How to Conduct and Write a Case-control Study

Endrik H. Sy, MD, FPAFP[2], Jan Vittorio Quizon, MD, DFM[3] and Noel L. Espallardo, MD, FPAFP

A case control study is a type of observational study. In this study design, participants are selected to participate depending on their outcome status. Cases are participants with outcome of interest whereas controls are participants who do not have the outcome of interest. These studies estimate the odds ratio or the odds between the exposure and health outcome, however they cannot prove causality. Advantages of case control studies include the following: inexpensive, easy to design and implement, and are used to study rare outcomes. Case control studies are prone to certain research bias but can be addressed by the investigator through careful designing and planning. This paper describes the case control studies, their advantages, bias in case control studies and how to address them, and discuss the steps in how to conduct this type of study design.

Key words: Case control, odds ratio, observational study

Definition

Case-control studies are one of the observational study designs that family physicians should be familiar with. An advantage of this study design is that they are relatively quick to perform, economical, easy to design and implement.[1] In this article, we describe case-control study design, how to avoid or address bias and guide the reader on how to write and evaluate the case report manuscripts. A case-control study is a type of observational study design where participants are selected for the study based on their outcome status.[2] A case-control study design is classified as observational study design because the researcher does not control the assignment of participants to different groups unlike experimental studies wherein you want the two groups to be similar as possible. In this study design you will have two groups of participants. Participants with the outcome or disease of interest are referred to as cases while those without are referred to as controls. The researcher then assesses the exposure in both of these groups. This is also different from retrospective cohort studies wherein you group participants according to their exposure and determine association of exposure to the outcome (Figure 1).

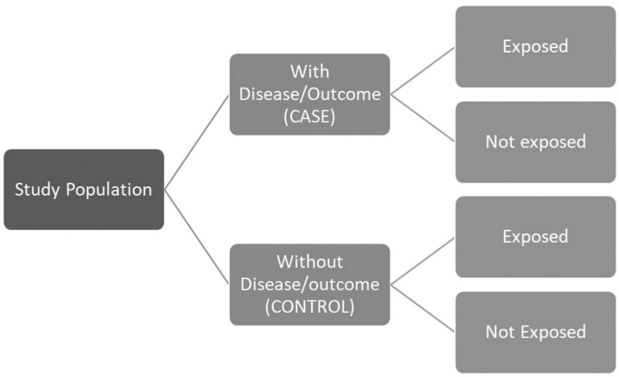

Figure 1. Diagram case control studies.

Case-control studies can be prospective or retrospective depending on the manner of patient recruitment or data collection.[3] If recruitment or data collection is being done as cases develop forward in time it is prospective, but if the cases have already developed in the past and patient recruitment is being done by reviewing existing clinical records then it is retrospective.

Nested case-control study is when case studies are performed within a cohort study. A nested case–control design has a case–control study "nested" within a defined cohort.[4] The cohort may already have been defined by the investigator as part of a formal cohort study, often including banking of specimens, images, and so on, to be analyzed in the future after outcomes occur. Alternatively,

[1]Baguio General Hospital and Medical Center
[2]Veterans Memorial Medical Center

the investigator can design a nested case–control study de novo, in a cohort that is not already defined, in which case defining the cohort will be the first step.

Why Conduct Case-Control Studies

Case control studies have advantages compared to other study designs as they are appropriate for investigating outbreaks, and studying rare diseases or outcomes.[5] This advantage comes from the design itself since we start with participants with known outcomes, hence making it possible to enroll sufficient patients with a rare disease. This study design can be used for investigation of diseases that occur infrequently or that develop years after exposure. It is also useful in the study of exposures that cannot be randomized for logistic or ethical reasons.[6] Case-control studies can be executed rapidly at a relatively low cost.[7] Case control study design allows the researcher to look at multiple risk factors at once. Hence, it is commonly used as one of the first studies to investigate association between exposure and an event or disease.[8]

Case-control study design establishes the cause-and-effect relation between risk factors and health outcomes. It can determine the relative importance of a predictor variable in relation to the presence or absence of the disease. But since the cases are deliberately chosen, it cannot establish the incidence or prevalence of the outcome. Case Control studies can however be used to calculate odds ratios, which usually approximate the relative risk when the outcome is rare. Case-control studies are conducted retrospectively, meaning that exposure and outcome data are collected after the outcome has occurred. This design allows researchers to quickly and efficiently gather information on past exposures and compare them between cases and controls. As a result, case-control studies are well-suited for generating initial hypotheses about potential associations between exposures and outcomes.

Example of Case-Control Studies

The landmark case-control studies of Wynder and Graham, Doll and Hill, and Levin in 1950 on the association between cigarette smoking and lung cancer lead to popularity of the use of this type of study.[9] There are two recent publications of case control studies in The Filipino Family Physician. In 2019, a case control study by Narag-Manzano showed a positive relationship between utilization of anti-hypertensive medications and local health clinic services. An inverse relationship was seen between utilization and educational attainment, other sources of income, nuclear family structure, symptoms of elevated BP, other anti-hypertensive medicines, regular consultation and hospitalization for the past year.[10] A case control study by Zaldriaga et al in 2020, studied the risk of polypharmacy and risk of developing dementia. The study revealed that the risk of developing dementia does not have statistically significant association with the number of medications in the Filipino elderly.[11]

Bias in Case-Control Studies

Case-control studies are prone to have several types of bias, and this can be addressed by controlling or eliminating them during the design phase of the study. Evaluation of potential bias should be done early as there is little we can do once data has been collected. Table 1 summarizes the different types of bias encountered in case-control studies and strategies for how to address or avoid these types of biases.

Steps in Conducting Case-control Study

Step 1: Decide on the Research Question to be Answered.

The very first step in conducting a case-control study is to determine first the research question that will be answered by your study. Formulate a hypothesis and then decide what will be measured and how. Specify the characteristics of the study group and decide how to construct a valid control group by addressing sampling bias. Ideally the cases studied should be a random sample of all the patients with the disease. But when the outcome is rare, this is very difficult. In some cases, the disease may have not been diagnosed or have been misdiagnosed.

Step 2: Identify the Purpose

The definition of the hypothesis, or research question to be answered, is probably more important in this type of study than in any other, since the selection of study participants depends so closely on the question that the investigators want to answer.[12] As participants are selected on the basis of their outcome and will be compared by their exposure, it is important that the participants represent the exposure prevalence of the general population and that the study is not biased by the selection of cases and controls being influenced by their exposure (selection bias). Validate the purpose with a literature review.

Step 3: Formulate the Objectives

Formulating clear and specific objectives is essential in case-control studies to guide the research and ensure that the study's focus is well-defined. The objectives could be framed using the mnemonics SMART, in which the objectives should be Specific, Measurable, Attainable, Relevant, and Time-based. More specifically for case-control studies, the exposure and outcome, study population, and desired outcome should be clearly defined. Formulate the hypothetical results by crafting your dummy tables. Generally, there should be a dummy table dedicated for each of your specific objectives.

Step 4: Identify the Study Population

Selection of cases and control groups is a very important step to consider when conducting case control studies as this is a potential source of bias if the design is not well constructed. Group or frequency

Table 1. Summary of bias in case-control studies and strategies[11]

Bias	Definition	Strategies
Selection Bias	occurs when the procedures that were used to select participants and/or factors distort the results	Exclusion criteria for both cases and controls may be chosen to maximize the probability of participation.
Ascertainment Bias	occurs when there is inaccurate ascertainment of either the disease or exposure of interest. Chart reviews of data are susceptible to this bias because the investigator has no control on how the disease or exposure has been ascertained or recorded..	Random sampling to select participants and blinding to avoid selective recognition
Berkson Bias or also known as Admission Rate Bias	is based on the concept that patients with more than one disease or condition are more likely to be hospitalized than with only one disease. Therefore, this may produce an overestimation of exposed cases in the hospital population	Define controls by assessing if the individuals were admitted to the hospital with a disease of similar severity as the case
Neyman Bias	occurs when the case group is not representative of the intended population. Diseases that are either transient or lead to early death may be underrepresented in the case group.	Count only incident cases, if this it not possible discuss the use of prevalent cases for the outcome of interest
Recall Bias or Memory Bias	Collection of information depends on the memory of participants. Generally, cases or those with disease have better recall than controls about the disease of interest which influences	Use data that was collected in a systematic manner before the onset of disease
Observer Bias	The observer obtains information differs if the individual is a case or a control influencing the level of detail in data collection	Select controls and assure that all observations will be performed in both groups under the same conditions. Train and standardize the procedures and behaviors of the observer
Confounding Bias	present when there is a variable that is associated with the exposure and associated with the outcome but not directly in the causal pathway between the two	assure comparability between the groups of the study; Pair; Stratify

matching ensures that there are the same number of controls as cases within each level of confounder.[12]

Cases may consist of new cases (incidence) that show selected characteristics during a specific time period in a specified population and a particular area. Cases may also consist of existing cases at a point in time (prevalence). With prevalence data, it may be more difficult to link a specific cause with a disease outcome because it is influenced by both the development and duration of disease. For example, suppose researchers were interested in assessing whether an association existed between exercise and the prevalence of arthritis. It may be that exercise patterns before the development of arthritis are much different than after the onset of symptoms; thus, the timing of when the exposure was evaluated could have a large impact on the association. For this reason, whenever possible, incident cases are preferred to prevalent cases in case-control studies.[13] Sources for cases can come from records from public health clinics, physician offices, health maintenance organizations, hospitals, and industrial and government sources. Cases should be representative of all persons with the disease. In some situations, all persons with the disease may be included in the study. It is more common, however, that cases come from sampled data.

In order for the sampled data to reflect the population of interest, random selection is required. An adequately large random selection of cases from a population of interest ensures that the results of the study can be appropriately generalized. In some situations, researchers may use restriction, which involves limiting subjects in a study to those with certain characteristics, such as Black males in Atlanta aged 40–59 years, to reduce potential biases and increase feasibility. Although restriction may limit generalization, it may be necessary to ensure a valid study. Conducting a valid study with definitive results should always be the primary goal of any epidemiologist.

Controls should be chosen who are similar in many ways to the cases. Factors that are chosen (like age, sex and time of hospitalization) to define how the control group is similar to the case group is what we call the Matching Criteria.[3] The disease / outcome of interest should be defined in specific, unambiguous terms by operational diagnostic criteria. Similar diagnostic procedures and criteria should have been used between the two groups.[5] Matching controls to cases will mitigate the effect of confounders. A confounding variable is one which is associated with the exposure and is a cause of the outcome.[3]

To better ensure that a case-control study is valid and reliable, the control subjects should look like the case subjects except for not having the disease. This means that controls need to be selected from the same population from which the cases were drawn. An epidemiologic assumption is that controls are representative of the general population in terms of probability of exposure and have the same possibility of being selected or exposed as the cases. Controls drawn from a population of the same area or populace of the cases should reflect the same gender, age, and other significant factors. Controls from the general population are assumed to be normal, to be healthy, and to reflect the well population from the area. Sampling of controls from a general population for large studies is an expensive endeavor and thus is not always realistic or possible. Controls are typically drawn from the same hospital or general population as the cases. They may also be drawn from the family, friends, or relatives of the cases.[13]

Step 5: Sample Size Computation

As with other types of observational studies, the researcher should use an appropriate sample size to ensure reliability of the study findings.[1] Sample size calculation can be done using Epi Info, a statistical program developed by the Centers for Disease Control and Prevention available via the link: https://www.cdc.gov/epiinfo/index.html. The confidence level used for most researchers is set at 95% however some researchers want to be more confident and choose a 99% confidence interval. A confidence level set 95% is a safe choice for most research. The general rule of thumb is that the larger the sample size the greater the precision or power for a study to detect an effect of a given size.[1]

Step 6: Develop the Data Collection Method

The next step is to develop the data collection form. Once the case and control group has been defined. The next step is to decide on how to collect the data. It is important to take note that the data must be collected the same way from both groups.[4]

Step 7: Collect and Analyze the Data

Data collection methods are important, because how the information collected is used and what explanations it can generate are determined by the methodology and analytical approach applied by the researcher.[14] Analysis of case control studies are usually done using the 2x2 table. Groups are classified as exposed or unexposed depending on whether or not a risk factor under study is present. Compared to Randomized control trials, the risk of developing the outcome event in the exposed group is directly compared to the risk in the non-exposed group, therefore producing "risk ratio". In case-control studies, however, a true risk or risk ratio cannot be calculated since the research does not begin with groups of exposed and non-exposed patients. As a substitute for a risk ratio, the researcher may compare the likelihood (or odds) of exposure and non-exposure group.[5] The odds ratio serves as a good approximation of the risk ratio.

Table 2. 2x2 table for case control studies.

	With Disease (Case)	Without Disease (Control)	Total
Exposed	a	b	a+b
Not exposed	c	d	c+d
Total	a+c	b+d	N

Odds ratio (OR) can be computed with the following formula:
OR = (a/b)/ (c/d) = ad/bc for outcome
OR = (a/c)/ (b/d) = ad/bc for exposure

The odds ratio measures the strength of the association between a risk factor (exposure) and disease. Odds ratios are used to compare the relative odds of an outcome to occur given exposure to variable to interest.[15] The odds ratio may be used to determine whether particular exposure is a risk factor for the outcome of interest.[5]

- **OR = 1**, means the odds for the outcome are the same for both exposed and unexposed
- **OR > 1**, the odds of the outcomes is higher among the exposed group
- **OR < 1**, indicates reduced odds of the outcome with exposure to the risk factor.

Confidence interval should also be computed for each odds ratio. If the confidence interval that includes 1.0 means that the association between the exposure and outcome could have been found by chance alone and it is not statistically significant.[4] Statistical significance can be determined by the use of the X^2 test or by placing a 95% confidence interval of around the odds ratio.[5]

Step 8: Writing Case-Control Studies using STROBE Guidelines

An important aspect of research is to share the results of your studies by writing and publishing your research paper. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines can be used as a guide when writing the final report for case-control studies. These are recommendations that consist of a checklist of 22 items, which relate to the title, abstract, introduction, methods, results and discussion sections of articles.[16] For the full checklist and article on the STROBE guideline, you may access the link below: https://www.strobe-statement.org/

Ethical issues

Conducting case-control studies raises several ethical considerations and challenges. Ethical issues that researchers should be aware of and address when conducting case-control studies include informed consent, privacy and confidentiality, bias and selection of controls, equity and fairness, risk of harm, data sharing and secondary use. It is important for researchers to carefully consider and address these ethical issues throughout the entire research process, from study design and participant recruitment to data collection, analysis, and dissemination of results. Adhering to ethical principles helps protect the rights and well-being of study participants and ensures the integrity and validity of the research.

References

1. Dey T, et al. A practical overview of case-sontrol studies in clinical practice. Chest 2020; 158(1S): S57-S64. doi:10.1016/j.chest.2020.03.009
2. Setia MS. Methodology Series Module 2: Case-control Studies. Indian J Dermatol 2016 Mar-Apr;61(2):146-51. doi: 10.4103/0019-5154.177773.
3. Rothman KJ. Epidemiology: An introduction (2nd ed.) Oxford University Press, Inc. 2012
4. Hulley SB, Cummings SR, Browner WS, Grady DG & Newman TB. Designing Clinical Research (4th ed.) 2013; 97-110. Lippincott Williams & Wilkins, a Wolters Kluwer Business.
5. Lewallen S and Courtright P. Epidemiology in practice: case-control studies. Community Eyehealth 1998; 11(28): 57-8.
6. Hayden GF, Kramer MS & Horwitz RI. The case-control study. A practical review forthe clinician. JAMA 1982; 247(3): 326–31.
7. Sutton-Tyrrell K. Assessing bias in case-control studies. Proper selection of cases and controls. Stroke 1991; 22(7): 938–42. https://doi.org/10.1161/01.str.22.7.938
8. Tenny S, Kerndt CC & Hoffman MR. Case control studies. In StatPearls. StatPearls Publishing 2023.
9. Breslow N. Case control studies. Handbook of Epidemiology 2005; 288-319.
10. Narag-Manzano J. Patient-related factors associated with the utilization of anti-hypertensive medications from the PhilPEN Program Implemented in the Five Pilot Barangays of the AMIGA-UP Community Health and Development Program in Amadeo, Cavite: A case-control study. The Filipino Family Physician 2019; 57(2).
11. Zadriaga, et al. Association between polypharmacy and dementia among Filipino elderly in a tertiary care hospital in Metro Manila, Philippines: A case-control study. The Filipino Family Physician 2020; 58(2).
12. Souza FO, Miranda SS, Valente ALP, de Araújo TM & Smolen JR. Bias in case-control studies: is it possible to avoid? J Epidemiol Infect Control 2016; 6(2): 92-6. https://doi.org/10.17058/reci.v6i2.6370;
13. Bailey L, Vardulaki K, Langham J, Chandramohan D & Newman TB. Introduction to Epidemiology 2005; 74-85. Open University Press.
14. Merrill RM. Introduction to Epidemiology (7th ed., 2017; pp. 142-9). Jones & Bartlett Learning.
15. Szumilas M. Explaining odds ratios. J Can Acad Child Adolesc Psychiatr 2010 Aug;19(3):227-9. Erratum in: J Can Acad Child Adolesc Psychiatr 2015 Winter;24(1):58.
16. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M. STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. PLoS Med 2007 Oct 16;4(10):e297. doi:10.1371/journal.pmed.0040297.