ABSTRACT

Case-control studies are observational studies in which cases are subjects who have a characteristic of interest, such as a clinical diagnosis, and controls are (usually) matched subjects who do not have that characteristic. After cases and controls are identified, researchers “look back” to determine what past events (exposures), if any, are significantly associated with caseness. For “looking back,” data may be obtained by clinical history-taking or from medical records such as case files or large electronic health care databases. The data are analyzed using logistic regression, which adjusts for confounding variables and yields an odds ratio and a probability value for the association between the exposure of interest (independent variable) and caseness (dependent variable). Because case-control studies are not randomized controlled studies, cause–effect relationships do not necessarily explain significant associations detected in the regressions; unexplored confounding may be responsible. These concepts are explained with the help of examples.

Keywords: Case-control studies, research design, logistic regression

Case-Control Study: General Description

A case-control study is one in which cases are compared with controls to identify historical exposures that are significantly associated with a current state or, stated in different words, variables that are significantly associated with caseness. In case-control studies, cases are subjects with a particular characteristic. The characteristic that defines caseness may be a clinical diagnosis (e.g., schizophrenia [Sz]), a treatment outcome (e.g., treatment-resistance), a side effect (e.g., tardive dyskinesia), or any other characteristic that is the subject of interest. Controls are subjects who do not have the characteristic that defines caseness. For Sz, controls may be healthy controls; for treatment-resistance, controls would be subjects with the same diagnosis and who are treatment-responsive; for tardive dyskinesia, controls would be subjects who received the same treatment but did not develop this adverse outcome. Controls are commonly selected based on matching with cases for variables such as age, sex, site of recruitment, and other variables. Matching may be 1:1, but when data are drawn from large electronic databases, it is often possible to match five or even 10 controls with each case. In such studies, there may be thousands of cases and tens or even hundreds of thousands of controls.

As an actual example of a case-control study, children with autism spectrum disorder (ASD) may be compared with normally developing children to determine whether a history of maternal antidepressant use during pregnancy is more frequent among cases than among controls; if it is, and if the association remains statistically significant after adjusting for confounding variables, one may speculate that gestational exposure to antidepressants predisposes to autism spectrum disorder. Here, readers may note that there is only one exposure of interest: gestational exposure to antidepressants.

As a hypothetical example of a case-control study, patients with Sz may be compared with healthy controls to...
determine whether a family history of Sz, viral infection during pregnancy, season of birth, obstetric complications during pregnancy, brain insults in early childhood, and other variables are associated with Sz in the sample. Here, readers may note that all the variables listed are exposures of interest and corrections are desirable to protect against the risk of Type 1 statistical error associated with multiple hypothesis testing.

In summary, in case-control studies, there are cases and there are controls that are matched with cases. Researchers then “look back” to ascertain what past events (exposures) are associated with caseness. The exposures of interest may be one or many.

**Analysis of Case-Control Studies**

Case-control studies are analyzed using logistic regression. The dependent variable is the (dichotomous) grouping variable: case vs. control. The independent variables are the exposure(s) of interest plus the confounding variables whose effects must be adjusted for in the regression to understand the unique effect of the exposure variable(s). The logistic regression yields an odds ratio and a statistical significance (P) value for each independent variable; this allows us to understand whether or not the independent variables are significantly associated with caseness, and, if they are, what the effect sizes are, as exemplified by the odds ratios. Readers may note that whether a significant association is a marker of risk or a cause of the risk cannot be determined from an observational study; this was explained in an earlier article.

As a special note, when cases and controls are well matched on many important variables, a procedure known as conditional logistic regression analysis may be employed.

**Characteristics of Case-Control Studies**

How do case-control studies fit into classifications of research design described in an earlier article? Case-control studies are empirical studies that are based on samples, not individual cases or case series. They are cross-sectional because cases and controls are identified and evaluated for caseness, historical exposures, and confounding variables at a single point in time. They are observational; there is no intervention. They are prospective when cases and controls are identified and interviewed in real time, such as in an outpatient department, and retrospective when they are identified in and studied from medical records or electronic health care databases. Strengths and limitations of prospectively vs. retrospectively ascertained data were described in an earlier article.

The nested case-control study is a special situation in which cases and controls are both identified from within a cohort. So, instead of studying the entire cohort, which would be time- and labor-intensive, the researchers study only cases and matched controls within that cohort. To explain with the help of an actual example, Gronich et al. examined the electronic database of the largest health care provider in Israel and identified a cohort of 1,762,164 adults who did not have a diagnosis of Parkinson’s disease (PD). During follow-up, 11,314 patients were newly diagnosed with PD. Each patient (case) was matched with 10 randomly selected controls based on age, sex, ethnicity, and duration of follow-up. Thus, rather than extracting data for 11,314 cases and the rest of the 1,762,164 adults who did not develop PD and who were therefore noncases, the authors carved out a smaller sample of controls from within the cohort. Thus, the final sample of 11,314 cases and 113,140 controls was “nested” within the original cohort; studying this smaller sample took less time and was less labor-intensive than studying the entire cohort.

**Parting Notes**

There are two reasons why, in case-control studies, large samples are desirable, and why many controls may be matched to a single case. One reason is that patients are not randomized to be cases or controls. In such circumstances, as in quasi-controlled studies, there is bound to be confounding. With larger samples, statistical power to adjust for confounding will improve. The other reason is that, in case-control studies, data are usually drawn from medical records or databases. Information extracted from such sources is very unlikely to have been collected and recorded with the expectation of use in future research. So, there are bound to be inaccuracies. When data are blurred (inaccurate), there is statistical noise. When the sample size is large, it becomes easier to see a signal through the noise.

**Declaration of Conflicting Interests**

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author received no financial support for the research, authorship, and/or publication of this article.

**ORCID iD**

Chittaranjan Andrade https://orcid.org/0000-0003-1526-567X

**References**

1. Andrade C. Describing research design. Indian J Psychol Med 2019; 41: 201–202.
2. Andrade C. Simultaneous descriptors of research design. Indian J Psychol Med 2021; 43(6): 83–84.
3. Andrade C. Research design: cohort studies. Indian J Psychol Med 2022; 44(1): (in press).

4. Croen LA, Grether JK, Yoshida CK, et al. Antidepressant use during pregnancy and childhood autism spectrum disorders. Arch Gen Psychiatry 2011; 68(11): 1104–1112.

5. Andrade C. Author’s response to multiple testing and protection against type I error using P value correction: Application in cross-sectional study designs. Indian J Psychol Med 2019; 41(2): 198.

6. Kuo C-L, Duan Y, and Grady J. Unconditional or conditional logistic regression model for age-matched case-control data? Front Public Health 2018; 6: 57.

7. Ernster VL. Nested case-control studies. Prev Med 1994; 23(5): 587–590.

8. Gronich N, Abernethy DR, Auriel E, et al. Beta2-adrenoceptor agonists and antagonists and risk of Parkinson’s disease. Mov Disord 2018; 33(9): 1465–1471.

9. Andrade C. The limitations of quasi-experimental studies, and methods for data analysis when a quasi-experimental research design is unavoidable. Indian J Psychol Med 2021; 43(5): 451–452.