

ENDGAMES

STATISTICAL QUESTION

Case-control studies: advantages and disadvantages

Philip Sedgwick *reader in medical statistics and medical education*

Centre for Medical and Healthcare Education, St George's, University of London, London, UK

Researchers investigated the risk factors associated with the development of pulmonary tuberculosis in Russia. A case-control study was performed in the city of Samara, 700 miles south east of Moscow. Cases were 334 consecutive adults diagnosed as having culture confirmed pulmonary tuberculosis at any of the city's specialist tuberculosis clinics between 1 January 2003 and 31 December 2003. For each case, a control matched for year of birth and sex, and with no history of tuberculosis, was sampled randomly from a registry of the general population of Samara city. A questionnaire was used to collect information retrospectively about potential risk factors before and during the development of pulmonary tuberculosis. Controls were asked about exposure to risk factors before the index date for their matched case—that is, the date when tuberculosis was diagnosed.¹

The researchers reported that the most important risk factors associated with the development of pulmonary tuberculosis were raw milk and unemployment.

Which of the following statements, if any, are true?

- The sampling of the controls was prone to selection bias
- The information collected by the questionnaire was prone to recall bias
- It was possible to estimate the population at risk of pulmonary tuberculosis
- It can be inferred that raw milk and unemployment cause pulmonary tuberculosis

Answers

Statement *b* is true, whereas *a*, *c*, and *d* are false.

The purpose of the study was to establish those risk factors associated with the development of pulmonary tuberculosis. A case-control study design was used. Two groups of people were identified on the basis of their health status—those with tuberculosis (the cases) and otherwise healthy people with no history of pulmonary tuberculosis (the controls). A case-control study is retrospective in design. In the example above, information about past exposure to potential risk factors before and during the development of pulmonary tuberculosis was collected by questionnaire. The cases and controls were

compared to ascertain whether particular risk factors were more common in one group than in the other.

It is important that consideration is given to the selection of controls for a case-control study. Typically, the controls will have had no history of the disease or condition of interest. Furthermore, the controls should be representative of the population. However, controls are often recruited through convenience sampling—for example, from a hospital clinic or a general practice. Therefore, any resulting sample of controls would not be representative of the general population in terms of health. Hence, the recruitment of controls is typically prone to selection bias—that is, the controls are systematically different from the population they are meant to represent. Any observed differences between cases and controls in the measured risk factors would not reflect that in the general population. In the example above, the controls were selected at random from a registry of the general population of Samara city. Because the controls were selected at random, they would not be prone to selection bias (*a* is false). Case and controls were matched for year of birth and sex. The advantages of matching in case-control studies have been described in a previous question.²

A questionnaire was used to collect information about past exposure to a variety of risk factors. The study was retrospective in design, with participants reporting exposure to risk factors before and during the development of pulmonary tuberculosis. Therefore, the information recorded would have been prone to recall bias (*b* is true). Recall bias, described in a previous question,³ is the systematic difference between the cases and controls in the accuracy of reported information about past exposure to risk factors. Recall bias will be present if participants have selective preconceptions about the association between pulmonary tuberculosis and past exposure to the risk factor(s).

Relative risk would have been the preferred measure of the association between pulmonary tuberculosis and each recorded risk factor.⁴ However, in the example above it was not possible to calculate the relative risk of pulmonary tuberculosis for those with a particular risk factor present relative to those without. This is because it is not possible to estimate directly the population at risk in a case-control study (*c* is false), as described

in a previous question.⁵ Estimating the population at risk would involve estimating the incidence or prevalence of pulmonary tuberculosis in the population. This would be not only for the entire population, but also those with and without each risk factor present. For a case-control study the odds ratio can be derived instead as an estimate of the relative risk. Odds and odds ratios have been described previously.⁶ Adjusted odds ratios were derived in the example above—that is, confounding was adjusted for to allow for the simultaneous effects of other variables studied. Odds ratios can be adjusted for confounding using a statistical method known as logistic regression.⁷

The researchers reported that development of pulmonary tuberculosis was associated with exposure to the risk factors of raw milk and unemployment. However, it cannot be inferred that raw milk or unemployment causes pulmonary tuberculosis (*d* is false), only that those people who had drunk raw milk or who had been unemployed were more likely to have developed pulmonary tuberculosis. This is because it is not always possible in case-control studies to predict whether exposure to the risk factors preceded development of the disease or condition. Furthermore, it was not possible to measure and then control for, through statistical analysis, all factors that may have affected the development of pulmonary tuberculosis. The observed associations between pulmonary tuberculosis and raw milk plus unemployment may have been the result of confounding—other risk factors that were not measured may have been associated with raw milk or unemployment and been the cause of pulmonary tuberculosis. Only an association between a risk factor and disease or condition, and not causation, can be inferred from the results of a case-control study. This is in contrast to an experimental study, such as a clinical trial, that uses random allocation to control for confounding at baseline. Case-control studies are observational by design. Other types of observational studies include prospective cohort studies.⁸ An

observational study is one in which researchers do not intervene in any way but simply observe and record people's behaviour, symptoms, attitudes, or other characteristics.

Case-control studies are generally quick, cheap, and easy to perform. Cases and controls are often sampled from, for example, an existing database of health records on a group of patients. Furthermore, case-control studies are particularly suitable for studying risk factors associated with rare diseases or conditions. In contrast, an observational design such as a prospective cohort study would not be suitable if the disease or condition is rare because it is unlikely that many members of a cohort will develop the disease or condition of interest. Case-control studies are not prone to loss to follow-up, unlike cohort studies. Sometimes case-control studies are performed as initial studies to establish potential associations before undertaking larger and more expensive studies. A disadvantage of case-control studies, in addition to those described above, is that they are not suitable when exposure to any of the risk factors is rare because few, if any, of the cases or controls are likely to have been exposed to them.

Competing interests: None declared.

- 1 Coker R, McKee M, Atun R, Dimitrova B, Dodonova E, Kuznetsov S, et al. Risk factors for pulmonary tuberculosis in Russia: case-control study. *BMJ* 2006;332:85.
- 2 Sedgwick P. Why match in case-control studies? *BMJ* 2012;344:e691.
- 3 Sedgwick P. What is recall bias? *BMJ* 2012;344:e3519.
- 4 Sedgwick P, Marston L. Relative risks. *BMJ* 2010;341:c3983.
- 5 Sedgwick P. Case-control studies: measures of risk. *BMJ* 2013;346:f1185.
- 6 Sedgwick P. Odds and odds ratios. *BMJ* 2013;347:f5067.
- 7 Sedgwick P. Logistic regression. *BMJ* 2013;347:f4488.
- 8 Sedgwick P. Prospective cohort studies: advantages and disadvantages. *BMJ* 2013;347:f6726.

Cite this as: *BMJ* 2014;348:f7707

© BMJ Publishing Group Ltd 2013