for this new disease. It is quickly noted by monitoring that homosexuals have a high risk of developing the disease. Subsequent case reports and serial case reports indicate that AIDS can also occur through blood transmission in intravenous drug users, blood transfusion patients receiving blood transfusions, and hemophilia patients with blood products. This descriptive data provided clues for the design and implementation of analytical studies and subsequently identified a range of specific risk factors for AIDS. Serum obtained from these cases and comparable controls helped identify the pathogen of AIDS, the human immunodeficiency virus (HIV).

3.2.4 Bias

1. The results are of high promiscuity. The patient is in a natural clinical environment, and the doctor may not be able to control the patient's ability to seek and receive other treatment or control the patient's diet and daily life, which may affect the clinical outcome of the disease.
2. The absence of a control group precluded causal inference.
3. The results are less generalizable. Because cases and case series reports are individual. Strictly, it is almost impossible to find other cases of the same

condition in reality. Usually, based on their own knowledge and experience, doctors would choose the case reports which have the most consistent key characteristics for reference.

4. There is a serious publication bias.

3.2.5 Limitation

Although case reports and case series reports are useful in forming etiological hypotheses, their limitations may overrule causal inference.

1. The incidence of disease cannot be obtained from case reports and case series reports. The case report/case series report lacks the population of patients with a disease that is necessary to calculate the disease rate. For example, when calculating the proportion or incidence of breast cancer in women exposed to BPA, the total number of people exposed to BPA or the total number of years must be clear.
2. Case reports and case series reports lack a control group. In the above example, 60% (9/15) of the 15 breast cancer cases were exposed to BPA. The exposure rate appears to be high, but what is the exposure rate in women who do not have breast cancer? This comparison is key to the hypothesis that BPA may be the cause of breast cancer, but it is absent in case reports and case series reports.
3. The cases described in case reports and case series reports are often highly selective subjects, which could not represent the general population well. For example, 15 cases of breast cancer may be from a community hospital with the same severe air pollution or other potential carcinogen concentrations. In this case, a reasonable estimate of the incidence of breast cancer in women in the same community that is not exposed to BPA is needed to infer the relationship between BPA and breast cancer so as to avoid overestimating the link between the two. At the same time, these highly selected cases are highly likely to be reported early, and more cases need to be accumulated, including atypical cases of clinical stages (especially in the middle and late stages), to see the complete history of the disease.
4. There is sampling variability in case reports and case series reports because there might be large natural variations as the disease progresses. The number of cases needs to be increased to estimate the incidence of disease accurately and eliminate the effects caused by chance or sample variation.

3.3 Cross-Sectional Study

3.3.1 Concept

Cross-sectional study is an epidemiological study that describes the distribution of disease or health status among a specific group of population at a specific time and explores the relationship between variables and disease or health status. Cross-sectional study can get the prevalence of diseases, so it is known as prevalence study. Through cross-sectional study, the occurrence of certain diseases, abnormalities, and vital events in the population can be learned about.

3.3.2 Application

1. To describe the distribution of diseases or health status and provide clues for disease etiology study.
2. Identifying high-risk groups is the first step in early detection, diagnosis, and treatment for chronic diseases.
3. Repeated cross-sectional surveys at different stable stages can not only obtain baseline data of other types of epidemiological studies but also can evaluate the effectiveness of disease monitoring, vaccination, and other prevention and control measures by comparing the prevalence differences at different stages.

Cross-sectional study is the basis and starting point of epidemiological research as well as one of the foothold of public health decision-making. It is a prominent position in epidemiology. Cross-sectional study could not only accurately describe the distribution of disease or health status in a population but also explore the relationship between multiple exposure and disease. But the statistical correlations between disease and exposure revealed by cross-sectional study, which only provides clues to establish causal associations, are derived from analytical studies and cannot be used to make causal inferences.

3.3.3 Classification

Cross-sectional study can be divided into census and sampling survey according to the scope of research objects involved. In the actual work, the use of census or sampling survey mainly depends on the purpose of the research, the characteristics of the research topic, funds, manpower, material resources, and implementation difficulty.

3.3.3.1 Census

Concept

Census refers to the survey of all the people in a specific time or period and within a specific range as research objects. A specific time or period means a short time. It can be a certain time or a few days. For too long, disease or health conditions in the population could change, which may affect census results. A specific range refers to a particular area or population.

Purpose

① Early detection, diagnosis, and treatment can be achieved through census, such as cervical cancer screening in women.
② The prevalence of chronic diseases and the distribution of acute infectious diseases, such as the prevalence of hypertension among the elderly and the distribution of measles in children, can be obtained through the census.
③ Through a census, the health status of local residents can be obtained, such as residents' diet and nutrition status survey.
④ The distribution of disease and its risk factors can be comprehensively understood through census, and the relationship between risk factors and disease can be preliminarily analyzed to provide clues for etiological research.
⑤ In a census, all subjects are investigated through a questionnaire or physical examination. In this process, health education could be conducted to popularize medical knowledge.
⑥ The normal range of index of all sorts of physiology and biochemistry of the human body can be obtained, just like the measurements of teenage height and weight.

Conditions for Carrying out the Census

① Sufficient manpower, material resources, and equipment are available for case detection and treatment.
② The prevalence of diseases should be higher so that more patients can be found and the benefits of census can be improved.
③ The disease detection method should be simple, easy to operate, and easy to implement in the field. The experiment should have high sensitivity and specificity.

Strengths and Limitations

Strengths

Census surveys all members of a defined population, and there is therefore no sampling error in the census, and it is relatively simple to determine the respondents. Census can provide a comprehensive understanding of the health status and the distribution of diseases or risk factors in a population to establish physiological reference values.

All cases in the population can be found through the census, which provides clues for etiological analysis and research to help with prevention.

Through the census, a comprehensive health education and health promotion activities can be carried out to publicize and popularize the medical knowledge.

Limitations

① It is not suitable for the investigation of disease with low prevalence and complex diagnosis methods.
② Due to the heavy workload and short survey period, it is difficult to carry out an in-depth and detailed investigation, and there may be missed diagnosis and misdiagnosis. The proportion of no response may be high, affecting the representativeness of the research results.
③ Due to the large number of staff participating in the census, the variety of their proficiency in techniques and methods would increase the difficulties to control the quality of the survey.
④ Only prevalence or positive rate can be obtained, but not incidence.
⑤ Due to the relatively large scope of population involved in census, more manpower, material resources, and time are consumed in research.

3.3.3.2 Sampling Survey

Concept

Through random sampling, a representative sample of the population at a specific time point and within a specific range is investigated, and the range of parameters is estimated by the sample statistics, i.e., the overall situation of the population is inferred through the investigation of the research subjects in the sample. In the actual investigation work, there is no need to carry out the census if it is not for the purpose of early detection and early treatment of patients but only to describe the distribution of disease.

The basic requirement of the sampling survey is that the results obtained from the sample can be extrapolated to the entire population. For this reason, the sampling must be randomized, and the sample size must be sufficient (representative).

Strengths and Limits

Compared with the census, the sampling survey has the advantages of saving time, manpower, and material resources. At the same time, due to the small scope of the investigation, it is easy to do it in detail. However, the design, organization, and implementation of the sampling survey and data analysis are complex. Duplications and omissions are not easy to find, so they are not suitable for populations with large variations. Sampling is also not appropriate for diseases with low prevalence because small samples could not provide the information required. In addition, if the sample size is large enough up to 75% of the population, a census may be a better choice.

3.3.4 Design and Implementation

An excellent design scheme is the premise of successful implementation in a research project. It is necessary to pay special attention to the representativeness of the selected subjects in the sampling survey, which is the prerequisite for the overall inference of the research results. Random sampling and avoiding selection bias are important conditions to ensure the representativeness of the research objects.

3.3.4.1 Clarifying the Purpose and Type of Investigation

According to the research problems expected to be solved, the purpose of the survey should be confirmed, such as to know the distribution characteristics of diseases and the exposure of risk factors or to carry out group health examination. Then, census or sampling survey can be determined to choose based on the specific research purpose according to the specific research purpose.

3.3.4.2 Identifying Subjects

According to the research objective and the distribution characteristics, geographical scope and time point of investigation, and the feasibility of carrying out the survey in the target population, the research object was determined. In the design, the research object can be defined as all or part of the residents in a certain area. It can also be composed of the floating population at a certain point. It can also be selected as the research object for some special groups. For example, the professional workers exposed to a certain chemical substance can be collected to study skin cancer.

3.3.4.3 Determining Sample Size

The sample size is the minimum number of observations required to ensure the reliability of the research results. The factors determining the sample size of the present study come from many aspects, but the main influencing factors include the following aspects:

1. Expected incidence (P) or standard deviation (S). The largest sample size is required when the current incidence rate is 50%.
2. The accuracy of the results of the survey requirements, i.e., the greater the allowable error (d), the smaller the sample size required.

3. Significance level (α) or the probability of the type I error. The smaller the test level, the more samples are required. For the same test level, the sample content required by the bilateral test is larger than that required by the unilateral test, which is usually taken as 0.05 or 0.01.

Statistical variables are generally divided into two categories: numerical variables and categorical variables. Therefore, the formula used to estimate the corresponding sample size is also different.

Estimating Sample Size for Numerical Variables

The following formula is used to estimate the sample size for random sampling:

$$n = \left(\frac{Z_\alpha S}{d}\right)^2 \tag{3.1}$$

In the formula, n is the sample size, d is the allowable error, i.e., the difference between the sample mean and the overall mean, which is determined by the survey designer according to the actual situation. S is the standard deviation, Z_α is the normal critical value at the test level and α is usually taken as 0.05 and $Z_\alpha = 1.96$.

Estimating of Sample Size for Categorical Variables

The following formula is used to estimate the sample

$$n = \frac{t^2PQ}{d^2} \tag{3.2}$$

In the formula, n is the sample size, P is the estimated overall prevalence, $Q = 1 - P$, d is the allowable error, t is the statistic of hypothesis testing.

Assuming that d is a fraction of P, when the permissible error $d = 0.1P$, $\alpha = 0.05$, $t = 1.96 \approx 2$, then formula (3.2) can be written as:

$$n = 400 \times Q/P \tag{3.3}$$

The above sample size estimation formula only applies to the data of Binomial distribution, i.e., $np > 5$, $n(1 - p) > 5$. Otherwise, it is advisable to estimate the sample size by Poisson distribution. The expected value of Poisson distribution and the confidence interval table can be used to estimate the sample size.

The sample size calculation method introduced above is only applicable to simple random sampling. However, in field epidemiologic investigation, cluster sampling is more commonly adopted because it is easy to organize and implement. The sampling error of cluster sampling is large. If the sample size of simple random sampling is

Table 3.1 The confidence interval of expected value for the Poisson distribution

	0.95			0.90	
$1 - 2\alpha$	Lower	Upper		Lower	Upper
0	0.00	3.69		0.00	3.00
1	0.03	5.57		0.05	4.74
2	0.24	7.22		0.36	6.30
3	0.62	8.77		0.82	7.75
4	1.09	10.24		1.37	9.15
5	1.62	11.67		1.97	10.51
6	2.20	13.06		2.61	11.84
7	2.81	14.42		3.29	13.15
8	3.45	15.76		3.93	14.43
9	4.12	17.08		4.70	15.71
10	4.30	18.29		5.43	16.96
11	5.49	19.68		6.17	18.21
12	6.20	20.96		6.92	19.44
13	6.92	22.23		7.69	20.67
14	7.65	23.49		8.46	21.89
15	8.40	24.74		9.25	23.10
20	12.22	30.89		13.25	29.06
25	16.18	36.90		17.38	34.92
30	20.24	42.83		21.59	40.69
35	24.38	48.68		25.87	46.40
40	28.58	54.47		30.20	54.07
45	32.82	60.21		34.56	57.69
50	37.11	65.92		38.96	63.29

calculated to estimate the sample size of cluster sampling, the sample size will be smaller. Therefore, it is advocated to multiply the sample size for simple random sampling by 1.5 as the sample size for cluster sampling.

Example The estimated incidence of colorectal cancer in a city is 30/100,000. How many people should be sampled?

If you take a random sample of 10,000 people, according to the incidence, 30/100,000, the expected number of survey cases is 3. As Table 3.1 shows, when the expected number of cases (1 − 2α) is 2, the 95% confidence interval is 0.24–7.22, which means there may be no case. If there was at least 1 case with colorectal cancer, the lower limit of 95% confidence interval is 1.09, and the expected number of cases will be 4. In order to reach at least 4 cases of colorectal cancer patients in the survey results, $4/X = 30/100,000$, so $X = 13,334$. Finally, 13,334 people should be investigated at least.

In actual work, the sample size can be appropriately increased to avoid errors between the estimated and actual incidence rate.

3.3.4.4 Determining the Sampling Method

There is non-random sampling and random sampling. The former includes a typical investigation. A random sampling must follow the randomization principle, which means ensuring that everyone in the population has a known, non-zero probability of being selected as the research object to ensure representativeness. If the sample size is large enough, the data are reliable, and the analysis is correct, the results can then be extrapolated to the population.

The commonly used random sampling methods are simple random sampling, systematic sampling, stratified sampling, cluster sampling, and multi-stage sampling.

Simple Random Sampling

Simple random sampling is the simplest and most basic sampling method. The important principle is that each subject is selected with the equal probability (n/N). The specific method is as follows: first, all observation objects are numbered to form a sampling frame. Then, some observation objects are randomly selected from the sampling frame by drawing lots or using random number table to form samples.

Simple random sampling is the most basic sampling method and the basis of other sampling methods. However, when the overall number of the survey is large, it is difficult to number each individual in the population. Moreover, the sample is scattered, which is not easy to organize and implement. Therefore, it is rarely used in epidemiological studies.

Systematic Sampling

Systematic sampling, also known as mechanical sampling, is to number individuals in a population in order and then randomly select a number as the first survey individual, while the others are selected according to some rules. The most commonly used systematic sampling is isometric sampling, in which all units within the population are sorted and numbered. According to the sample size, the corresponding individual samples are mechanically selected in specific sampling space. The sample numbers taken are:

$$i, i + k, i + 2k, i + 3k, \ \ldots, i + (n - 1)k \tag{3.4}$$

k is sampling space; n is sample size; i is the randomly selected starting number.

Example If there are 250,000 observation objects in a population, 1000 objects are to be selected for investigation. Sampling can be carried out through systematic sampling. Firstly, the sampling interval is $k = 250{,}000/1000 = 250$, and then one number was randomly selected from the first unit by the simple random sampling method as the starting point. If *i* is 25, the individual numbers were selected successively: 25, 275, 525, 775, 1025, etc.

Compared with simple random sampling, systematic sampling saves time, and the sample distribution is more uniform and representative. However, the disadvantage of systematic sampling is that when the observed individuals in the population have a periodic increasing or decreasing trend, it would produce bias, and the representativeness of the obtained samples will be declined, e.g., the seasonality of diseases, periodic changes of investigation factors.

Stratified Sampling

Stratified sampling refers to dividing the population into several subpopulations according to certain characteristics, and then conducting simple random sampling from each subpopulation to form a sample. The smaller the intra-group variation and the greater the inter-group variation, the better. Stratified sampling is more accurate than simple random sampling. Moreover, it is more convenient for organization and management.

Stratified sampling is divided into two categories: one is called proportional allocation stratified random sampling, i.e., the sampling proportion within each subpopulation is equal. The other is called optimum allocation stratified random sampling, i.e., the sampling proportion within each subpopulation is unequal. The sampling proportion with small inter-group variation is small, and with large inter-group variation is large.

Cluster Sampling

Cluster sampling refers to dividing the population into several groups and selecting some groups as observation samples. If all the selected groups are all the respondents, it will be a simple cluster sampling. If some individuals are investigated after sampling again, it is called two-stage sampling. The characteristics of cluster sampling are as follows:

① It is easy to organize, convenient to try, and implement;
② The smaller the difference between groups, the more groups are extracted, and the higher the accuracy will be;
③ The sampling error is large, so it is usually increased by 1/2 on the basis of the simple random sample size estimation.

The above-mentioned four basic sampling methods are introduced. When the sampling method is fixed, the order of sampling error is from large to small: cluster sampling, simple random sampling, systematic sampling, and stratified sampling.

Multi-Stage Sampling

The sampling process is carried out in multi-stages, with each stage using a different sampling method. Combined with the above sampling methods, multi-stage sampling is commonly used in large epidemiological studies. For example, the InterASIA Study (International Collaborative Study of Cardiovascular Disease in Asia) has adopted the following multi-stage sampling method:

The first stage: the sampling unit is the province and city. Four economically and geographically representative cities were drawn from the south and north, respectively. Beijing and Shanghai were included in the northern and southern samples, respectively, and a total of 10 provinces and municipalities were drawn. It should be noted that in order to fully consider the geographical and economic level of representation, random sampling is not used at this stage, but the random sampling method is used in the next three stages.

The second stage: the sampling units are counties and urban areas. A county and an urban area were randomly selected from the provinces and cities in the first stage, and ten counties and ten urban districts were drawn.

The third stage: the sampling unit is a street, town, or township. Street or town (or township) is randomly selected from each urban area and county, and a total of ten streets and ten towns (or townships) are drawn.

The fourth stage: the sampling unit is an individual. The list of residents of all streets or towns will be used as a sample source (limited to 35–74 years old), and each site will have 400 male and female residents.

The above sampling methods used in four stages are called multi-stage sampling.

Multi-stage sampling can make full use of the advantages of various sampling methods and overcome their shortcomings. The disadvantage of multi-stage sampling is that the demographic data and characteristics of each sub-group should be collected before sampling. Also, the statistical analysis of the data is complicated, such as the sampling weight of the complex sampling design when calculating the sampling error.

3.3.4.5 Data Collection, Collation, and Analysis

In a cross-sectional study, the method of data collection cannot be changed once it is determined. Consistency must be maintained throughout the study to avoid heterogeneity of data. The data collection process should pay attention to the unification definition of exposure and the criteria of disease. All personnel involved in the inspection or testing must be trained to avoid measurement bias with unified investigation and testing standards.

Determining the Data to be Collected

The most basic principle of the cross-sectional study is whether the subject has a certain disease or characteristics, and the investigator uses grading or quantitative methods as much as possible. In addition, other information, such as social and environmental factors, need to be collected to illustrate the distribution and related factors. The relevant information collected generally includes the following:

① Basic information about the individual, including age, gender, ethnicity, education level, marital status, per capita monthly income, etc.
② Occupational and exposure status, including nature, type, position, and working years.
③ Lifestyle and health conditions, including diet, smoking history, drinking history, physical exercise, depression, anxiety, medical history, disease history, etc.
④ Women's reproductive status, including menstrual and obstetrical histories, use of contraceptives, and hormones.
⑤ Environmental information, expressed in objective and quantitative indicators.
⑥ Prevalence, infection rate, etc.

Investigator Training

Before the investigation, the investigators should be trained uniformly following a standard protocol. The consistency of the methods and standards for collecting data can be guaranteed. Investigator training is an important part to ensure the accuracy of data.

Data Collection Methods

In a cross-sectional study, there are three methods for collecting data. The first one is by laboratory measurement or examination, such as blood glucose detection, blood lipid detection, etc. The second way is to investigate the subject through the use of a questionnaire to obtain information on exposure or disease. The third way is to use routine data. For example, get data from the Center for disease control (CDC) and electronic disease records.

Data Collation and Analysis Methods

Data collation refers to checking the integrity and accuracy of the original data carefully, filling in the missing items, deleting the duplicates, and correcting the errors. Disease or a state of health is verified and classified according to clearly defined criteria. Then it can be described according to different spaces, time, and the distribution in the crowd.

In data analysis, the population can be further divided into exposed and non-exposed groups or different levels of exposure population. The differences in disease rate or health status between the groups can be compared and analyzed. The subjects can also be divided into disease and non-disease groups to evaluate the relationship between factors (exposure) and disease.

① Description of the demographic characteristics. A detailed description of the demographic characteristics (e.g., gender, age, education level, occupation, marital status, and socioeconomic status) can help to easily understand the basic characteristics of the research object and can be used to compare with other studies.

② Analysis of the distribution characteristics of the disease: According to the characteristics of the different subjects (gender, age, education level, occupation, marital status, socioeconomic status, etc.), regional characteristics (urban, urban, north, south, mountain, plain, or administrative division, etc.) and time characteristics (season, month, year, etc.) are grouped, the prevalence of a disease or the mean and sampling error of a certain variable are calculated and compared and the correct statistical method is used to test the differences between the different groups.

③ Analysis of the relationship between exposure factors and disease: Compare the prevalence of a disease or the mean value of a numerical variable according to the presence or absence of exposure factors or the level of exposure. It is also possible to calculate an odds ratio (OR) to estimate the association and association strength in an epidemiological method (such as a case-control study). Not only univariate analysis but also multivariable adjustments to calculate the ORs are required. What needs to be emphasized here is that cross-sectional study can only provide preliminary clues to the cause.

3.3.5 Bias and Control

Bias is the systematic errors generated in the process from design, implementation, to analysis, as well as the one-sidedness in the interpretation or inference of the results, which leads to the tendency of a difference between the research results and the true value, thus mischaracterizing the relationship between exposure and disease. The common bias in cross-sectional studies includes selection bias, information bias, and confounding bias.

3.3.5.1 Selection Bias

Selection bias is the systematic error caused by the differences of the characteristics between the included subjects and those who were not included in the study. It mainly includes the following aspects:

1. Selective bias: In object selection process, due to not strictly sampling, the objects are selected subjectively, which results in the deviation of the research samples from the population. For example, when you want to know the prevalence of hepatitis B in one city last year, if the sample were only information collected from the hepatitis specialized hospital, the prevalence must be higher than the actual rate in the general population.
2. Non-response bias: During the investigation, the subjects did not cooperate or were unable or unwilling to participate for various reasons, resulting in a missed investigation. If the response rate is too low (less than 80% or even 85%), it could produce a non-response bias, and it is more difficult to apply the results to estimate the source population.
3. Survivor bias: In cross-sectional study, survivors of disease are often selected as subjects. Current cases and deaths may have different characteristics, which could not summarize the overall situation. Therefore, the results have some limitations and one-sidedness.

3.3.5.2 Information Bias

Information bias is a systematic error that occurs when information is obtained from a research subject during the investigation process. Information bias can come from subjects, investigators, measuring instruments, equipment, and methods.

1. Respondent bias includes recall bias and reporting bias: The subjects were biased by unclear or completely forgotten disease history, drug application history, and risk exposure history.
2. Investigator bias: The bias occurs in the process of collecting, recording, and explaining information from respondents. One reason is, different investigators have different results for the same subject, the other is the same investigator has different results for several surveys of the same subject.
3. Measurement bias is a systematic error caused by inaccurate instrument and incorrect operation procedure. For example, if the sphygmomanometer is not calibrated, all measurement results are higher or lower than true value. The methods of investigation used are not uniform, and bias may occur.

3.3.5.3 Confounding Bias

Confounding bias is caused by confounding factors. If the association between exposure and disease is analyzed, then there will be confounding bias.

Bias can be avoided or reduced, so it is necessary to carry out quality control in the research to minimize the occurrence of bias.

1. In the sampling process, keep the randomization principle strictly to ensure the representativeness of the sample.
2. To improve the compliance and test rate of the respondents, each subject should be investigated.
3. To correct the measuring instruments, equipment, and testing methods, including the preparation of questionnaires.
4. To train the investigators, unify survey standards and conduct mutual supervision and spot checks.
5. After the investigation, reviewing and checking the information is needed.
6. In the process of data collation, the correct statistical analysis method should be selected, pay attention and identify confounding and influencing factors.

3.3.6 Strengths and Limitations

3.3.6.1 Strengths

1. The implementation time is short, and the results can be obtained quickly. Thus, the research task can be completed in a short time.
2. Compared with other research types, a cross-sectional study is a relatively inexpensive method.
3. It could investigate the association between disease and factors and establish a preliminary etiological hypothesis.
4. A cross-sectional study can provide a basis for making disease prevention and control plans.

3.3.6.2 Limitations

1. Prevalence, instead of incidence, can only be obtained from a cross-sectional study.
2. Low-prevalence disease and its influencing factors are not suitable for cross-sectional study.
3. The time sequence between exposure and disease cannot be determined, so there is no causal association, and only preliminary etiological clues can be provided.

3.3.7 Cases

Due to the rapid development of the Chinese economy, the diet and lifestyle have changed greatly. Diabetes, hypertension, hyperlipidemia, and many other diseases related to diet and lifestyle increased significantly. Li Liming et al. conducted a survey on the situation among Chinese people in 2002.

3.3.7.1 Purpose and Type of Study

The purpose of this study is to investigate the nutrition and health status of Chinese residents and to analyze the main factors affecting the nutrition and health status. Therefore, the cross-sectional study was adopted. Compared with the census, the sampling survey saves more time and cost. Therefore, multi-stage stratified cluster sampling was selected.

3.3.7.2 Subjects and Sample Size

The target population was the national resident population. With 95% accuracy and 90% precision, the minimum sample size was estimated at 225,000. In addition, assuming no response rate of 10%, the final sample size was 250,000.

3.3.7.3 Research Content and Data Collection, Collation, and Analysis

Data collection included questionnaires, medical examinations, laboratory tests, and dietary surveys. Firstly, demographic characteristics, socioeconomic status, disease history, smoking, drinking, and physical activity of the individuals were collected through questionnaires. Secondly, the height, weight, waist circumference, and blood pressure of the individuals were tested. Thirdly, laboratory tests were performed on the serum indexes, including hemoglobin, TC, TG, HDL-C, and LDL-C. Fourthly, the 24-hour retrospective method, food frequency method, and weighing method were used to carry out the dietary survey.

Finally, 243,206 people were enrolled in this study. After adjusting for age and region, the national adjustment rate was calculated by direct standardization method.

The results showed that the consumption of cereals was the maximum, and the dietary structure showed a significant regional difference. The consumption of meat, fruit, and vegetable oil in urban areas is higher than that in rural areas. While the consumption of cereals, tubers, and vegetables in rural areas was higher than that in urban areas. The overweight rate in China is 17.6%, and the obesity rate is 5.6%. Both overweight and obesity rates increase with age. Obesity rates are higher in cities than in rural areas among people over the age of 7. The prevalence of anemia is 15.2%, which is significantly higher in young and middle-aged women than in men. The prevalence of diabetes among Chinese adults is 2.6%, which increases with age. The prevalence in cities rises faster than in rural population.

3.3.7.4 Conclusion

The nutrition and health status of Chinese population are gradually changing. The prevalence of anemia reflects the lack of trace elements like iron in Chinese population. The prevalence of chronic diseases such as overweight, obesity, and diabetes is increasing rapidly, which has been a threat affecting the health of the Chinese people. In addition, disease prevention should be targeted because of the differences in nutrition and health levels between urban and rural populations.

3.4 Ecological Study

3.4.1 Concept

Ecological study is also called correlational study. It is used to analyze the relationship between exposures and diseases by describing the exposure and the frequency of diseases in different populations.

3.4.2 Type of Study Design

3.4.2.1 Ecological Comparison Study

Ecological comparison study is often used to compare the relationship between the exposure and the disease frequency in different population groups to provide clues for the disease cause. For example, the National Cancer Research Center of the United States has drawn a statistically significant regional difference in the age-adjusted mortality map of oral cancer between 1950 and 1969. The highest morality is in the urban areas which are dominated by men in the northeast and women in the southeast. Smoking may be a risk factor for oral cancer from this distribution, as smoking in the South is common. Later case-control studies also supported this cause hypothesis. Immigration epidemiological research method can also be applied in ecological studies. It is usually used to analyze the relationship between genetic factors or environmental factors and diseases by comparing the incidence of immigrants and their children with the incidence of residents of the original place of residence and residents of the settlement in different areas.

3.4.2.2 Ecological Trend Study

The ecological trend study is to continuously observe the changes in exposure levels and diseases in the population and describe their trends. The relationship between factors and disease is judged by comparing changes in disease before and after exposure changes.

In the implementation of ecological studies, the above two types are often combined. The suspicious etiology of the disease is explored by studying the frequency of occurrence of a disease in multiple regions (multiple groups of comparisons) and at different times (time trends). For example, some researchers analyzed the relationship between water hardness and cardiovascular mortality for gender and age in 63 towns in the UK from 1948 to 1964. It was found that cardiovascular mortality and water hardness were negatively correlated in all genders and ages, especially in men. In urban areas with high water hardness, the increase in cardiovascular mortality was less than in towns with low water hardness.

3.4.3 The Main Application

1. Etiological assumptions related to the disease distribution can be found through ecological studies. Ecological studies have found that colorectal cancer was more common in developed countries than in developing countries, considering that dietary habits or environmental pollution might be related to the incidence of colorectal cancer.
2. To provide positive or negative evidence for some existing disease causal hypotheses.
3. It can be used to evaluate the effects of intervention experiments or field experiments. For example, promote low sodium intake in the population and then compare the changes in the average sodium intake level before and after the promotion of low sodium salt and the trend of the average blood pressure of the population to evaluate the effect of low sodium salt intervention.
4. To estimate trends in disease changes. Applying ecological trend studies in disease surveillance to estimate trends in a disease can help prevent and control disease. Between 1959 and 1966, the number of deaths from asthma in England and Wales was associated with a simultaneous increase in sales of bronchodilators. After the cessation of bronchodilators in the pharmacy in 1968, the mortality rate of asthma decreased significantly. Therefore, the development of a ban on bronchodilators without prescription was the result of ecological research.

3.4.4 Bias

Disease information in ecological studies is often derived from historical records (cancer registers, medical records, etc.), while exposure information is often derived from government agency data (tobacco and alcohol sales, etc.). The accuracy of these data directly affects the reliability of research results.

3.4.5 Strengths and Limitations

3.4.5.1 Strengths

1. Ecological study can be carried out using existing routine data to save time and money and then quickly yield results. It takes a long time to measure the relationship between a biological indicator and a disease through a prospective study. The preliminary study using an ecological method can narrow the research risk.
2. For unknown disease etiology, etiological clues for further research can be provided by an ecological study. This is the most striking feature of ecological study.
3. In the field of environment or other research, an ecological study is the only alternative research method when the cumulative exposure of an individual is not easy to measure. For example, in the study of the relationship between urban air pollution and lung cancer, it is difficult to estimate the amount of polluted air inhaled by individuals accurately. At this time, multiple methods of ecological comparison can be applied for research.
4. When the range of individual exposure in a population is not large enough, it is difficult to estimate the relationship between exposure and disease. In this case, ecological comparison studies with multiple populations are more appropriate. For example, not only are high-fat diet habits similar, but also the intake is generally high in Western countries. If the relationship between individual fat intake and coronary heart disease were studied only in Western countries, it would be difficult to find a relationship. If a comparative study of the low-fat diet of the Eastern countries was chosen, meaningful results might be found.
5. Ecological studies are appropriate for evaluating population intervention measures. For example, folic acid deficiency in humans can lead to fetal neural tube defect, which was first hypothesized through ecological studies. The addition of folic acid in the pregnant population led to a significant decrease in the incidence of fetal neural tube defects.

3.4.5.2 Limitations

1. Ecological fallacy: Etiological clues suggested by ecological studies may be either a true or a false association between disease and exposure. The ecological fallacy is a misinterpretation of the association between exposure and outcomes due to an inaccurate assessment.
2. Confounding factors are often difficult to control, especially socio-demographic and environmental variables. Multicollinearity may affect the correct analysis of the relationship between disease and exposure.
3. Because the timing sequence between exposure and disease is not easy to determine, it is difficult to determine the causal relationship between the two variables.

When conducting an ecological study, do not set too many research questions in a study. The differences between population groups should be minimized. The interpretation of the results should be compared with other non-ecological results as far as possible and combined with professional knowledge for comprehensive analysis and judgment.

3.4.6 Case

In order to analyze and evaluate the relationship between life expectancy and fine particulate pollutants in the air, a study from the United States compiled the life expectancy, socioeconomic status, and society of 211 counties in 51 urban areas in 1980 and 2000. Demographic characteristics and concentration of airborne fine particle contaminants were analyzed using regression models to analyze the relationship between airborne fine particle concentration and life expectancy, after adjusting socioeconomic status and demographic variables, as well as the prevalence of smoking. The results of the study showed that a 10 ug/m^3 reduction in the concentration of fine particulate contaminants was associated with an increase in life expectancy of 0.61 ± 0.20 years ($P = 0.004$). And after adjusting the multivariate in the model, the results still remained significant. The results suggested that reduced fine particulate contaminants in the air can increase life expectancy by 15%.