Kidney Disease and Population Health

nephron Clinical Practice

Nephron Clin Pract 2010;115:c94–c99
DOI: 10.1159/000312871

Published online: April 21, 2010

Selection Bias and Information Bias in Clinical Research

Giovanni Tripepi[a] Kitty J. Jager[b] Friedo W. Dekker[b, c] Carmine Zoccali[a]

[a]CNR-IBIM, Clinical Epidemiology and Physiopathology of Renal Diseases and Hypertension of Reggio Calabria, Reggio Calabria, Italy; [b]ERA-EDTA Registry, Department of Medical Informatics, Academic Medical Center, University of Amsterdam, Amsterdam, and [c]Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands

Key Words
Bias · Information bias · Selection bias · Systematic error

Abstract
The internal validity of an epidemiological study can be affected by *random error* and *systematic error*. Random error reflects a problem of precision in assessing a given exposure-disease relationship and can be reduced by increasing the sample size. On the other hand, systematic error or bias reflects a problem of validity of the study and arises because of any error resulting from methods used by the investigator when recruiting individuals for the study, from factors affecting the study participation *(selection bias)* or from systematic distortions when collecting information about exposures and outcomes *(information bias)*. Another important factor which may affect the internal validity of a clinical study is *confounding*. In this article, we focus on two categories of bias: selection bias and information bias. Confounding will be described in a future article of this series.

Introduction

In designing or interpreting a clinical study a researcher has two concerns: the external and the internal validity of the study. In the modern perspective proposed by Rothman [1], external validity includes scientific and statistical generalisation. Scientific generalisation is the characteristic of an epidemiological study whereby it may generate a coherent, potentially causal, biological hypothesis applicable to a more general set of clinical or epidemiological circumstances than the specific population under investigation [1]. Statistical generalisation is fundamental in survey sampling in which the resulting sample must be statistically representative of the source (or target) population [1]. The key difference between the two features of external validity is that scientific generalisation rests on biological rather than on statistical representativeness of the sample.

The internal validity, i.e. the characteristic of a clinical study to produce valid results, can be affected by random and systematic (bias) errors. Random error is due to chance and can be minimised by increasing the sample size or by decreasing the variation in measurements (reducing measurement error). Bias is any error resulting from methods used by the investigator to recruit individuals for the study, from factors affecting the study par-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1660–2110/10/1152–0094$26.00/0

Accessible online at:
www.karger.com/nec

Dr. Giovanni Tripepi, CNR-IBIM
Istituto di Biomedicina, Epidemiologia Clinica e Fisiopatologia delle Malattie Renali e dell'Ipertensione Arteriosa, c/o Euroline di Ascrizzi Vincenzo
Via Vallone Petrara n. 55/57, IT–89124 Reggio Calabria (Italy)
Tel. +39 0965 397 010, Fax +39 0965 26879, E-Mail gtripepi@ibim.cnr.it

ticipation *(selection bias)* or from systematic distortions when collecting information about exposures and diseases *(information bias)*. More generally, bias is any deviation in the collection, analysis, interpretation and publication of data leading to conclusions that systematically underestimate or overestimate the true relationship between a given exposure and a specific disease or any other outcome [2]. Bias cannot be minimised by increasing the sample size. Most violations of internal validity can be attributed to selection bias, information bias or confounding. In this article, we focus on some examples of selection bias and information bias.

Selection Bias

A selection bias comes from any error in selecting the study participants and/or from factors affecting the study participation. As a consequence, the relationship between exposure and disease differs between those included in the study and those potentially eligible for the study (including non-participants or non-responders). From this perspective, particularly for aetiological research, internal validity is a prerequisite for external validity. Because the exposure-disease relationship in non-participants is usually unknown, selection bias can only be hypothesised. In this article, we consider 5 types of selection bias: the non-response bias (example 1), the incidence-prevalence bias (examples 2 and 3), the loss-to-follow-up bias (example 4), the confounding by indication bias (example 5) and the volunteer bias (example 6).

Non-Response Bias

Example 1

A non-response bias occurs when the non-participation (non-response) is related to the exposure and, independently of exposure, to the disease/outcome. If the non-response is only related to the exposure and not to the disease, this affects the distribution of the exposure in the study but not the observed effect, e.g. the relative risk, and thus it does not affect the internal validity of the study. In a hypothetical study investigating the relationship between smoking and 10-year risk of renal dysfunction, we consider two scenarios: the first is the ideal scenario (universal agreement to take part in the study); the second one is a situation in which 20% of smokers with severe hypertension (i.e. smokers with increased risk of renal dysfunction) do not respond.

Ideal Scenario. In our hypothetical population there are 1,000 smokers and 1,000 non-smokers. All people agree to take part in the study (table 1). In smokers the 10-year risk of renal dysfunction is 0.10 or 10% [(100/1,000) × 100] and in non-smokers it is 0.05 or 5% [(50/1,000) × 100]. The risk ratio (RR) [3] is: RR = 0.10/0.05 = 2. Thus, the risk of renal dysfunction is 2 times higher in smokers than in non-smokers. Since this RR has been calculated from all eligible smokers and non-smokers, we consider this figure as the true RR.

Table 1. Ideal scenario, investigating the relationship between smoking and 10-year risk of renal dysfunction

	Individuals with renal dysfunction	Risk of renal dysfunction
Smokers (n = 1,000)	100	10%
Non-smokers (n = 1,000)	50	5%

Table 2. Second scenario, investigating the relationship between smoking and 10-year risk of renal dysfunction

	Individuals with renal dysfunction	Risk of renal dysfunction
Smokers (n = 800)	60	7.5%
Non-smokers (n = 1,000)	50	5%

Second Scenario. Twenty percent of smokers with severe hypertension (i.e. individuals with increased risk of renal dysfunction) do not accept to participate in the study (table 2). In this scenario, the non-response rate is related to the smoking status and to the risk of disease (in fact, independently of smoking, individuals with severe hypertension are more likely to have renal dysfunction). The RR is: RR = 0.075/0.050 = 1.5. This RR underestimates the true RR of the disease because the numerator does not include cases with renal dysfunction that could occur in non-participants (20% of smokers with severe hypertension, that is a population at high risk of renal disease). Although in this example the non-response leads to an underestimation of the true RR, this type of bias may also generate an overestimation of the RR, depending on the direction of the bias.

Incidence-Prevalence Bias

A selection bias particularly common in cross-sectional studies is the incidence-prevalence bias (also called the Neyman bias or survival bias). This bias occurs when the estimation of the risk of a disease is made by using

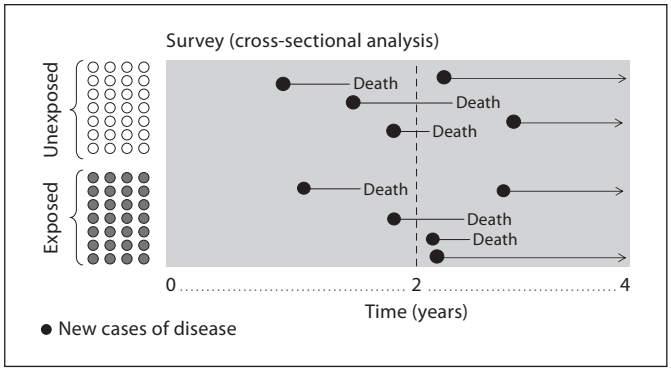

Fig. 1. Hypothetical example of incidence-prevalence bias (see text for details).

Table 3. Disease status in relation to multivitamin supplements

	Disease status	
	with ESRD (cases)	without ESRD (controls)
Users of multivitamin supplements	16	67
Non-users	184	133
Total	200	200

Table 4. Disease status in relation to multivitamin supplements (biased scenario)

	Disease status	
	with ESRD (cases)	without ESRD (controls)
Users of multivitamin supplements	32	67
Non-users	168	133
Total	200	200

data collected at a given point in time in a series of survivors rather than being based on data collected during a time period (example 2) or when the biased selection of cases produces a distorted frequency of exposure (example 3).

Example 2 (fig. 1)

We consider a hypothetical cohort study including 56 individuals: 28 exposed and 28 unexposed to a given risk factor. The study's aim is to investigate the relationship between the exposure to the risk factor and the risk of disease. The cohort is followed up for 4 years. During this period, the disease of interest occurs in 5 individuals in the exposed group and in 5 individuals in the unexposed group. Therefore, the RR of the disease (exposed vs. unexposed) is 1. We consider this figure (RR = 1) as the true RR. During the first 2 years, there is one death in the exposed group and another death in the unexposed group (fig. 1). If we perform a survey after 2 years, the prevalence ratio of disease between exposed and unexposed individuals is 0.5 [in fact, at 2 years of follow-up we have only 1 case of disease in 27 exposed individuals (prevalence of disease: 1/27 = 0.037 or 3.7%) and only 2 cases of disease in 27 unexposed individuals (prevalence of disease: 2/27 = 0.074 or 7.4%)], a value that deviates from the true RR (i.e. 1). For this reason, we should estimate the occurrence of a disease in terms of incidence (i.e. new cases occurring in a sample in a given time interval) rather than in terms of prevalence (i.e. cases counted at a given point in time) [4].

Example 3

In a case-control study, Tsai et al. [5] investigated the association between lifestyle factors and the odds of end-stage renal disease (ESRD) in Taiwan. Among the potential risk factors these authors considered the association between multivitamin supplements and ESRD (table 3). In this study, the odds ratio (OR) [3] of the use of multivitamin supplements is calculated by the standard formula: OR = (16/184)/(67/133) = 0.087/0.504 = 0.17. An OR of 0.17 means that the odds of exposure to multivitamin supplements were 83% lower in individuals with ESRD than in those without this complication. In this study, the selection of cases and controls and the assessment of multivitamin use were assumed to be unbiased, and we consider that this OR is the true OR.

In a hypothetical scenario where the selection of cases, but not that of controls, is biased (the investigator, influenced by previous knowledge of the exposure status, may collect cases mainly among individuals known to be multivitamin users), the frequency of cases may be spuriously higher (table 4). The OR in this situation is: OR = (32/168)/(67/133) = 0.190/0.504 = 0.38. In this scenario, the biased selection process of cases produces an important alteration in the estimate of OR (0.38 vs. 0.17). To avoid this problem, the selection process of cases and controls should be identical and should be independent of the exposure status (i.e. the investigator should be blinded to the exposure status).

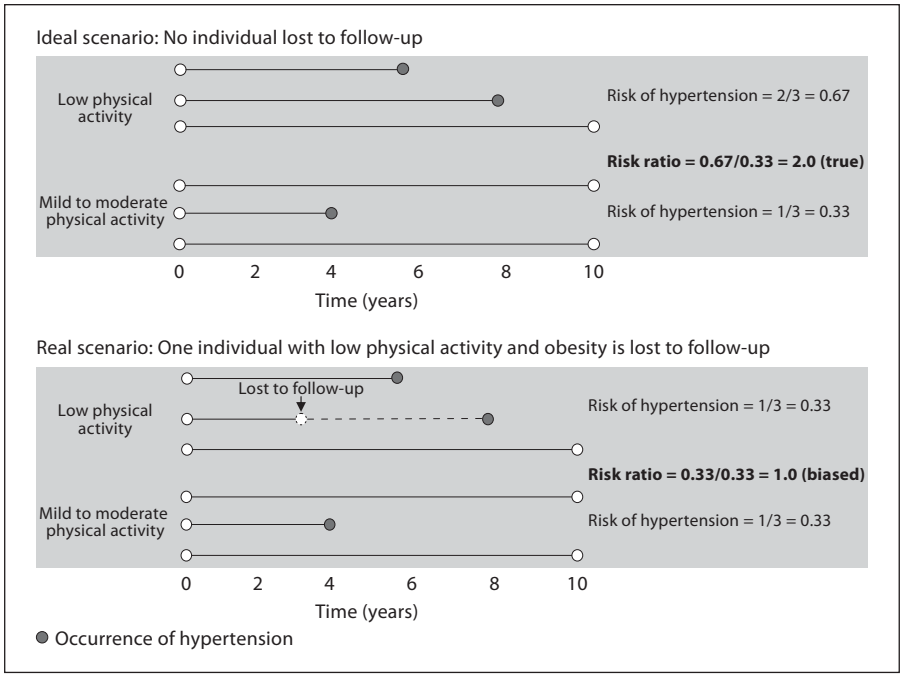

Fig. 2. Hypothetical example of loss-to-follow-up bias (see text for details).

Loss-to-Follow-Up Bias

Example 4

A loss-to-follow-up bias occurs in prospective cohort studies. With this type of bias, the true relationship between exposure and disease will only be distorted if the losses during follow-up are selective (non-random) with respect to both exposure and outcome. We consider a hypothetical cohort study investigating the relationship between physical activity and the 10-year risk of hypertension. Again, we imagine two situations (fig. 2): the first is the ideal scenario; the second one is a situation in which one individual with low physical activity and obesity (that is an individual at high risk for hypertension) is lost to follow-up. In the ideal scenario, the 10-year RR of hypertension in the low physical activity group as compared to that with mild to moderate physical activity is 2.0. In the second scenario, since the loss to follow-up of individuals with low physical activity is also affected by obesity (that is a co-morbid condition predisposing to hypertension independent of physical activity), the resulting RR (of 1) is overtly underestimated.

Confounding by Indication

Example 5

Confounding by indication is a type of bias that is generated when the indication to treat is a confounder for the treatment-outcome relationship. This bias occurs in observational studies of treatment efficacy, i.e. studies in which the allocation of patients to a specific treatment depends on an arbitrary decision of the investigator rather than on chance as in randomised clinical trials [6]. As a consequence the two study arms (active vs. placebo) are not comparable for relevant prognostic factors at baseline. Here, we consider 2 hypothetical studies investigating the effect of a new class of statins on the incidence rate of myocardial infarction in patients with hypercholesterolaemia: the first one is a randomised clinical trial, and the second one includes patients who are not randomly allocated to the active arm (that is they receive treatment on the basis of the doctor's decision). In the randomised clinical trial, the hazard ratio of myocardial infarction is 0.70, i.e. patients receiving the new statin have a hazard rate of myocardial infarction that is 30% lower than that in patients receiving the placebo. In the observational study the results are surprising because the hazard ratio of myocardial infarction (active arm vs. placebo) is 1.10, i.e. patients on treatment with the new statin have a hazard rate of myocardial infarction that is 10% higher than that in patients on placebo. This is because in the observational study the investigator treats patients with a more compromised prognosis at baseline more frequently with the new statin (a drug that he considers of higher efficacy in comparison to previous treatments), thus generating a bias due to the absence of comparability between the two study groups.

Volunteer Bias

Example 6

A volunteer bias (or self-selection bias) occurs when individuals who volunteer for a study differ in relevant clinical characteristics from those who do not. The self-selection is a threat for the internal validity of the study if it is related to the exposure and, independently of exposure, to the disease/outcome. In a prospective, observational, study in the general population, Ganguli et al. [7] assessed the prognostic implications of the volunteer bias by comparing the mortality rate in 1,366 individuals recruited through intensive enrollment efforts and in 315 volunteers who agreed to take part in the study after just one mailing. At enrollment, the volunteers were more frequently women, with higher education and cognitive test score and less likely to use the health service when compared to non-volunteers. During 6–8 years' follow-up, the mortality rate was much lower in the volunteers than in the remaining individuals. The authors concluded that health-related studies with populations composed partly or entirely of volunteers should take a potential volunteer bias into account when analysing and interpreting data. A volunteer bias cannot occur in randomised studies in which subjects are randomised only after agreeing to participate [8].

Information Bias

An information bias occurs during data collection. The most important type of information bias is the misclassification bias. A misclassification bias is present when the detection of the exposure status (exposure identification bias) and/or the disease assessment (disease identification bias) is biased, i.e. exposed/diseased individuals are classified as non-exposed/non-diseased and vice versa. In clinical practice, a common source of misclassification derives from the inaccuracy of some diagnostic tests. Misclassification can be *non-differential* or *differential*.

Non-Differential Misclassification

Example 7

In clinical research the accuracy of any exposure-disease relationship depends on the performance of the diagnostic test used for assessing the exposure or for establishing the disease. Here we focus on exposure misclassification. In non-differential misclassification the performance of the diagnostic test (that is the ability of the test to correctly classify individuals as truly exposed/unexposed to a given risk factor) is the same in cases and in controls. In differential misclassification, the performance of the diagnostic test for the exposure identification differs between cases and controls.

We consider a case-control study [adapted from 9] investigating the association between AIDS and oesophageal candidiasis. In the first scenario (ideal scenario) the presence of oesophageal candidiasis was ascertained by the gold standard (biopsy) and in the second scenario by a self-report questionnaire, a method which may generate non-differential misclassification (table 5). The OR of oesophageal candidiasis between cases and controls is calculated as the ratio between the two odds: OR = (20/480)/(5/995) = 0.042/0.005 = 8.4. Since oesophageal candidiasis was assessed by the gold standard (biopsy) and since we assume there is no selection bias, we consider 8.4 as the true OR.

To elucidate the distortion of the OR estimate attributable to the use of the questionnaire, we consider table 6 in which individuals are allocated on the basis of a self-report questionnaire. In this instance, the OR of oesopha-

Table 5. Ideal scenario: oesophageal candidiasis established by biopsy

Oesophageal candidiasis (biopsy)	AIDS cases	Controls
Present	20	5
Absent	480	995
Total	500	1,000

Table 6. Second scenario: oesophageal candidiasis assessed by a self-report questionnaire

Oesophageal candidiasis (questionnaire)	AIDS cases	Controls
Present	86	161
Absent	414	839
Total	500	1,000

Table 7. Oesophageal candidiasis assessed by a self-report questionnaire (differential misclassification)

Oesophageal candidiasis (questionnaire)	AIDS cases	Controls
Present	86	400
Absent	414	600
Total	500	1,000

geal candidiasis as assessed by the questionnaire is: OR = (86/414)/(161/839) = 0.21/0.19 = 1.1. Due to the fact that the self-report questionnaire is an inaccurate method to identify individuals affected by oesophageal candidiasis, the resulting OR is biased. With two exposure categories (presence/absence of oesophageal candidiasis) non-differential misclassification always affects the OR toward 1.

Differential Misclassification

Example 8

In differential misclassification the performance of a test for exposure identification differs between cases and controls. We consider a new hypothetical scenario in which the self-report questionnaire for identifying individuals affected by oesophageal candidiasis has a different performance in cases and controls (table 7). The OR of oesophageal candidiasis as assessed by the self-report questionnaire is: OR = (86/414)/(400/600) = 0.21/0.67 = 0.3. Here, differential misclassification leads to an underestimation of the strength of the association between exposure and disease.

In general, differential misclassification may either increase or decrease the strength of reported associations, depending on the direction of the misclassification. Common causes producing misclassification are: *recall bias* (a bias that results from imprecise memory of past exposures); *interviewer bias* (the tendency of the interviewer to obtain answers that support preconceived notions); *observer bias* (resulting from the outcome assessor's knowledge of exposure status), and *regression dilution bias* (a bias related to regression to the mean which originates in longitudinal studies investigating the association between baseline measurements of a continuous variable and the risk of a given outcome) [for a complete review of these biases, see 10].

Conclusions

Bias is an unavoidable problem in clinical and epidemiological research. However, the correct selection of the study design, the careful choice of procedures of data collection and handling and the correct definition of exposure and disease represent important prevention strategies for minimising systematic errors in clinical research.

References

1 Rothman KJ: Epidemiology. An Introduction. New York, Oxford University Press, 2002, pp 20–21.
2 Last J (ed): A Dictionary of Epidemiology, ed 3. Oxford, Oxford University Press, 1988.
3 Tripepi G, Jager KJ, Dekker FW, Zoccali C: Measures of effect in epidemiological research. Nephron Clin Pract 2010;115:c91–c93.
4 Noordzij M, Dekker FW, Zoccali C, Jager KJ: Measures of disease frequency: prevalence and incidence. Nephron Clin Pract 2010; 115:c17–c20.
5 Tsai SY, Tseng HF, Tan HF, Chien YS, Chang CC: End-stage renal disease in Taiwan: a case-control study. J Epidemiol 2009;19:169–176.
6 Stel VS, Zoccali C, Dekker FW, Jager KJ: The randomized controlled trial. Nephron Clin Pract 2009;113:c337–c342.
7 Ganguli M, Lutle ME, Reynolds MD, Dodge HH: Random versus volunteer selection for a community-based study. J Gerontol A Biol Sci Med Sci 1998;53:M39–M46.
8 Hernan MA, Hernandez-Diaz S, Robins JM: A structural approach to selection bias. Epidemiology 2004;14:615–625.
9 Hessol NA, Schwarcz S, Ameli N, Oliver G, Greenblatt RM: Accuracy of self-reports of acquired immunodeficiency syndrome and acquired immunodeficiency syndrome-related conditions in women. Am J Epidemiol 2001;153:1128–1133.
10 Tripepi G, Jager KJ, Dekker FW, Zoccali C: Bias in clinical research. Kidney Int 2008;73: 148–153.