Neuroepidemiology 2010;34:264–266

DOI: [10.1159/000297757](https://doi.org/10.1159/000297757)

Case-Control Studies: The Importance of Design and Conduct

Amanda G. Thrift

Baker IDI Heart and Diabetes Institute, and Department of
Epidemiology and Preventive Medicine, Monash University,
Melbourne, Vic., Australia

Introduction

Case-control studies, also termed retrospective studies, are observational studies designed primarily to identify risk factors and causes of disease, particularly rare diseases [1]. They are popular because they are relatively inexpensive to undertake and can provide information while an issue is still topical [2–4].

Case-control studies have several major disadvantages that restrict their applicability and reliability. The most important of these is their potential to generate spurious associations because of various forms of bias, by the action of ‘confounding variables’, or because many possible risk factors have been studied in a single investigation [3, 4]. This article comprises an overview of the methods used in a case-control study.

Selection of Cases

Cases should be representative of all cases that occur [5]. Ideally, a defined population is chosen and a system is established to identify new incident cases as they occur [6]. Incident cases are generally preferred to prevalent cases as prevalent cases over-represent mild cases with a good prognosis [1, 4, 5].

A major strength of a case-control study is that cases with a single pathological entity can be selected [4, 5]. Participants may be required to undergo diagnostic tests to confirm the presence of the disease, such as the use of imaging or autopsy to confirm the presence of intracerebral haemorrhage [7]. These tests enable investigators to be confident that they are studying the disease in question.

Selection of Controls

Controls should be a group of individuals similar to cases but free from the disease in question [3, 8]. Essentially, the population from which the controls are chosen should include all those, and only those, who would have been included as cases if they had developed the disease [4, 8]. When cases have been identified within a defined population, the appropriate controls would be those randomly selected from a register of the same community [4], such as an electoral roll. If this is the method used, then care

should also be taken to include only those cases who also appear on this register.

When cases have been selected from hospitals, it may be difficult to define the exact population from which they have been drawn. In this situation, alternative strategies would be to use neighbourhood controls or random telephone dialling [1, 4, 9–11]. Neighbourhood controls are obtained by visiting the street where the case lived and identifying the closest individual according to a carefully defined protocol. Random telephone dialling is unpopular as it is likely to result in low response rates. This is because of increases in (1) telemarketing and (2) the use of mobile telephones, with consequent reduced coverage of domestic telephones. Other methods of control selection include the use of friends, family members or workmates [1, 4], but their similarity in drinking and smoking habits, and possibly in their choice of hobbies and other activities, might lead to overmatching. Hospital controls are also sometimes used because of their similarity in terms of any selection processes that are involved [12]. However, the difficulty in restricting controls to an unbiased group of hospital admissions has made this strategy unpopular.

In some instances, it may be necessary to screen controls to make sure that they do not have the disease. For most serious illnesses, e.g. intracerebral haemorrhage, it is reasonable to assume that if the condition has not been expressed clinically then it is not likely to be present. In these instances screening controls is unnecessary.

Exposure Ascertainment

Past exposure is usually determined by questionnaire. It is important that questions focus on the relevant time period [3, 13]. For example, most known causes of solid malignancy operate 20–30 years before clinical manifestations become evident, and questions should be directed towards this time period. When the pathophysiology of the condition suggests that aetiological factors operating at the moment of event onset may be relevant, questions should instead be focussed on both recent and past events.

Bias Control

Each of the major processes involved in conducting a case-control study can involve bias. This, in turn, can produce a spurious finding of an association between an exposure and a disease, when in fact none exists [1]. Alternatively, the bias might conceal a relationship.

Case Selection

Selection bias may occur when included cases differ systematically from those not included. This occurs particularly when cases are chosen from diagnosed and accessible cases rather

than from a register of all cases [1, 5]. When cases are selected from hospital sources, bias may occur if these cases differ from cases not hospitalised. For example, cases of dementia are likely to present at hospital mainly because of co-existing illness. Cases chosen from this source may be unrepresentative in terms of smoking and alcohol since these are commonly associated with the typical respiratory and cardiovascular illnesses that would lead to hospitalisation. Particular care should be taken if the reason for non-availability is relatively common among people with the disease. Every effort should be made to include those people so that the study sample is representative and bias is minimised.

Control Selection

Selection of controls may also produce bias if the controls are not representative of the population from which the cases were derived. This form of bias may occur in many situations. For example, bias may occur if neighbourhood controls are selected from only those who are at home during the day. On average, this will over-represent the unemployed and shift workers who may be of a generally lower socioeconomic strata. To minimise this form of bias, interviewers may be instructed to make repeated after-hours visits to 'empty' houses to ensure that potential controls who work during the day can be included. A high response rate and a careful protocol will improve the representativeness of the controls.

Information Bias

Information/measurement bias refers to the potential to create or conceal a relationship between an exposure and a disease as a result of a systematic error in determining exposure [5]. Information bias is particularly likely to occur if different sources of information are used [3]. For example, if cases are interviewed in hospital and controls are interviewed in their homes, interviewers may supplement the information obtained from cases by referring to the case's hospital record. This source would not be available for controls. To avoid this type of bias, interviewers must be trained to question cases and controls with similar degrees of probing, and avoid utilising any sources of data that are not equally accessible for cases and controls [1, 3]. The use of identical sources of data for cases and controls is critical [3].

Case-control studies are reliant on memory of past events, and bias may result from differential recall of information [1, 2, 6, 8, 12, 14]. Data collection may be particularly unreliable when cases are neurologically impaired as this impairment may affect the case's ability to recall past events. Poor recall may make cases respond positively less often to questions about past exposure, leading to an underestimation in the apparent level of association between exposure and disease [1, 2, 6, 8, 12]. Alternatively, bias in an 'opposite' direction might be expected to occur whenever cases are more likely to have seriously contemplated their prior exposures in a more intense fashion than the healthy control. Most people developing a serious illness probably consider the possible causes and therefore this type of bias is difficult to avoid.

Numerous strategies can be employed to counter the effect of recall bias. Instead of interviewing people with memory impairment, information could be collected from a proxy respondent familiar with the habits of the case [10]. Matched controls should also have a similarly related proxy to answer questions to ensure equivalence of information sources. Proxies may be unaware of

detailed and specific information. In this instance, alternative sources of information could be used. For example, doctors' prescription records could be used to obtain medication history for both cases and controls [15].

Another useful strategy is to include questions that can be regarded as 'positive control' and 'negative control', i.e. factors known to have (and known not to have) an association with the disease in question [3]. For example, an association identified between aspirin and intracerebral haemorrhage will have more credibility if there is no association with paracetamol, as paracetamol has no platelet effect [16]. On the other hand, a global finding of a weak positive or negative association with most of the elements in the questionnaire will be more suggestive of a systematic bias.

The success, or otherwise, in controlling information bias can never be precisely determined and for this reason the various forms of bias can never be excluded as a cause of spurious associations.

Confounding

Another problem with case-control studies is the difficulty in isolating the effect of the factor of interest from others that it may be associated with [4, 5, 10, 14]. For example, investigators have shown that hypertension occurs more commonly in older people than in younger people. For this reason, an association between hypertension and a disease may be affected by the causal association between older age and the disease. In this situation, the relationship between hypertension and the disease is said to be confounded by age.

The influence of confounding variables identified in advance can be eliminated in the design stages of an epidemiological study by:

- exclusion from the study, e.g. if smoking is a potential confounding factor, then the study can be conducted entirely in nonsmokers [14];
 - matching to ensure that the potentially confounding variables are evenly distributed in the 2 groups [3, 4, 10, 14].
- At the analysis stage, confounding can be controlled by either:
- stratification, i.e. measurement of the strength of association in well-defined strata of the confounding variable [14];
 - multiple logistic regression, i.e. estimating the strength of association while controlling for a number of confounding variables simultaneously [3, 14].

The incidence of many diseases differs between the two sexes and increases sharply with age. Thus, differences between the age and sex distribution of the cases and controls are particularly likely to produce confounding. For this reason, cases are often age- and sex-matched to controls to eliminate this source of confounding [3, 5]. Such matching is not essential, but when a disease has an age or sex distribution very different to the general population the accuracy of statistical adjustment is substantially reduced.

Socioeconomic status is another potential confounding factor because many diseases are more likely to occur in lower socioeconomic groups [5, 17]. This potential source of confounding can be minimised by matching cases and controls on neighbourhood of residence. Care must be taken not to overmatch or the controls may become too similar to the cases leading to an underestimation of risk [4, 5].

When potential confounding factors have been considered in advance and measured, their influence can generally be elimi-

Table 1. Criteria for causation

-
- Temporal relationship
(risk factor precedes disease development)
 - Odds ratio is large
 - Dose-response relationship is evident
 - Association is biologically plausible
 - Consistent findings between studies
 - Removal of the factor results in a reduced disease risk
 - Relationship is specific to a single exposure and absent for 'control' exposures
-

nated during the analysis [5]. However, the possibility that an apparent relationship is in reality due to the action of an unknown and unrecognised confounding factor affects most case-control studies [4].

Conclusions

Case-control studies represent a strategy to identify or confirm risk factors as causes of diseases. These studies allow many potential factors to be related to the occurrence of a single disease. Many associations can be generated by chance, and therefore most significance must be attached to associations that are strong and declared as hypotheses to be examined in advance of the analysis being undertaken.

Case-control studies are susceptible to many sources of bias and to confounding. Some sources of bias can be minimised by appropriate design, while confounding can be minimised in either the design phase or in the analysis.

By themselves, case-control studies reveal associations and do not prove causation. However, causation might be suggested if certain criteria are met (table 1). Most concern about the validity of findings is generated by weak associations which can be easily generated by the types of bias described above. Weak associations require considerable support from the criteria for causation before they can be considered as evidence of causation.

References

- 1 Schlesselman JJ: Case-Control Studies: Design, Conduct, Analysis. New York, Oxford University Press, 1982.
- 2 Sackett DL: Bias in analytical research. *J Chronic Dis* 1979;32:51-63.
- 3 Jick H, Vessey MP: Case-control studies in the evaluation of drug-induced illness. *Am J Epidemiol* 1978;107:1-7.
- 4 Breslow NE, Day NE: Statistical Methods in Cancer Research. Lyon, International Agency for Research on Cancer Scientific Publications, 1980, vol 1: The Analysis of Case-Control Studies.
- 5 Ibrahim MA, Spitzer WO: The case control study: the problem and the prospect. *J Chronic Dis* 1979;32:139-144.
- 6 Horwitz RI, Feinstein AR: Methodologic standards and contradictory results in case-control research. *Am J Med* 1979;66:556-564.
- 7 Thrift AG, McNeil JJ, Donnan GA: The risk of intracerebral haemorrhage with smoking. *Cerebrovasc Dis* 1999;9:34-39.
- 8 Kopec JA, Esdaile JM: Bias in case-control studies. a review. *J Epidemiol Community Health* 1990;44:179-186.
- 9 Waksberg J: Sampling methods for random digit dialing. *J Am Stat Assoc* 1978;73:40-46.
- 10 Kelsey JL, Thompson WD, Evans AS: Methods on Observational Epidemiology. New York, Oxford University Press, 1986.
- 11 Greenberg ER: Random digit dialing for control selection: a review and a caution on its use in studies of childhood cancer. *Am J Epidemiol* 1990;131:1-5.
- 12 Stavray KM, Clarke EA: Hospital or population controls? An unanswered question. *J Chronic Dis* 1983;36:301-307.
- 13 Whittemore AS: The age distribution of human cancers for carcinogenic exposure of varying intensity. *Am J Epidemiol* 1977;106:418-432.
- 14 Beaglehole R, Bonita R, Kjellström T: Basic Epidemiology. Geneva, World Health Organization, 1993.
- 15 Thrift AG, McNeil JJ, Forbes A, Donnan G: Low risk of primary intracerebral haemorrhage among users of aspirin and non-steroidal anti-inflammatory drugs. *Br Med J* 1999;318:759-764.
- 16 Insel PA: Analgesic-antipyretics and antiinflammatory agents: drugs employed in the treatment of rheumatoid arthritis and gout; in Gilman AG, Rall TW, Nies AS, Taylor P (eds): The Pharmacological Basis of Therapeutics. New York, McGraw-Hill, 1992, vol 1, pp 638-674.
- 17 Thrift AG, Dewey HM, Sturm JW, Paul S, Gilligan AK, Srikanth VK, Macdonell RAL, McNeil JJ, Macleod MR, Donnan GA: Greater incidence of both fatal and non-fatal strokes in disadvantaged areas: The North East Melbourne Stroke Incidence Study. *Stroke* 2006;37:877-882.

Assoc. Prof. A. Thrift
Stroke Epidemiology, Baker IDI Heart and Diabetes Institute
75 Commercial Road
Melbourne, Vic. 3004 (Australia)
Tel. +61 3 8532 1100, Fax +61 3 8532 1111
E-Mail amanda.thrift@bakeridi.edu.au