Design and data analysis case-controlled study in clinical research

Sanjeev V. Thomas, Karthik Suresh[1], Geetha Suresh[2]

Department of Neurology, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, Kerala, India, [1]Department of Pulmonary and Critical Care Medicine, Johns Hopkins University School of Medicine, [2]Department of Justice Administration, University of Louisville, Louiseville, USA

Abstract

Clinicians during their training period and practice are often called upon to conduct studies to explore the association between certain exposures and disease states or interventions and outcomes. More often they need to interpret the results of research data published in the medical literature. Case-control studies are one of the most frequently used study designs for these purposes. This paper explains basic features of case control studies, rationality behind applying case control design with appropriate examples and limitations of this design. Analysis of sensitivity and specificity along with template to calculate various ratios are explained with user friendly tables and calculations in this article. The interpretation of some of the laboratory results requires sound knowledge of the various risk ratios and positive or negative predictive values for correct identification for unbiased analysis. A major advantage of case-control study is that they are small and retrospective and so they are economical than cohort studies and randomized controlled trials.

Key Words

Analysis, case-control study, design

For correspondence:
Dr. Sanjeev V. Thomas, Department of Neurology, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum - 695 011, Kerala, India.
E-mail: sanjeev.v.thomas@gmail.com

Ann Indian Acad Neurol 2013;16:483-7

Introduction

Clinicians think of case-control study when they want to ascertain association between one clinical condition and an exposure or when a researcher wants to compare patients with disease exposed to the risk factors to non-exposed control group. In other words, case-control study compares subjects who have disease or outcome (cases) with subjects who do not have the disease or outcome (controls). Historically, case control studies came into fashion in the early 20th century, when great interest arose in the role of environmental factors (such as pipe smoke) in the pathogenesis of disease. In the 1950s, case control studies were used to link cigarette smoke and lung cancer. Case-control studies look back in time to compare "what happened" in each group to determine the relationship between the risk factor and disease. The case-control study has important advantages, including cost and ease of deployment.

Access this article online	
Quick Response Code:	Website: www.annalsofian.org
	DOI: 10.4103/0972-2327.120429

However, it is important to note that a positive relationship between exposure and disease does not imply causality.

At the center of the case-control study is a collection of cases. [Figure 1] This explains why this type of study is often used to study rare diseases, where the prevalence of the disease may not be high enough to permit for a cohort study. A cohort study identifies patients with and without an exposure and then "looks forward" to see whether or not greater numbers of patients with an exposure develop disease.

For instance, Yang *et al*. studied antiepileptic drug (AED) associated rashes in Asians in a case-control study.[1] They collected cases of confirmed anti-epileptic induced severe cutaneous reactions (such as Stevens Johnson syndrome) and then, using appropriate controls, analyzed various exposures (including type of [AED] used) to look for risk factors to developing AED induced skin disease.

Controls

Choosing controls is very important aspect of case-control study design. The investigator must weigh the need for the controls to be relevant against the tendency to over match controls such that potential differences may become muted. In general, one may consider three populations: Cases, the

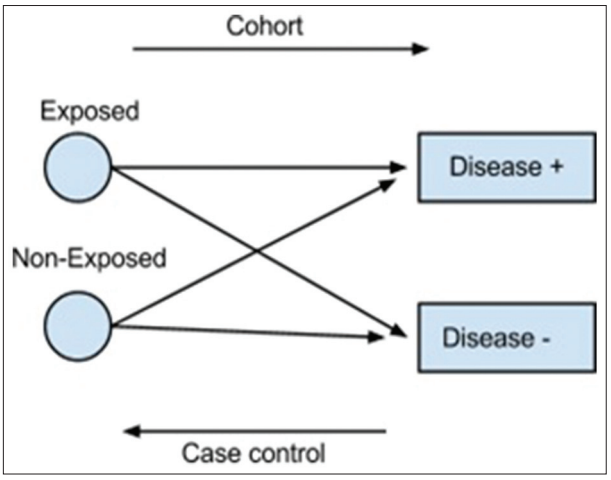

Figure 1: Comparison of cohort and case control studies

relevant control population and the population at large. For the study above, the cases include patients with AED skin disease. In this case, the relevant control population is a group of Asian patients without skin disease. It is important for controls to be relevant: In the anti-epileptic study, it would not be appropriate to choose a population across ethnicities since one of the premises of the paper revolves around particularly susceptibility to AED drug rashes in Asian populations.

One popular method of choosing controls is to choose patients from a geographic population at large. In studying the relationship between non-steroidal anti-inflammatory drugs and Parkinson's disease (PD), Wahner *et al*. chose a control population from several rural California counties.[2] There are other methods of choosing controls (using patients without disease admitted to the hospital during the time of study, neighbors of disease positive cases, using mail routes to identify disease negative cases). However, one must be careful not to introduce bias into control selection. For instance, a study that enrolls cases from a clinic population should not use a hospital population as control. Studies looking at geography specific population (e.g., Neurocysticercosis in India) cannot use controls from large studies done in other populations (registries of patients from countries where disease prevalence may be drastically different than in India). In general, geographic clustering is probably the easiest way to choose controls for case-control studies.

Two popular ways of choosing controls include hospitalized patients and patients from the general population. Choosing hospitalized, disease negative patients offers several advantages, including good rates of response (patients admitted to the hospital are generally already being examined and evaluated and often tend to be available to further questioning for a study, compared with the general population, where rates of response may be much lower) and possibly less amnestic bias (patients who are already in the hospital are, by default, being asked to remember details of their presenting illnesses and as such, may more reliably remember details of exposures). However, using hospitalized patients has one large disadvantage; these patients have higher severity of disease since they required hospitalization in the first place. In addition, patients may be hospitalized for disease processes that may share features with diseases under study, thus confounding results.

Using a general population offers the advantage of being a true control group, random in its choosing and without any common features that may confound associations. However, disadvantages include poor response rates and biasing based on geography. Administering long histories and questions regarding exposures are often hard to accomplish in the general population due to the number of people willing (or rather, not willing) to undergo testing. In addition, choosing cases from the general population from particular geographic areas may bias the population toward certain characteristics (such as a socio-economic status) of that geographic population. Consider a study that uses cases from a referral clinic population that draws patients from across socio-economic strata. Using a control group selected from a population from a very affluent or very impoverished area may be problematic unless the socio-economic status is included in the final analysis.

Cases

In case-controls studies, cases are usually available before controls. When studying specific diseases, cases are often collected from specialty clinics that see large numbers of patients with a specific disease. Consider for example, the study by Garwood *et al*.[3] which looked at patients with established PD and looked for associations between prior amphetamine use and subsequent development various neurologic disorders. Patients in this study were chosen from specialty clinics that see large numbers of patients with certain neurologic disorders. Case definitions are very important when planning to choose cases. For instance, in a hypothetical study aiming to study cases of peripheral neuropathy, will all patients who carry a diagnosis of peripheral neuropathy be included? Or, will only patients with definite electromyography evidence of neuropathy be included? If a disease process with known histopathology is being studied, will tissue diagnosis be required for all cases? More stringent case definitions that require multiple pieces of data to be present may limit the number of cases that can be used in the study. Less stringent criteria (for instance, counting all patients with the diagnosis of "peripheral neuropathy" listed in the chart) may inadvertently choose a group of cases that are too heterogeneous.

The disease history status of the chosen cases must also be decided. Will the cases being chosen have newly diagnosed disease, or will cases of ongoing/longstanding disease also be included? Will decedent cases be included? This is important when looking at exposures in the following fashion: Consider exposure X that is associated with disease Y. Suppose that exposure X negatively affects disease Y such that patients that are X + have more severe disease. Now, a case-control study that used only patients with long-standing or ongoing disease might miss a potential association between X and Y because X + patients, due to their more aggressive course of disease, are no longer alive and therefore were not included in the analysis. If this particular confounding effect is of concern, it can be circumvented by using incident cases only.

Selection bias occurs when the exposure of interest results in more careful screening of a population, thus mimicking an association. The classic example of this phenomenon was noted in the 70s, when certain studies noted a relationship

between estrogen use and endometrial cancer. However, on close analysis, it was noted that patients who used estrogen were more likely to experience vaginal bleeding, which in turn is often a cause for close examination by physicians to rule out endometrial cancer. This is often seen with certain drug exposures as well. A drug may produce various symptoms, which lead to closer physician evaluation, thus leading to more disease positive cases. Thus, when analyzed in a retrospective fashion, more of the cases may have a particular exposure only insofar as that particular exposure led to evaluations that resulted in a diagnosis, but without any direct association or causality between the exposure and disease.

Exposures

One advantage of case-control studies is the ability to study multiple exposures and other risk factors within one study. In addition, the "exposure" being studied can be biochemical in nature. Consider the study, which looked at a genetic variant of a kinase enzyme as a risk factor for development of Alzheimer's disease.[4] Compare this with the study mentioned earlier by Garwood *et al*.,[3] where exposure data was collected by surveys and questionnaires. In this study, the authors drew blood work on cases and controls in order to assess their polymorphism status. Indeed, more than one exposure can be assessed in the same study and with planning, a researcher may look at several variables, including biochemical ones, in single case-control study.

Matching

Matching is one of three ways (along with exclusion and statistical adjustment) to adjust for differences. Matching attempts to make sure that the control group is sufficiently similar to the cases group, with respects to variables such as age, sex, etc., Cases and controls should not be matched on variables that will be analyzed for possible associations to disease. Not only should exposure variables not be included, but neither should variables that are closely related to these variables. Lastly, overmatching should be avoided. If the control group is too similar to the cases group, the study may fail to detect the difference even if one exists. In addition, adding matching categories increases expense of the study.

Analysis

One measure of association derived from case control studies are sensitivity and specificity ratios. These measures are important to a researcher, to understand the correct classification. A good understanding of sensitivity and specificity is essential to understand receiver operating characteristic curve and in distinguishing correct classification of positive exposure and disease with negative exposure and no disease. Table 1 explains a hypothetical example and method of calculation of specificity and sensitivity analysis.

Interpretation of sensitivity, specificity and predictive values

Sensitivity and specificity are statistical measures of the performance of a two by two classification of cases and controls (sick or healthy) against positives and negatives (exposed or non-exposed).[5] Sensitivity measures or identifies the proportion of actual positives identified as the percentage of sick people who are correctly identified as sick. Specificity measures or identifies the proportion of negatives identified as the percentage of healthy people who are correctly identified as healthy. Theoretically, optimum prediction aims at 100% sensitivity and specificity with a minimum of margin of error. Table 1 also shows false positive rate, which is referred to as Type I error commonly stated as α "Alpha" is calculated using the following formula: 100 − specificity, which is equal to 100 − 90.80 = 9.20% for Table 1 example. Type 1 error is also known as false positive error is referred to as a false alarm, indicates that a condition is present when it is actually not present. In the above mentioned example, a false positive error indicates the percent falsely identified healthy as sick. The reason why we want Type 1 error to be as minimum as possible is because healthy should not get treatment.

The false negative rate, which is referred to as Type II error commonly stated as β "Beta" is calculated using the following formula: 100 − sensitivity which is equal to 100 − 73.30 = 26.70% for Table 1 example. Type II error is also known as false negative error indicates that a condition is not present when it should have been present. In the above mentioned example, a false negative error indicates percent falsely identified sick as healthy. A Type 1 error unnecessarily treats a healthy, which in turn increases the budget and Type II error would risk the sick, which would act against study objectives. A researcher wants to minimize both errors, which not a simple issue because an effort to decrease one type of error increases the other type of error. The only way to minimize both type of error statistically is by increasing sample size, which may be difficult sometimes not feasible or expensive. If the sample size is too low it lacks precision and it is too large, time and resources will be wasted. Hence, the question is what should be the sample size so that the study has the power to generalize the result obtained from the study. The researcher has to decide whether, the study has enough power to make a judgment of the population from their sample. The researcher has to decide this issue in the process of designing an experiment, how large a sample is needed to enable reliable judgment.

Statistical power is same as sensitivity (73.30%). In this example, large number of false positives and few false negatives indicate the test conducted alone is not the best test to confirm the disease. Higher statistical power increase statistical significance by reducing Type 1 error which increases confidence interval.

Table 1: Hypothetical example of sensitivity, specificity and predictive values

	Disease	No disease	Total
Positive	151 (TP)	550 (FP)	701
Negative	55 (FN)	5430 (TN)	5485
Total	206	5980	6186

Source www.pitt.edu/~super7/7011-8001/7251.ppt, Sensitivity=151/151+55=73.30%, Specificity=5430/5430+550=90.80%, Positive predictive value=151/151+550=11.8%, Negative predictive value=5430/55+5430=99.9%. TP=True positive, FN=False negative, FP=False positive, TN=True negative. Note: Authors have computed a template to calculate in excel. If you feed in the numbers it will automatically generate values. Copy of the template could be obtained directly from the authors on request

In other words, larger the power more accurately the study can mirror the behavior of the study population.

The positive predictive values (PPV) or the precision rate is referred to as the proportion of positive test results, which means correct diagnoses. If the test correctly identifies all positive conditions then the PPV would be 100% and negative predictive value (NPV) would be 0. The calculative PPV in Table 1 is 11.8%, which is not large enough to predict cases with test conducted alone. However, the NPV 99.9% indicates the test correctly identifies negative conditions.

Clinical interpretation of a test

In a sample, there are two groups those who have the disease and those who do not have the disease. A test designed to detect that disease can have two results a positive result that states that the disease is present and a negative result that states that the disease is absent. In an ideal situation, we would want the test to be positive for all persons who have the disease and test to be negative for all persons who do not have the disease. Unfortunately, reality is often far from ideal. The clinician who had ordered the test has the result as positive or negative. What conclusion can he or she make about the disease status for his patient? The first step would be to examine the reliability of the test in statistical terms. (1) What is the sensitivity of the test? (2) What is the specificity of the test? The second step is to examine it applicability to his patient. (3) What is the PPV of the test? (4) What is the NPV of the test?

Suppose the test result had come as positive. In this example the test has a sensitivity of 73.3% and specificity of 90.8%. This test is capable of detecting the disease status in 73% of cases only. It has a false positivity of 9.2%. The PPV of the test is 11.8%. In other words, there is a good possibility that the test result is false positive and the person does not have the disease. We need to look at other test results and the clinical situation. Suppose the PPV of this test was close to 80 or 90%, one could conclude that most likely the person has the disease state if the test result is positive.

Suppose the test result had come as negative. The NPV of this test is 99.9%, which means this test gave a negative result in a patient with the disease only very rarely. Hence, there is only 0.1% possibility that the person who tested negative has in fact the disease. Probably no further tests are required unless the clinical suspicion is very high.

It is very important how the clinician interprets the result of a test. The usefulness of a positive result or negative result depends upon the PPV or NPV of the test respectively. A screening test should have high sensitivity and high PPV. A confirmatory test should have high specificity and high NPV.

Case control method is most efficient, for the study of rare diseases and most common diseases. Other measures of association from case control studies are calculation of odds ratio (OR) and risk ratio which is presented in Table 2.

Absolute risk means the probability of an event occurring and are not compared with any other type of risk. Absolute risk is expressed as a ratio or percent. In the example, absolute risk reduction indicates 27.37% decline in risk. Relative risk (RR) on the other hand compares the risk among exposed and non-exposed. In the example provided in Table 2, the non-exposed control group is 69.93% less likely compared to exposed cases. Reader should keep in mind that RR does not mean increase in risk. This means that while a 100% likely risk among those exposed cases, unexposed control is less likely by 69.93%. RR does not explain actual risk but is expressed as relative increase or decrease in risk of exposed compared to non-exposed.

Table 2: Different ratio calculation templates with sample calculation

	Scenario 1=Risk reduction		Total	Scenario 2=Risk increase		Total
	Cases (E)	Controls (C)		E	C	
Exposed	EE=20	CE=90	110	EE=70	CE=110	180
Non-exposed	EN=150	CN=140	290	EN=80	CN=140	220
Total subjects	ES=EE+EN=170	CS=CE+CN=230	400	ES=150	CS=250	400
ER %	EER=EE/ES=11.76	CER=CE/CS=39.13		EER=46.67	CER=44	

Other ratios

Ratio	Variable		Equation	Scenario 1 (decrease)	Scenario 2 (increase)
ARR	<0 absolute risk reduction	ARR	CER-EER	−27.37%	
	>0 absolute risk increase	ARI		-	2.67%
RRR/increase	<0 relative risk reduction	RRR	(CER-EER)/CER	−69.93%	-
	>0 relative risk increase	RRI		-	6.06%
NNT	<0 number needed to treat	NNT	1/(CER-EER)	−3.65%	-
	>0 number needed to harm	NNH		-	37.50%
RR		RR	EER/CER	0.30	1.06
OR		OR	(EE/EN)/(CE/CN)	0.207	1.11
AR (%)		AR	EER-CER	−27.37	2.67
ARP		ARP	(RR−1)/RR	-	5.71
PF (%)		PF	1−RR or 1−OR	0.70	-

ER=Event rate, ARR=Absolute risk ratio, RRR=Relative risk reduction, NNT=Number needed to treat, RR=Relative risk, OR=Odds ratio, AR=Attributable risk, ARP=Attributable risk percent, PF=Preventive fraction

OR help the researcher to conclude whether the odds of a certain event or outcome are same for two groups. It calculates the odds of a health outcome when exposed compared to non-exposed. In our example an OR of. 207 can be interpreted as the non-exposed group is less likely to experience the event compared to the exposed group. If the OR is greater than 1 (example 1.11) means that the exposed are 1.11 times more likely to be riskier than the non-exposed.

Event rate for cases (E) and controls (C) in biostatistics explains how event ratio is a measure of how often a particular statistical exposure results in occurrence of disease within the experimental group (cases) of an experiment. This value in our example is 11.76%. This value or percent explains the extent of risk to patients exposed, compared with the non-exposed.

The statistical tests that can be used for ascertain an association depends upon the variable characteristics also. If the researcher wants to find the association between two categorical variables (e.g., a positive versus negative test result and disease state expressed as present or absent), Cochran-Armitage test, which is same as Pearson Chi-squared test can be used. When the objective is to find the association between two interval or ratio level (continuous) variables, correlation and regression analysis can be performed. In order to evaluate statistical significant difference between the means of cases and control, a test of group difference can be performed. If the researcher wants to find statically significant difference among means of more than two groups, analysis of variance can be performed. A detailed explanation and how to calculate various statistical tests will be published in later issues. The success of the research directly and indirectly depends on how the following biases or systematic errors, are controlled.

Biases

When selecting cases and controls, based on exposed or not-exposed factors, the ability of subjects to recall information on exposure is collected retrospectively and often forms the basis for recall bias. Recall bias is a methodological issue. Problems of recall method are: Limitations in human ability to recall and cases may remember their exposure with more accuracy than the controls. Other possible bias is the selection bias. In case-control studies, the cases and controls are selected from the same inherited characteristics. For instance, cases collected from referral clinics often exposed to selection bias cases. If selection bias is not controlled, the findings of association, most likely may be due to of chance resulting from the study design. Another possible bias is information bias, which arises because of misclassification of the level of exposure or misclassification of disease or other symptoms of outcome itself.

Conclusion

Case control studies are good for studying rare diseases, but they are not generally used to study rare exposures. As Kaelin and Bayona explains[6] if a researcher want to study the risk of asthma from working in a nuclear submarine shipyard, a case control study may not be a best option because a very small proportion of people with asthma might be exposed. Similarly, case-control studies cannot be the best option to study multiple diseases or conditions because the selection of the control group may not be comparable for multiple disease or conditions selected. The major advantage of case-control study is that they are small and retrospective and so they are economical than cohort studies and randomized controlled trials.

References

1. Yang CY, Dao RL, Lee TJ, Lu CW, Yang CH, Hung SI, *et al*. Severe cutaneous adverse reactions to antiepileptic drugs in Asians. Neurology 2011;77:2025-33.
2. Wahner AD, Bronstein JM, Bordelon YM, Ritz B. Nonsteroidal anti-inflammatory drugs may protect against Parkinson disease. Neurology 2007;69:1836-42.
3. Garwood ER, Bekele W, McCulloch CE, Christine CW. Amphetamine exposure is elevated in Parkinson's disease. Neurotoxicology 2006;27:1003-6.
4. Vázquez MC, Vargas LM, Inestrosa NC, Alvarez AR. c-Abl modulates AICD dependent cellular responses: Transcriptional induction and apoptosis. J Cell Physiol 2009;220:136-43.
5. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ 1994;308:1552.
6. Kaelin AM, Bayona M. Case control study. The young epidemiology scholars program (Yes) is supported by the Robertwood Johnson Foundation and administered by the College Board, 2004. Available from: http://www.collegeboard.com/prod_downloads/yes/4297_MODULE-06.pdf.last accessed on 26 August 2013.

How to cite this article: Thomas SV, Suresh K, Suresh G. Design and data analysis case-controlled study in clinical research. Ann Indian Acad Neurol 2013;16:483-7.
Received: 01-03-13, **Revised:** 03-07-13, **Accepted:** 03-08-13
Source of Support: Nil, **Conflict of Interest:** Nil