Case control design: an efficient method to identify risk factors

Kingshuk Roy Choudhury and Daniel Barboriak,
Department of Radiology, Duke University

Executive summary

A case control study is a special type of retrospective study which can be used to identify risk factors or biomarkers for a disease. It requires relatively small sample sizes relative to a cohort study, particularly when the disease of interest is rare. This document describes how to design and perform a case control study and gives an example of how it might be relevant to clinical research in Radiology.

Identifying risk factors or biomarkers using a cohort study

In its simplest form, the identification of a risk factor or biomarker for a particular disease or clinical outcome involves studying a cohort of subjects, on which two types of information are recorded: a) the disease or clinical outcome (*D*), which may be present (*D* = 1)or absent (*D* = 0) b) the presence or absence of the potential risk factor *X*. Consider a cohort study to identify the risk of cigarette smoking on developing coronary heart disease (CHD) over a period of 20 years (Kanchanaraksa, 2008).

Table 1: Results of cohort study of the effect of cigarette smoking on developing coronary heart disease (CHD)

	Get CHD (*D* = 1)	Don't get CHD (*D* = 0)	Total	Incidence
Smoke cigarettes (*X* = 1)	84	2916	3000	84/3000
Don't smoke cigarettes (*X* = 0)	87	4913	5000	87/3000

The determination of whether the risk factor affects disease outcome is done by correlating *D* with *X*, to produce a summary measure like relative risk (*RR*) of disease (for those with the risk factor relative to those without), or equivalently, an odds ratio.

$$RR = \frac{\text{Risk of dis. w. risk factor}}{\text{Risk of dis. w.o. risk factor}} = \frac{n(D=1 \& X=1)/n(X=1)}{n(D=1 \& X=0)/n(X=0)} = \frac{84/3000}{87/5000} = 1.61 \qquad (1.1)$$

Alternatively, the added risk can be expressed in terms of an odds ratio:

$$OR = \frac{\text{Odds of dis. w. risk f.}}{\text{Odds of dis. w.o. risk f.}} = \frac{n(D=1 \& X=1)/n(D=0 \& X=1)}{n(D=1 \& X=0)/n(D=0 \& X=0)} = \frac{84/2916}{87/4913} = 1.63 \quad (1.2)$$

The 95% confidence intervals for *RR* (1.20,2.17) and *OR* (1.20,2.21) are similar and both suggest that cigarette smoking is a significant risk factor (p-value = 0.002) for developing coronary heart disease over a period of 20 years. The technique of multivariate logistic regression can extend this

methodology to the consideration of multiple risk factors for a given outcome and/or some of the risk factors are quantitative measurements.

Problems with cohort studies

While a cohort study like the one described above can yield strong conclusions, it typically requires large resources of time and effort to have sufficient power to draw conclusions. Namely, it took a) 20 years' worth of follow up b) 8000 study subjects to generate enough data (cases with CHD) for a significant result. The difficulty lies in the fact that most diseases or clinical outcomes are (thankfully) rare, thus taking a random of sample from a general population (or even a sick population such as hospital patients), would result in very few subjects with the disease, unless the sample size is very large.

Case control study

The solution to the follow up time issue is to look at a retrospective study, i.e. one where we look at past records to identify subjects with and without the disease outcome. The problem with standard retrospective studies is that information about the risk factor of interest may not always be available. For instance, the database may not record the smoking status of subjects for the study in Table 1. In the case of radiology research, identification of the risk factor may involve re-reading scans. In either case, obtaining the risk factor for 8000 subjects may be prohibitive.

An efficient solution involves selecting an enriched sample of subjects with the disease or clinical outcome. If only diseased cases are selected, however, we won't be able to be determining the effect of the risk factor: we need subjects both with and without the disease. The solution is to select (an equal) number of non-diseased cases (called controls). The key principle governing choice of controls is that they should be well matched with the diseased subjects in aspects other than the risk factor of interest. A study with retrospectively selected diseased cases and matched non-disease controls is called a **case control study**. Case control studies allow the assessment of risk in smaller size retrospective studies, when large and long term cohort studies are not feasible. For the coronary heart disease study in Table 1, a case control version might look like the following:

Table 2: Results of case control study of the effect of cigarette smoking on developing coronary heart disease (CHD)

	Get CHD (D = 1)	Don't get CHD (D = 0)
Smoke cigarettes (X = 1)	112	176
Don't smoke cigarettes (X = 0)	88	224
Total	200	400

$$OR_2 = \frac{\text{Odds of dis. w. risk f.}}{\text{Odds of dis. w.o. risk f.}} = \frac{n(D=1 \,\&\, X=1)/n(D=0 \,\&\, X=1)}{n(D=1 \,\&\, X=0)/n(D=0 \,\&\, X=0)} = \frac{112/176}{88/224} = 1.62 \quad (1.3)$$

The confidence interval for OR_2 (1.16, 2.29) leads to virtually the same conclusion as was obtained in the large cohort study in Table 1. Thus a similar conclusion could be reached by spending much less time and effort.

The principle of design of a cohort study is reversed in a case control study. In a cohort study, we start with a fixed number of subject with and without the risk factor (e.g. 3000 smokers and 5000 non-smokers, Table 1) and then follow up to determine how many develop the disease. In a case control study, we start with a fixed number of diseased and non-diseased subjects (200 cases and 400 controls in Table 2). We then determine how many of these have the risk factor of interest.

As might be expected, a case control study isn't a complete free lunch. It has some limitations. The first is that one can't estimate the disease incidence using a case control study. In Table 1, we learnt that the incidence of CHD was 84/3000 = 2.8/1000 subjects in smokers and 87/5000 = 1.7/1000 in non-smokers. This isn't possible in the case control study because we don't have the denominator, i.e. the number of subjects without the disease. This also means that we can't calculate the relative risk *RR* (1.1). We can, however, calculate the odds ratio *OR* and a little bit of algebra shows that when disease incidence is small, *RR* and *OR* give very similar values (compare results of (1.1) and (1.2)).

How to design and perform a case control study

Sample size

For of a case control study, the sample size required to achieve a given amount of power primarily depends on the size of the effect, i.e. the odds ratio. An odds ratio of 1.0 means that the prevalence of the risk factor in the group with the disease was the same as in the group without the disease. If we want our study to detect a small but statistically significantly difference in the odds ratio from 1.0 (for example, an odds ratio 1.1 or greater), the study will need a much larger sample size than for a study designed to detect a larger odds ratio (for example, an odds ratio of 2.0 or greater). The power also depends on other factors like the ratio of cases to controls, as well as the incidence. The most efficient ratio of cases to controls is 1, but cost or design considerations may dictate the use of a different ratio. In particular, for the purpose of greater generalizability, sometimes multiple control groups are used in a study. In the study by Teo et al (2006) described above, one control group could have been patients hospitalized with other diseases, while a second control group could have been visitors of patients. Due to such complications, it is best to consult a statistician when planning a case control study.

Selecting subjects for case control studies using DEDUCE

DEDUCE (Duke Enterprise Data Unified Content Explorer) is an on-line research tool providing Duke investigators with access to clinical information collected as a by-product of patient care. DEDUCE comprises over two decades worth of DUHS medical records. Current data sources include lab data, demographics, ICD-9 diagnoses, CPT procedures, inpatient medications, and CPOE orders. Access to DEDUCE requires enrolling in a training course and registration.

DEDUCE is particularly advantageous in designing case control studies in two ways: a) one can identify a set of cases quite easily by typing in an appropriate set of search terms. b) Once cases have been selected, matching controls (see below) can then be selected by searching using appropriate matching criteria. Note that the potential set of controls thus identified may be quite large, so a subsequent random sub-sample might need to extracted from this large set. Also, note that extraction of patient records from DEDUCE requires prior IRB approval. However, an assessment of the number of cases available can be made without IRB approval.

Choice of controls

The second limitation of case control studies lies the appropriateness of choice of controls, namely a) how well are they matched to cases. b) Is there a possibility that a third 'confounding' variable could be giving rise to the association we see in the *OR*? These limitations can be avoided in the cohort study, typically by choosing a random sample of subjects. In the case control study, since a random sample isn't possible, care must be exercised in the choice of controls. For instance, Teo et al (2006), who conducted a multi center (global) case control study for a similar question, used the following criteria: a) For cases, they chose 'all consenting cases without cardiogenic shock or history of major chronic diseases' b) For controls, they chose: 'At least one age-matched (plus or minus 5 years) and sex-matched control (without a history of heart disease or exertional chest pain) was recruited per case by use of specific criteria. A community-based control was either a visitor or relative of a patient from a non-cardiac ward or an unrelated visitor of another cardiac patient. A hospital-based control was selected from those at the same centre with illnesses not obviously related to coronary heart disease or its risk factors'. Notice that care was taken to exclude from controls those with any known history of heart disease. In a study of whether quantitative MRI measurements can predict knee replacement, similar criteria were used for control selection (Eckstein et al 2013). Note that only two factors (age and sex) were considered for matching. Matching on a large number of factors or 'overmatching', is to be avoided because it can result in loss of power in the study, for various reasons (Breslow, 2005). In any case, the effect of observable confounding factors can be adjusted in a multivariate analysis. Thus the principles for choice of controls are a) basic, e.g. age, gender matching with cases b) avoiding controls with other obvious risk factors for the disease outcome of interest and c) not less than the number of cases.

Analysis

A simple unmatched case control study, such as that in Table 1, can be analysed using a chi-squared test for association. In studies where the effect of other confounding variables needs to be accounted for, a multivariate modelling strategy, like logistic regression, can be used. The exact type of analysis will depend on the particulars of the study, so it is advisable to consult a statistician.

Case Control in Radiology

The value of the case control design in strengthening a retrospective study can be illustrated in an example. Suppose a radiologist has the impression that a finding on a head CT study is present more often in patients who subsequently develop dementia. One could do a retrospective study of patients who are diagnosed with dementia, using DEDUCE to identify patients who had obtained CT studies (for example) three or more years before the diagnosis, and record how often the finding in question was present in that cohort. This retrospective design could suggest that the finding often preceded the development of dementia, but without information on how often the finding was present in patients who did not become demented over that time period it would be difficult to judge whether the finding on CT is useful information that could diagnose these groups.

The study could be improved by using DEDUCE to find patients who had obtained a CT study and had not (within three years) developed dementia, and matching these control patients by age and sex with those in the case cohort. The reports of the CT images of both cases and controls could be reviewed or (even better) the images themselves reviewed in a blinded fashion to determine whether the finding was present, or not.

In this case, it would be crucial that the controls not have dementia developing within three years of the CT study. If the point of the research was to investigate whether the CT finding was an *independent* indicator of subsequent dementia, an attempt could be made using DEDUCE to match the controls with cases for known risk factors (for example, whether or not the patient complained of memory problems, or whether or not the patient had a family history of dementia).

References:

Breslow, N., 2005, Case-Control Studies, *Handbook of Epidemiology*, Springer, pp 287-319

Eckstein F, Kwoh CK, Boudreau RM, OAI investigators, 2013, Quantitative MRI measures of cartilage predict knee replacement: a case-control study from the Osteoarthritis Initiative. *Ann Rheum Dis.,* 72(5):707-14.

Kanchanaraksa, S., 2008, Fundamentals of Epidemiology I, http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/FundEpi/coursePage/index/

Teo, K. K., Ounpuu, S., Hawken, S., Pandey, M., Valentin, V., & Hunt, D. et al. (2006). Tobacco use and risk of myocardial infarction in 52 countries in the INTERHEART study: A case-control study. *Lancet*, 368(9536), 647-658.