Chapter 8
Bias

Lu Long

Key Points

- Bias refers to various influencing factors in epidemiological research, including design, implementation, analysis, and inference, also known as systematic error. Three major threats to validity are selection bias, information bias, and confounding bias.
- Selection bias occurs when the characteristics of the subjects are different from the source population, which leads to the deviation of the research results from the real situation.
- Information bias, known as observational bias, refers to the inaccuracy or incompleteness of the exposure or outcome information obtained during the implementation of the research, which results in the misclassification of the exposure or disease of the research subjects and affects the validity of the results.
- Confounding bias is due to the existence of one or more external factors that mask or exaggerate the link between research factors and diseases, thus partially or wholly distorting the actual association.

8.1 Introduction of Bias

Bias, also known as systematic error, refers to various influencing factors in epidemiological research, including design, conduct, analysis, and inference.

The existence of these influencing factors, including design errors, data acquisition distortion, incorrect analysis, or not logical inference, leading to the association between exposure and outcome is misestimated, and this actual relationship is systematically distorted, which leads to the wrong conclusion. Bias is an important issue that affects the authenticity of the results. Thus, we must fully understand the

L. Long (✉)
West China School of Public Health, Sichuan University, Chengdu, China

C. Wang, F. Liu (eds.), *Textbook of Clinical Epidemiology*,
https://doi.org/10.1007/978-981-99-3622-9_8

source of bias and its causes and minimize the occurrence of bias in our studies to ensure the authenticity of the study. There are two directions of bias, i.e., positive bias and negative bias. Positive bias means that the measured value of the study overestimates the true value, on the contrary, it is negative bias.

We generally classify bias as selection bias, information bias, and confounding bias.

8.2 Selection Bias

8.2.1 Definition

Selection bias means that the characteristics of the selected subjects are different from those of the unselected, which leads to a deviation of the research results from the real situation.

8.2.2 Classification

8.2.2.1 Self-Selection Bias

Self-selection bias, or volunteer bias, is one source of selection bias. Self-selection bias is one type of bias that results from individuals disproportionately selecting themselves to join a group. For example, researchers selected soldiers from the Smoky Atomic Test in Nevada to investigate the leukemia incidence. In this study, 76% of the soldiers are members of the cohort with known outcomes, and the remaining 24% were identified as the cohort without known outcomes. Those who knew the outcomes, 82% were traced by the investigators, while others reached out to surveyors. We ordinarily consider self-reported subjects as a threat to validity because self-reporting may be related to the study results.

In the Smoky Atomic Test study, among the 62% (2% × 76%) of cohort members, investigators traced four target cases and the 14% (18% × 76%) of cohort members also traced four target cases who reported themselves. We assume the leukemia incidence without known outcomes (24%) is similar to that of the subjects traced by the investigators. In that case, we should expect that only (24%/62%) × 4 = 1.5, meaning that about one or two cases occurred among this 24% of the members without known outcomes, only a total of nine or ten cases in the entire cohort. If instead, we suppose that the 24% without known outcomes had the same incidence of leukemia as subjects with known outcomes. We would calculate that 8(24%/76%) = 2.5, meaning that about two or three cases occurred among this 24%, in the entire cohort we will observe 10 or 11 cases. However, among the 24% + 14% of the cohort, all cases were untraced among the self-reported, leaving no case among those without known outcome. T. The total number of cases will be only

8 in the entire cohort. This example indicates that self-selection bias is a small but a real problem in research.

8.2.2.2 Berksonian Bias

Berkson's bias or Berksonian bias is also known as admission rate bias. It usually occurs in a hospital-based case-control study because the selected case or controls represent only a subset of patients with a disease rather than an unbiased sample of the corresponding target population. Affected by medical conditions, residence, socio-economy, education, and other factors, patients have specific selectivity to hospitals, and hospitals also have a specific selectivity to patients, which results in problems in sample representativeness and bias in a hospital-based case-control study.

For example, hospital-based case-control study was used to explore the relationship between birth control pills and thrombophlebitis. Cases were recruited from people with thrombophlebitis in a hospital. And randomly selected as patients without thrombophlebitis in a certain ward of the same hospital as a control group. Suppose there are 5000 patients with thrombophlebitis and 5000 patients without thrombophlebitis. Oral contraceptive accounts for 15% in each of them. It is assumed that admission rates for these three conditions are relatively independent (Table 8.1).

It can be calculated from Table 8.1 that the correlation of thrombophlebitis and oral contraceptive, OR = (750 × 4250)/(4250 × 750) = 1.0, which indicates that there is no correlation among oral contraceptive and thrombophlebitis.

Now assume the admission rate of case group was 25% while control group was 60%, and the admission rate of oral contraceptive was 40%. The composition of the comparative study samples is shown in Table 8.2.

The admission rate of the 750 patients with thrombophlebitis and exposure to contraceptive was 25%, So the number of thrombophlebitis hospitalizations were 750 × 25% = 187.5 ≈ 188; and 40% of the remaining were hospitalized due to exposure to contraceptive, and the number of hospitalized patients was (750–750 × 25%) × 40% = 225, and the total hospitalizations was 413.

Table 8.1 Exposure and disease distribution in the total population

	Oral contraceptive		
Group	Yes	No	Total
Case	750	4250	5000
Control	750	4250	5000

Table 8.2 Distribution of exposure and disease in the hospital

	Oral contraceptive		
Group	Yes	No	Total
Case	413	1063	1476
Control	570	2550	3120

The admission rate of the 4250 patients with thrombophlebitis rather than exposure to contraceptive was 25%, So the number of cases group was $4250 \times 25\% = 1062.5 \approx 1063$.

The admission rate of the 750 patients without thrombophlebitis who were exposed to contraceptives was 60%, so the hospitalizations were $750 \times 60\% = 450$, and 40% of the remaining patients were hospitalized because of exposure to contraceptives, the hospitalizations was $(750–750 \times 60\%) \times 40\% = 120$, with a total hospitalization of 570.

The admission rate of the 4250 patients without thrombophlebitis and the oral contraceptives was 60%, So the number of total hospitalizations was $4250 \times 60\% = 2550$.

According to the above data, OR = $(2250 \times 413)/(570 \times 1063) = 1.53$, oral contraceptive was positively correlated with thrombophlebitis.

There was no association between oral contraceptives and thrombophlebitis in the total population, but a case-control study using hospital samples found a positive correlation. The degree of the association was influenced by the admission rate, which deviated from the true association in the population. This is Berksonian Bias.

8.2.2.3 Detection Signal Bias

Detection signal bias, known as unmasking bias, is also a common selection bias. If the exposure factor to be studied has no turel causal relationship to the disease, however, its presence may cause the subject to develop symptoms or sighs related to the disease to be studied, leading to earlier or more frequent visits to the doctor, which increases the detection rate of the disease and makes it more likely to be included as a case in the study. Suppose these patients are taken as case groups in case-control studies. In those cases, there will be systematic differences in certain characteristics (such as exposure factors) between admitted patients and non-admitted patients, leading to misestimating the true associations between exposure factors and outcomes. For example, several studies found that oral estrogen was associated with endometrial cancer and believed that oral estrogen was a risk factor for endometrial cancer. However, many scholars later proposed that estrogens do not cause cancer to occur, but only allow cancer to be diagnosed. Because estrogen can stimulate the growth of the endometrium, making the uterus prone to bleeding. The women who take estrogen are more likely to seek medical attention, this made early-stage endometrial cancer patients easier to be identified. In contrast, while case-control studies with such patients as case group led to an increased proportion of oral androgens in endometrial cancer patients, thereby overestimating the association between estrogen and endometrial cancer.

8.2.2.4 Neyman Bias

Neyman bias, called prevalence-incidence bias, was first described by Neyman in 1955 and occurred in the case-control study design. When carrying out case-control studies, we can select three cases: cases-incident cases, prevalent cases, and death cases. If all the admitted cases are survived cases, especially the cases with a long disease course, may be related to the survival, but not to the onset of the disease. It may, thus misestimate the etiological effect of these factors. On the other hand, survivors of disease may change some of their existing exposure. When they are investigated, they may mistake these changed exposure characteristics as their disease conditions, resulting in errors in the correlation between these factors and the disease.

8.2.2.5 Loss of Follow-Up

Cohort studies, clinical trials, and clinical prognosis studies generally require follow-up of subjects. For the long observation period, the follow-up process cannot avoid the absence of outcome events due to relocation of subjects, death due to other reasons (competitive risk), or withdrawal from the study due to poor treatment effects, adverse reactions, and other reasons. Loss of follow-up will affect the representativeness of the research objects, thus affecting the authenticity of the results. Therefore, this bias is called loss of follow-up bias.

8.2.3 Control

It is difficult to eliminate or correct its effects on the results once select bias occurs. Therefore, scientific research design should be performed to reduce and avoid such bias.

8.2.3.1 Scientific Research Design

In the research process, we(researchers) should clear the global and the sample population and predict the various bias that may be generated in the sample selection process based on the nature of the study. In the case-control study, we should avoid selecting cases in a single hospital, and we can set up community control and hospital control at the same time. Even if the cases can only be selected from the hospital, they should also be randomly sampled in the different areas and different levels of hospitals. In the cohort study, we can establish various controls, including comparing incidence in exposed populations and all populations or compare incidence in exposed populations and other unexposed populations, to reduce the effects of selective bias.

8.2.3.2 Develop Strict Inclusion and Exclusion Standards

In both observational research and experimental research, we must have developed a strict, clear unified standard about inclusion and exclusion, including disease diagnostic criteria and exposure criteria, enabling the selected research object to better represent the overall. After the exclusion standard determines the selection, it strictly complies with the study's implementation phase and cannot be changed casually.

8.2.3.3 Maximize Response Rates

Various measures should be taken to obtain the cooperation of the subjects as far as possible, improve the response rate, reduce or prevent the occurrence of loss of follow-up, and control the selection bias. During the study, we should increase the subjects' understanding of the significance of the study through various ways. When the non-response rate or loss of follow-up rate is more than 10%, we should be cautious in analyzing the research results. A random sampling survey should be conducted on the non-responders or lost respondents if possible, and the results of the sampling survey should be compared with those responders. If there is no significant difference, it shows that the non-response or loss of follow-up has little effect on the results; oppositely, we should explain appropriately. Strategies to reduce loss to follow-up include: screening of willingness prior to registration, detailed collection of participants' contact information, using effective incentives, and maintaining regular contact with participants. In addition, the sample size can be appropriately increased to reduce the impact of the loss of follow-up or non-response on the results after the corresponding sample size is calculated in the design stage.

8.2.3.4 Randomization Principle

Randomization can be divided into two different forms of random sampling and random allocation. Random sampling means the opportunity of each target object extracted into the study queue is equal, making the research sample representative, avoiding bias due to the subjective, arbitrary choice of research objects; random allocation is the equivalent opportunity for participants to be assigned to the experimental group or control group without the effect of researchers and participants' subjective wishes or unconscious objective reasons. The purpose of random distribution is to make the non-research factors evenly distributed in each group and to increase the transferability among groups.

8.3 Information Bias

8.3.1 Definition

Information bias, known as observational bias, refers to the inaccuracy or incompleteness of the exposure or disease information obtained during the implementation of the research, which results in the wrong classification of the exposure or outcome of the research subjects and affects the authenticity of the results. Information bias generally occurs when there are errors in the measurement, which is also known as classification error or misclassification for discrete variables. Misclassification can divide into differential misclassification and nondifferential misclassification. Compared with nondifferential misclassification, differential misclassification has a greater impact on study results. Due to the directions of differences in the misclassification among groups, the effect value may be overestimated or underestimated.

8.3.2 Classification

8.3.2.1 Differential Misclassification

Differential misclassification refers to classification errors that rely on the actual values of other variables. The most common differential misclassification is recall bias. Suppose an interview of congenital malformations in a case-control study, we generally obtain the etiological information from the mother. We selected mothers who have recently given birth to a deformed baby as a case, whereas mothers who had recently given birth to an apparently healthy baby as a control. The mothers of deformed infants are better able to recall exposures than mothers of healthy infants, leading to a kind of differential misclassification, referred to as recall bias. Because the birth of a deformed infant can stimulate the mother to recall all events that may have played some role in the unfortunate outcome. The difference produced by this recall bias is an apparent effect unrelated to any biological effect. Recall bias is likely to arise in any case-control study that requires recall of past experiences. Klemetti and Saxen [9] considered time as a critical indicator of recall accuracy.

When establishing or verifying a research hypothesis, if personal biased views are reflected in the process of data collection, it will lead to interviewer bias. The resulting inducement bias is also classified as interviewer bias if the researcher intentionally induces the subject to provide the required information. In cohort studies or experimental epidemiological studies, more detailed examination of exposure or intervention group may be performed if the investigator has previously assumed that the exposure or intervention is associated with the occurrence of outcome. It leads to a misjudgment of the study results.

Not all misclassification will exaggerate the association under study, but examples of the opposite can also be found. When investigating sensitive issues with the subjects, they will deliberately minimize the information. For example, patients with sexually transmitted diseases such as syphilis and gonorrhea may be reluctant to let investigators know about their history of exposure to unprotected sex because of stigma, and the resulting bias may underestimate the association between unprotected sex and sexually transmitted diseases.

8.3.2.2 Nondifferential Misclassification

Classification error that is independent of other variables is called nondifferential misclassification.

Presumably, bias due to independent nondifferential misclassification of exposure or disease is predictable in the direction, i.e., toward the null. Some researchers have used complex procedures to demonstrate that misclassification is nondifferential, Unfortunately, decomposing continuous or categorical data into fewer categories can transform non-differential errors into differential misclassifications even under blinding is accomplished or in cohort studies where disease outcomes have not yet emerged. Non-differentially alone does not guarantee a bias toward the null. Even if nondifferential misclassification is implemented, it may come at the cost of increasing the total bias.

Both disease and exposure can occur nondifference misclassification. When the proportion of subjects misclassified by disease does not depend on the subject's status with respect to other variables in the analysis, including exposure, it will occur non-differential disease misclassification. Similarly, when the proportion of subjects misclassified by exposure does not depend on subject status to other variables in the analysis, including disease, it will occur nondifferential exposure misclassification.

We will give an example to illustrate how an independent nondifferential disease misclassification with full specificity does not bias the risk ratio estimate but rather biases the absolute magnitude of the risk difference estimate downward by a factor, equal to the probability of false negatives. Suppose there is a cohort study in which 30 cases occur in 300 unexposed subjects and 60 cases occur among 200 exposed subjects. The actual risk ratio is 3, and the actual risk difference is 0.20. Assumes no false positives for disease detection, sensitivity is only 70% for both exposure groups. The expected numbers of exposure cases detected will be 0.70×60 and unexposed cases detected will be 0.70×30, which means that the expected risk ratio is estimated to be $((0.70 \times 60)/200)/((0.70 \times 30)/300) = 3$ and the expected risk difference is estimated to be $(0.70 \times 60)/200 - (0.70 \times 30)/300 = 0.14$. Thus, although disease misclassification did not bias the risk ratio but the expected risk difference estimate was 0.14/0.20 of the actual risk difference.

The effects of nondifferential misclassification of exposure are similar to the effect of nondifferential misclassification of disease. We hypothesized a cohort study comparing the incidence of liver cancer in smokers with the incidence among nonsmokers to explore nondifferential exposure misclassification. The incidence

rate was assumed to be 0.01% per year for nonsmokers, and 0.05% per year for smokers. We surppose2/3 of the study population are smokers, but only 50% admit this. This would then result in only 1/3 of subjects being identified as smokers with a disease incidence of 0.05% per year. And the remaining 2/3 of the population is made up of equal numbers of smokers and nonsmokers. Among those classified as nonsmokers, their average incidence would be 0.03% per year rather than 0.01% per year. The rate difference has been reduced by misclassification from 0.04% to 0. 02%, while the rate ratio has been reduced from 5 to 1.7.

These examples present how a nondifferential misclassification of a dichotomous exposure will produce a bias toward the null value (no relationship) if the misclassification is unrelated to other errors. The association will be completely obliterated and the direction of association will be reversed by bias, if the misclassification is severe enough (although the reversal will only occur if the classification method is worse than randomly classifying people as "exposed" or "unexposed").

We cannot dismiss a study simply because of the presence of substantial non-differential misclassification of exposure, it is incorrect. This is because the implications may be greater if there is no misclassification, which provides a probability of misclassification that applies uniformly to all subjects. Thus, the impact of nondifferential misclassification depends heavily on whether the study is considered positive or negative. Emphasizing measurement rather than qualitative descriptions of study results can reduce the likelihood of misinterpretation, but even so, it is important to keep in mind the direction and possible magnitude of bias.

8.3.3 Control

Whether differential misclassification or nondifferential misclassification is mainly due to problems in measurement or data collection methods, resulting in errors in acquired data. Therefore, we mainly adopt the following methods to control information bias.

8.3.3.1 Material Collection

The main purpose of the survey design is to standardize the tables in the study, which is crucial for internal validity, so that valid, reliable, and complete data could be collected efficiently. In addition, pretesting survey instrument in populations similar to the study population can identify flaws in the survey design and instruments before full data collection begins. We'd better use the blinding method to collect data to avoid the influence of subjective psychology of research objects and investigators on the survey results.

8.3.3.2 Objective Research Indicators

Try to use objective indicators or quantitative indicators to avoid information bias, such as applying laboratory examination results and consulting the medical records or health examination records of the subjects as the source of investigation information. Suppose it is necessary to collect data by means of inquiry. In that case, we should adopt closed questions and answers as far as possible to prevent the occurrence of report bias and measurer bias. For questionnaires concerning lifestyle and privacy, the respondents should be informed in advance that all responses are confidential and will be properly kept appropriately.

8.3.3.3 Investigation Skills

The investigative skills of investigators are particularly important when obtaining information, especially the research that requires the participation of investigators. We can improve their investigation level by training investigators and formulating investigators' manuals to reduce information bias.

8.4 Confounding Bias

8.4.1 Definition

Confounding bias is due to one or more external factors that mask or exaggerate the link between research factors and diseases, thus partially or entirely distorting the actual relationship between them. Confounding is produced by confounders (exposures, interventions, treatments, etc.).

Taking Stark and Mantel's study on neonatal Down's syndrome as an example. Population monitoring data indicated that Down's syndrome was associated with birth order. Assume the incidence of Down's syndrome in the first-born child was 0.06% while in the fifth-born child was 0.17%. The risk of Down's syndrome increased with the increase of birth sequence, which seemed birth order to be a risk factor for Down's syndrome. However, we should consider maternal age at delivery as a confounder, closely related to birth order and Down's syndrome risk. Further study found that the incidence of Down's syndrome in children delivered by pregnant women younger than 20 years old was 0.02%, and gradually increased with the age of delivery, and the incidence of Down's syndrome in children delivered by pregnant women over 40 years old was as high as 0.85%. The study indicates that the maternal age at childbirth is related to the occurrence of the disease. Therefore, it is suggested that the association of birth sequence with Down's syndrome risk may be influenced by the confounding factor of maternal age at birth.

In this part, we briefly refer to confounding bias, but we will discuss confounding and how to control it in the next part.

8.4.2 Confounding

When estimating the effect of an exposure on exposed individuals, Confounding can occur when the exposed and nonexposed subgroups of the population have different background disease risks. These subgroups can have different disease risks even if they are not exposed to any of the effects in both subpopulations. More generally, confounding occurs when the exposed and unexposed groups are not fully comparable or "exchangeable" in terms of exposure response, i.e., the exposed and unexposed groups may exhibit different risks even if both experience the same level of exposure. In general, a factor associated with both the exposure and the outcome could be a confounder. The following are three necessary but not sufficient conditions to be a confounder of the effect of the exposer.

First, confounder must be predictors of the disease without the exposure under study. Confounders are not necessarily the genuine cause of the disease under study. However, they are only "predictive" within the level of exposure apart from casual relations. For example, race, age, gender, etc., may be considered as potential confounders. Thus, one almost always sees adjustments made for age and sex.

Second, the confounder must be related to the study exposure. For example, confounder should be related to exposures in the control group in case-control study. If the factor is not associated with exposure in the control group, an association between cases may still occur because both the study factor and the potential confounder are risk factors for disease, but this is a consequence of those effects and therefore does not cause confounding.

Third, confounder cannot be intermediate variables between exposure and outcome. In other words, confounders cannot be intermediates in the causal pathway between exposure and disease, or a condition caused by the outcome. To do otherwise would introduce a serious bias. Hypothetically, in a study of overweight and the risk of cardiovascular disease, it would be inappropriate to control for diabetes as confounder if diabetes was a consequence of being overweight and is also a part of the causal chain leading to overweight and cardiovascular disease. On the other hand, assuming diabetes is studied directly as a primary interest, overweight would be regarded as a potential confounder if it also involved exposure to other risk factors for cardiovascular disease.

We discussed the misclassification of disease and exposure in information bias. Here we need to refer to the misclassification of confounders. The ability to control confounding in the analysis will be hindered If a confounder is misclassified. Although independent nondifferential misclassification of exposure or disease usually causes the study results to be biased in the direction of the null hypothesis, independent nondifferential misclassification of a confounder usually reduces the degree of control for confounding, which may lead to bias in either direction. For this

reason, misclassification of confounder can be a serious concern. If the confounding is robust and the exposure–disease relationship is weak or zero, misclassification of the confounder can yield highly misleading results, even if such misclassification is independent and non-differential.

8.4.3 Control

In the study design and analysis, confounding bias can be controlled by adjusting for all confounders or a sufficient subset of them at the same time. There are usually three methods to control for bias during the design stage.

8.4.3.1 Random Allocation

The first method is randomization, where participants are randomly assigned to exposure categories (applicable to experiments only). Ideally, we can create study cohorts with the equal incidence rate and eliminate the potential for confounding. But it must be practically and ethically feasible to assign exposure subjects. If just a few factors determine incidence, and the investigation personnel are aware of these factors, the ideal plan might call for exposure assignment that would result in the identical, balanced distributions of these disease causes in each group. Nonetheless, in studies of human disease, there are always immeasurable causes of disease that cannot be forced to be balanced amongst treatment groups. Randomization is one approach that permits one to probabilistically limit the confounding of unmeasured factors and to quantitatively account for the potential residual confounding arising from these unmeasured factors. However, this is usually only one alternative that may be beneficial for potential exposures. For instance, it is impractical and unethical to conduct randomized trials of the health effects of smoking, and therefore randomized trials may fail to prevent all confounding.

8.4.3.2 Restrict

The second control method is restriction, i.e., limiting the conditions of the study subject to a narrow range of values of the potential confounders. If a variable is prohibited from changing, it will not generate confounding if it is prohibited from varying. The restriction is a promising way to prevent or at least reduce confounding by known factors, it is both extremely effective and inexpensive. However, the advantages of restricting the study must be balanced against the disadvantages of reducing the study population when potential subjects are less plentiful. This approach has several conceptual and computational advantages, but may severely reduce the number of study subjects available and ultimately limit the extrapolation of results.

8.4.3.3 Matching

The third control method is matching, where study subjects are matched on the basis of potential confounders. Matching may be done by subject to subject, called individual matching, or for groups to groups, called frequency matching. Individual matching refers to the selection of one or more reference subjects with equal matching factor values to those of the index subject, whereas frequency matching refers to the selection of a whole stratum of reference subjects with similar matching-factor values to that of a stratum of index subjects. Individual matching would prevent age-gender-race confounding in cohort studies but is seldom done because it is very labor-intensive. In addition, matching does not completely eliminate confounding but does facilitate its control in case-control studies because matching for strong confounder will usually improve the precision of effect estimates. We have to discuss the concept of overmatching, which is often occurred in matched studies. In case-control studies, matching may be less accurate if the match factor related to exposure is only a weak risk factor for the disease of interest. When the number of matching factors exceeds 3, finding a suitable control becomes increasingly difficult.

8.4.3.4 Data Analysis

The above three control methods are usually implemented during the design phase. The analysis phase can also employ a number of methods to control for confounding bias. In the most straightforward situation, controlling for confounding in the analysis includes stratifying the data according to the level of confounders and calculating an effect estimate that summarizes the association between the strata of confounding factors. In a stratified analysis, it is usually not possible to control for more than two or three confounders at the same time, because finer stratification often results in many strata that contain no exposed or non-exposed individuals. Such strata are noninformative; therefore, a stratification that is too fine is a waste of information. In addition, we can use multi-factor analysis and standardized analysis to control confounding bias.