BMJ

BMJ 2013;346:f25 doi: 10.1136/bmj.f25 (Published 4 January 2013)

ENDGAMES

STATISTICAL QUESTION

Analysing case-control studies: adjusting for confounding

Philip Sedgwick *reader in medical statistics and medical education*

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

A case-control study examined the association between chronic *Helicobacter pylori* infection and coronary heart disease at young ages. In total, 1122 survivors of suspected acute myocardial infarction at age 30-49 years were recruited. For each case, a control matched for age and sex with no history of coronary heart disease was enrolled. Chronic infection with *H pylori* was confirmed serologically. Information on other risk factors for coronary heart disease was collected, including smoking behaviour, indicators of socioeconomic status, obesity, and blood lipid concentrations. Controls were asked about their current habits and history, whereas cases were asked about their habits and history just before their index myocardial infarction. Blood samples were obtained from cases within 24 hours of the onset of symptoms and from controls after collection of the information about risk factors.[1]

Early onset myocardial infarction was significantly associated with seropositive *H pylori* infection antibodies (odds ratio 2.28 (99% confidence interval 1.8 to 2.9)). The odds ratio was reduced to 1.87 (1.42 to 2.47) after adjustment for smoking and indicators of socioeconomic status and to 1.75 (1.29 to 2.36) after additional adjustment for blood lipid concentrations and obesity. Therefore, a moderate association existed between coronary heart disease and *H pylori* infection seropositivity that could not be fully explained by other risk factors.

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

a) Matching ensured that any differences between cases and controls were not due to differences in age and sex.

b) The case-control study estimated the population at risk.

c) The adjusted odds ratios could have been derived using logistic regression.

d) The association between *H pylori* seropositivity and coronary heart disease was independent of smoking and indicators of socioeconomic status.

Answers

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

The purpose of the case-control study was to investigate whether *H pylori* seropositivity was a potential risk factor for acute myocardial infarction at young ages. Two groups of patients were selected on the basis of their disease status: the cases, who were survivors of suspected acute myocardial infarction and were aged between 30 and 49 years; and healthy controls matched for age and sex. The cases and controls were compared to ascertain whether *H pylori* seropositivity occurred more often in one group than the other. If so, it would be a potential risk factor for acute myocardial infarction at young ages.

In the investigation of the association between *H pylori* seropositivity and coronary heart disease there was potential for confounding. Any confounding could result in a spurious statistical association or even cause an association to be missed. Confounding would have occurred if there was a difference between cases and controls in demographic characteristics or in prognostic factors that influence the association between *H pylori* seropositivity and coronary heart disease at young ages. These factors would have included those risk factors on which data were collected, such as age, sex, smoking behaviour, indicators of socioeconomic status, obesity, and blood lipid concentrations.

To illustrate the phenomenon of confounding in the above study, consider smoking status. Smoking is associated with an increased risk of coronary heart disease and is also related to *H pylori* infection: smokers are more likely to have *H pylori* infection than non-smokers. If *H pylori* infection was found to occur more often among cases than controls, it might be difficult to ascertain whether coronary heart disease was associated with the increased frequency of chronic *H pylori* infection or with the increased frequency of smoking. Therefore, unless the effects of smoking were adjusted for it may confound any association between *H pylori* infection and coronary heart disease. Confounding in the context of clinical trials was discussed in a previous question.[2] In the above study confounding was accounted for in two ways: matching cases and controls with regard to age and sex at the design stage; and adjusting for potential confounders at the analysis stage.

p.sedgwick@sgul.ac.uk

Matching is done on suspected confounding variables, and in the above study these were age and sex. Matching cases and controls with regard to age and sex meant that for each case who was recruited a control of the same age and sex was found. In the above study, matching on age involved finding a control whose age was within five years of that of the matching case. As the cases and controls were comparable for age and sex, any differences between them in *H pylori* seropositivity could not be due to differences in age and sex—that is, potential confounding resulting from age and sex differences was minimised (*a* is true). Matching cases and controls for age and sex was more efficient than adjusting for potential confounding related to age and sex differences during the statistical analysis, described below.

It was not possible to estimate the population at risk from the case-control study (*b* is false). Estimating the population at risk has been described in a previous question.[3] It would involve estimating what proportion of the population, with and without the risk factor (*H pylori* seropositivity), would develop coronary heart disease at young ages. In the above study, participants were initially identified by their disease status (case or control) in equal proportions. Information about potential risk factors for coronary heart disease was subsequently collected retrospectively. Cases and controls would obviously not be in equal proportions in the population. Therefore, the proportion of study participants with and without the risk factor who experienced coronary heart disease would not estimate the population at risk, and relative risks could not be obtained to estimate the association between *H pylori* seropositivity and coronary heart disease. Odds ratios, described in a previous question,[4] were derived instead.

The unadjusted odds ratio for the association between *H pylori* seropositivity and coronary heart disease was presented first. The unadjusted odds ratio, sometimes referred to as a crude odds ratio, had not been adjusted for potential confounding. The odds ratio was subsequently adjusted for potential confounding through use of a statistical method known as logistic regression (*c* is true). Adjustment for confounding was done in two stages: firstly for smoking and indicators of socioeconomic status; and then additionally for blood lipid concentrations and obesity. This allowed for the magnitude of confounding for each set of variables to be investigated in a stepwise fashion. By adjusting or controlling for a number of confounders simultaneously, the true association between chronic *H pylori* infection and coronary heart disease could be estimated. Adjusting for confounding quantifies the association in participants with assumed similar smoking and socioeconomic status, together with equal blood lipid concentrations and degree of obesity. In effect, it removes any differences between the categories or values of each confounding variable in the association between *H pylori* seropositivity and coronary heart disease.

The researchers presented odds ratios with 99% confidence intervals rather than standard 95% ones. The critical level of significance for the study was therefore set at 1%. This therefore meant that the association between *H pylori* seropositivity and coronary heart disease had to be stronger, in a statistical sense, for it to be significant—that is, the associated P value had to be less than 0.01 instead of 0.05, as is typical. The researchers did this because the study was exploratory. Setting the critical level of significance at 1% meant that it was less likely that significant relationships would be found or that type I errors (when the null hypothesis of no difference between cases and controls is erroneously rejected) would occur.

If the association between a risk factor and disease remains significant after adjustment for confounding, it is said to be independent of the potential confounding variables that were controlled for. In the above example, the unadjusted odds ratio for the association of *H pylori* seropositivity with coronary heart disease was 2.28 (99% confidence interval 1.8 to 2.9). The odds ratio was significant at the 1% level because the 99% confidence interval did not include one (unity). The odds ratio was reduced to 1.87 (1.42 to 2.47) after smoking and indicators of socioeconomic status were controlled for. Because the association remained significant after adjustment for confounding, it was independent of smoking and socioeconomic status (*d* is true). However, because the odds ratio was reduced in value after controlling for confounding, the association was partly explained by smoking and socioeconomic status. After the additional adjustment, for blood lipid concentrations and obesity, the odds ratio was further reduced to 1.75 (1.29 to 2.36). Therefore, the association was also partly explained by blood lipid concentrations and obesity. However, the association between *H pylori* infection and coronary heart disease remained significant and therefore was independent of blood lipid concentrations and obesity.

The sample odds ratio estimates the relative risk in the population. It has been proposed that the odds ratio is a good estimate when the disease or outcome is rare in the population, typically considered when the prevalence is less than 10%. This would no doubt be true for acute myocardial infarction at young ages. From the fully adjusted odds ratio it could be estimated that there is a 75% higher risk of coronary heart disease if seropositive *H pylori* antibodies are present than if there is no infection.

In the above example, only *H pylori* infection as a risk factor for coronary heart disease was reported. Other variables were adjusted for, as potential confounders, but their effects as risk factors were not reported. Typically in a case-control study the effects of more than one potential risk factor would be investigated. The logistic regression model explores the effects of potential risk factors for a disease simultaneously, adjusting for confounding by all the other variables included in the study. Cases and controls were matched for age and sex in the above study. Therefore, because the two groups were similar in these variables, their effects as risk factors or confounders could not be examined through the use of logistic regression. Although age and sex are risk factors for coronary heart disease, they were not of interest in this study. Cases and controls tend not to be matched on more than three variables, because to match on any more typically makes it difficult to find enough controls. A large number of potential controls is needed when matching on three variables, such as age, sex, and ethnicity.

1 Danesh J, Youngman L, Clark S, Parish S, Peto R, Collins R for the International Studies of Infarct Survival (ISIS) Collaborative Group. *Helicobacter pylori* infection and early onset myocardial infarction: case-control and sibling pairs study. *BMJ* 1999;319:1157.
2 Sedgwick P. Confounding in clinical trials. *BMJ* 2012;345:e7951.
3 Sedgwick P. Estimating the population at risk. *BMJ* 2012;345:e6859.
4 Sedgwick P. Odds ratios II. *BMJ* 2010;341:c4971.

Cite this as: *BMJ* 2013;346:f25