Original Article

Selecting the appropriate study design: Case–control and cohort study designs

Aamir Omair

Department of Medical Education, Research Unit, King Saud Bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia

ABSTRACT

This article discusses the observational analytic study designs, i.e., case–control and cohort studies. These two study designs are useful for testing a hypothesis to determine the association between a risk factor and a disease. The analysis for both the studies is based on the conventional 2 × 2 table with the disease status in columns and the risk factor status in rows. The case–control studies start from the disease status and compare the exposure to the risk factor(s) between the diseased (cases) and the not diseased (controls) groups. The odds ratio is determined to compare the proportion of exposed persons in the two groups. The cohort studies start from the exposure to the risk factor status and compare the incidence of the disease in the exposed and not exposed groups. The relative risk compares the incidence between the two groups. The 95% confidence interval is estimated for both studies to determine an actual association between the risk factor and the disease. The strengths and limitations of the two study designs differ based on the direction of the two designs. The case–control study goes backward from the disease status so is more useful for rare diseases and for evaluating multiple risk factors, but it cannot determine causality, and there are chances of recall bias affecting the results of the study. The cohort studies are generally prospective in design from the exposure status and can determine the causal association between the risk factor and the disease. However, the cohort studies are more expensive and require a longer time as well as a larger sample size; the loss to follow-up and misclassification biases can affect the results of the cohort studies.

Keywords: 2 × 2 table, case–control, cohort, odds ratio, relative risk

INTRODUCTION

This is the second part of the article on epidemiological study designs. The previous article discussed the different types of 'descriptive studies' including the case report/case series, correlational and cross-sectional study designs.[1] This article shall describe observational analytic study designs, which includes case–control and cohort studies; clinical trials will be discussed in detail in the next article. The first two study designs are part of the 'observational' group of study designs along with the rest of the descriptive studies. Clinical trials are classified as 'interventional studies', which are also called as 'experimental' study designs. In the observational studies, the researcher classifies the study subjects into groups of diseased/not diseased or exposed/not exposed to risk factors, based on observation of their natural state. In the interventional study designs, the subjects are distributed into the intervention or the non-intervention/standard treatment group by the investigator themselves.[2]

2 × 2 CONTINGENCY TABLE

Descriptive studies are based on a single sample and are useful for identifying risk factors and determining the prevalence as well as generating hypotheses for the association between the risk factor and the

Address for correspondence:
Dr. Aamir Omair, Department of Medical Education, Research Unit, King Saud Bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia.
E-mail: amir.omair@gmail.com

Access this article online

Quick Response Code:

Website: www.thejhs.org

DOI: 10.4103/1658-600X.173842

How to cite this article: Omair A. Selecting the appropriate study design: Case–control and cohort study designs. J Health Spec 2016;4:37-41.

disease/outcome. However, the associations identified in descriptive studies cannot be definitively attributed to the identified risk factor(s) as there is no comparison group. The analytical study designs, on the other hand, have a comparison group and consequently are useful for testing the hypothesis to determine the association between the risk factor and the disease/outcome of interest.[3] The conventional way of analysing case–control and cohort studies is through a 2 × 2 table [Table 1]. The disease status is given at the top of the column under the headings of 'diseased/not diseased' while the exposure to the risk factor is given in rows as 'exposed/not exposed'. The four cells in the 2 × 2 table are labelled as 'a-b-c-d'; cell 'a' represents the subjects who are exposed to the risk factor and are diseased while cell 'b' is the group that is exposed but not diseased. Similarly, cell 'c' includes the subjects that are not exposed but diseased, while cell 'd' is the group that is not exposed and not diseased.[4]

The observational analytic study designs basically test the hypothesis to see if there is a relationship between the disease and the risk factor. If the hypothesis is true, there should be a greater proportion of subjects in cells 'a' and 'd'. At the same time, there will always be subjects who are exposed to the risk factor but not diseased (cell 'b') as well as those that are not exposed to the risk factor but are diseased (cell 'c').[5] As shown in Table 2, there are people who smoke and have heart disease (cell 'a') but there are also many smokers who do not have heart disease (cell 'b'). Similarly, there are also many non-smokers who have heart disease (cell 'c') but the majority of the non-smokers are not diseased (cell 'd'). The explanation for this is that most of the non-communicable diseases are multifactorial in nature while even for the communicable diseases, exposure to the organism does not necessarily lead to the development of the disease. This 2 × 2 table forms the basis for determining the epidemiological measures of association of the odds ratio (OR) and the relative risk (RR) for the case–control and cohort studies, respectively.[4]

CASE–CONTROL STUDY

This is the simplest analytical study design and is based on comparing the group of patients who are diseased (cases) with a similar comparison group and who are not diseased (controls). The direction of the study is from the disease status to the exposure status [Table 3] and is useful for determining the risk factors that are associated with a disease. The two groups are compared with regards to the proportion of exposure to the risk factors(s) of interest in each group.[6]

Table 1: The elements of a simple 2×2 table for analysing epidemiological studies

	Diseased	Not diseased	Row total
Exposed	a Exposed and diseased	b Exposed and not diseased	a + b Total exposed
Not exposed	c Not exposed and diseased	d Not exposed and not diseased	c + d Total not exposed
Column total	a + c Total diseased	b + d Total not diseased	

Table 2: Example of 2×2 table for association between smoking and heart disease

	Heart disease		Row total
	Yes	No	
Smoker	a Smoker and heart disease	b Smoker and no heart disease	a + b Total smokers
Nonsmoker	c Nonsmoker and heart disease	d Nonsmoker and no heart disease	c + d Total nonsmokers
Column total	a + c Total heart disease	b + d Total no heart disease	

Table 3: The direction and evaluation of a case-control study design

	Diseased	Not diseased	Row total
Exposed	a Exposed and diseased	b Exposed and not diseased	a + b Total exposed
Not exposed	c Not exposed and diseased	d Not exposed and not diseased	c + d Total not exposed
Column total	a + c Total diseased	b + d Total not diseased	

OR calculation: $OR = \frac{a \times d}{b \times c}$

The OR is used for comparing the proportion of exposure between the two groups of cases and controls. The OR is determined by comparing the ratio of exposed with the not exposed in the diseased group with the ratio of exposed with the not exposed in the controls. This is given by the following formula: (a × d)/(b × c) [Table 3].[7] If the ratio of exposed to not exposed is similar between the diseased and not diseased groups, the OR will be close to '1'. The greater the difference in the exposure between the two groups, the higher the value of the OR. If the OR is significantly <'1', this means that the factor under consideration is actually a 'protective' factor, i.e., people who are diseased are less likely to have this factor as compared to people who are not diseased.[8] The OR alone is not sufficient to determine the association and the 95% confidence interval (95% CI) is also reported along with the OR.[7]

The CI gives a lower and an upper limit of the expected values for the OR based on the results of the study and the sample size. If both the CI values are above '1', it indicates that there is a positive association between the disease and the risk factor. On the other hand, if both the values are <'1', this shows a 'negative' association and the variable is considered a protective factor, i.e., the cases are less likely to have the factor as compared to the controls. If the 95% CI 'includes 1', i.e., the lower value is >'1' and the upper value is more than 1' this indicates 'no association' between the outcome and the risk factor under study.[7]

The case–control study design can test the hypotheses for determining the association between a disease and a risk factor. This study design can be used to test for several risk factors for a single disease and is especially useful for rare diseases. It requires less time and is less expensive to conduct as compared to the other analytical study designs. The main disadvantage of the case–control study designs is that they cannot determine if the risk factor causes the disease since it cannot be determined whether most of the risk factors occurred before the disease.[9] Hence, this study design can only determine if there is an association between the disease and the risk factor. This is important to consider when interpreting the results of case–control studies that should not be stated as to imply that there is causative relationship between the risk factor and the disease, e.g. '30% of the exposed group developed the disease' or 'exposed group were 'x' times more likely to develop the disease'. Instead, the appropriate statement for a case–control study should be that 'there is an association between the disease and the risk factor' or 'cases are 2 times more likely to be exposed to the risk factor as compared to controls'.[10]

Some biases that may affect the results of case–control studies include sampling bias and recall bias.[11] The sampling bias may occur if the cases and controls are taken from different subgroups of the population. In this case, the difference in the exposure may be due to some other inherent differences between the groups rather than the risk factor being studied. The recall bias is most applicable to case–control studies where the subjects are asked to answer questions about exposure to risk factors in the past. It is likely that people who are diseased (cases) remember their exposure to the risk factor more accurately as compared to the controls. This may result in the OR value being higher than the actual value. It is important to note that case–control study designs cannot determine neither the prevalence or incidence of the disease nor the risk factors as the subjects are generally collected by purposive sampling.[9]

A useful thing to consider is that the descriptive case series study can be converted into an analytical case–control study with the addition of a similar group of 'controls'.[12] The case series study consists of a group of patients having the same disease. If the number of cases is sufficient, a group of controls can be selected who are matched for certain criteria such as gender, age and group (which are not part of the variables being considered risk factors in the study). This makes the case–control study a practical and convenient study design to be applied in hospital or clinic settings where both diseased (cases) and non-diseased (controls) can be easily accessed.[12]

COHORT STUDIES

A 'cohort' is a group of persons sharing the same characteristics, and in terms of epidemiological study designs, this refers to 'being exposed to the same risk factor'.[13] This study design compares group of subjects who are exposed to a certain risk factor with a similar comparison group who are not exposed to the risk factor. The two groups are longitudinally followed-up over time to observe the occurrence of the outcome of interest in each group. The direction of the study is from the exposure status to the outcome status [Table 4] and is useful for comparing the incidence of disease in the two groups.[14] It is important to remember that in the cohort study, all the subjects in the exposed and non-exposed groups are 'not diseased' at the start of the study.

The RR is the epidemiological measure of association that is applied for the analysis of the results in cohort studies.[15] It compares the incidence of the disease in the exposed group with the incidence in the non-exposed group, hence the name RR or risk ratio. If the incidence in the two groups is equal, the value for RR will be '1' but if the value is greater than '1', this indicates a positive 'causal' relationship between the risk factor and the disease. In some cases, when the value of RR is <'1',

Table 4: The direction and evaluation of a cohort study design

	Diseased	Not diseased	Row total
Exposed	a Exposed and diseased	b Exposed and not diseased	a + b Total exposed
Not exposed	c Not exposed and diseased	d Not exposed and not diseased	c + d Total not exposed
Column total	a + c Total diseased	b + d Total not diseased	

$$\text{Relative risk} = \frac{\text{Incidence in exposed}}{\text{Incidence in not exposed}} = \frac{Ie}{Ine} = \frac{a/(a+b)}{c/(c+d)} = RR$$

this may indicate a protective relationship between the variable under study and the disease.[15] Similar to the OR, the RR alone is not sufficient to determine the actual relationship and must be accompanied with the 95% CI. The statistical interpretation based on the 95% CI is the same as for the OR.[7]

The cohort studies can be used to compare the incidence of more than one outcome for a single risk factor between the exposed and the not exposed groups.[11] This study design is most appropriate where rare exposures are being studied but can be applied for common risk factors such as smoking as well. The main advantage is that all the subjects are disease-free at the beginning of the study, so causality of the risk factor can be determined since the exposure precedes the outcome.[16] The main disadvantage of the cohort studies is due to the long follow-up period for some outcomes. This requires a relatively larger sample size depending upon the incidence rate of the outcome and also the expected loss to follow-up rate due to subjects dropping out from any or both of the groups. The cohort study is not suitable for studying rare diseases or outcomes since this will require a very large sample size to get sufficient number of outcomes for analysing the data.[14]

The cohort studies are generally prospective studies since it is important to establish that the exposure occurred first. However, for certain exposures such as blood group, genetic markers or other factors that clearly occurred earlier, it may be possible for conducting retrospective cohort studies.[16] This type of cohort study may especially be useful for outcomes that take a long time to develop after the exposure. The exposure status is established in the past from medical records or medical history, and the outcome status is determined at the time of the study and after follow-up for a period if required. The biases that may affect the results of the cohort studies include loss to follow-up bias, especially if the loss to follow-up is more in one group as compared to the other group. The other bias is related to the selection bias - the two groups must be comparable to each other except for the exposure status.[13] It is also important to screen both the exposed and non-exposed groups at the start of the study using the appropriate inclusion/exclusion criteria to make sure that there is no misclassification bias. This may also be due to the fact that during the follow-up period, the exposure status of the subjects may change leading to inappropriate analysis of the results.[14]

SUMMARY

The two observational analytic study designs, i.e., the case–control and the cohort studies, play an important part in testing the hypotheses for determining the association between exposure to risk factors and disease/outcome of interest. However, the two studies are methodologically different in that the case–control study starts from the outcome and goes 'back' to determine the exposure to the risk factor, while the cohort study starts from the exposure status and goes 'forward' to determine the incidence of outcome in the groups to be compared. In this way, the two study designs are more suitable for different types of outcomes and risk factors, and each one has its own strengths and limitations as shown in Table 5. Both study designs are observational studies, so the chance of confounding due to factors inherent to the group classification is still present. However, these two still constitute the most common study designs that are used in the epidemiological field along with the cross-sectional studies and the clinical trials.

Table 5: Strengths and limitations of the case-control and cohort study designs

Strengths	Limitations
Case-control studies	
Require less time and less expensive	Cannot determine incidence or prevalence
Require smaller sample size	Cannot determine causality
Can evaluate multiple exposures	Not useful for rare exposures
Useful for rare diseases/outcomes	Recall bias
	Selection bias
Cohort studies	
Can determine incidence	Requires more time and more expensive
Can determine causality of exposure	Requires larger sample size
Can evaluate multiple outcomes	Not useful for rare diseases/outcomes
Useful for rare exposures	Loss to follow-up bias
	Misclassification bias

Acknowledgment

This article is the fourth one in the series on articles on Research Methods. I would like to dedicate this article series to Professor James Ware, who was the main motivator behind these articles. His continuous support and valuable feedback on each article were instrumental in improving the quality of the articles. Professor James Ware's lively personality will be dearly missed by everyone who had the opportunity of working with him. He had an exuberance which he transferred to the people around him. His loss will be a difficult one to compensate in the field of medical education and research. We should strive to carry on his work in the same manner by providing support and guidance to anyone that needs assistance in the things that we learned with Professor James Ware.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Omair A. Selecting the appropriate study design for your research: Descriptive study designs. J Health Spec 2015;3:153-6.
2. Thiese MS. Observational and interventional study design types; an overview. Biochem Med (Zagreb) 2014;24:199-210.
3. Grimes DA, Schulz KF. Descriptive studies: What they can and cannot do. Lancet 2002;359:145-9.
4. Ingersoll GM. Analysis of 2 × 2 contingency tables in educational research and evaluation. Int J Res Educ 2010;27:1-14. Available from: http://www.cedu.uaeu.ac.ae/journal/issue27/ch5_27en.pdf. [Last accessed on 2015 Nov 18].
5. Sauerbrei W, Blettner M. Interpreting results in 2 × 2 tables: Part 9 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009;106:795-800.
6. Kanchanaraksa S. Case-Control Studies. Johns Hopkins University. Available from: http://www.ocw.jhsph.edu/courses/FundEpiII/PDFs/Lecture14.pdf. [Last accessed on 2015 Nov 18].
7. Szumilas M. Explaining odds ratios. J Can Acad Child Adolesc Psychiatry 2010;19:227-9.
8. Spitalnic S. Risk assessment II: Odds ratio. Hosp Physician 2006;42:23-6.
9. Lewallen S, Courtright P. Epidemiology in practice: Case-control studies. Community Eye Health 1998;11:57-8.
10. International Agency for Research on Cancer. Case-control studies. In: Cancer Epidemiology. Ch. 9. World Health Organization. Available from: http://www.iarc.fr/en/publications/pdfs-online/epi/cancerepi/CancerEpi-9.pdf. [Last accessed on 2015 Nov 18].
11. Mann CJ. Observational research methods. Research design II: Cohort, cross sectional, and case-control studies. Emerg Med J 2003;20:54-60.
12. Paneth N, Susser E, Susser M. Origins and early development of the case-control study. In: Morabia A, editor. A History of Epidemiologic Methods and Concepts. Basel, Switzerland: Birkhluser Verlag; 2004. p. 291-311. Available from: http://www.tc.umn.edu/~alonso/Paneth_case-control.pdf. [Last accessed on 2015 Nov 18].
13. Song JW, Chung KC. Observational studies: Cohort and case-control studies. Plast Reconstr Surg 2010;126:2234-42.
14. HealthKnowledge. Introduction to Study Designs – Cohort Studies; 2011. Available from: http://www.healthknowledge.org.uk/e-learning/epidemiology/practitioners/introduction-study-design-cs. [Last accessed on 2015 Nov 18].
15. Gay J. Clinical Epidemiology and Evidence-based Medicine Glossary: Terminology Specific to Epidemiology; 2005. Available from: http://www.people.vetmed.wsu.edu/jmgay/courses/GlossEpiTerminology.htm. [Last accessed on 2015 Nov 18].
16. Boston University School of Public Health. Advantages and Disadvantages of Cohort Studies; 2015. Available from: http://www.sphweb.bumc.bu.edu/otlt/MPH-Modules/EP/EP713_CohortStudies/EP713_CohortStudies5.html. [Last accessed on 2015 Nov 18].