Case-control study is also known as retrospective study. Unlike prospective study, the case-control study starts with cases (having the disease of interest) and a suitable control group (without the disease) and investigator obtains information regarding past exposure of possible aetiological factor(s) in both the groups. The investigator then compares the rates of exposure in both the groups. Objective of the case-control study is to find association between exposure and outcome (disease).

**Design of a case-control study:**

To examine the possible association between an exposure and disease, a group of people are selected who have the disease (cases) and, for the purpose of comparison, another group of people is identified who do not have the disease (controls). The proportion of people exposed to the factor of interest in the past is then ascertained in both the cases and controls. If the exposure is related to the disease, it is anticipated that the proportion of cases exposed will be higher than the proportion of controls exposed. Because of the design, it is not possible to calculate the incidence and prevalence rates of a disease from the data of a case-control study. Thus, the main features of a case-control study are:

- All relevant outcome has already been occurred by the time the study begins
- The study involves a group of people with the disease of interest and an appropriate control group (without the disease)
- This design is an indirect approach to measure the risk (in terms of odds ratio) of developing a disease for specific exposure
- Incidence and prevalence rate of a disease cannot be calculated from the data of a case-control study
- Relative risk cannot be measured from this design directly

Following table shows the data from a case-control study to determine the relationship between cigarette smoking and lung cancer. The data show that there are fewer heavy smokers among the controls and very few non-smokers among the lung cancer cases. Such findings strongly suggest an association (measured by Odds Ratio) between smoking and lung cancer.

<table>
<thead>
<tr>
<th>No. of cigarette smoked per day</th>
<th>Lung cancer patients</th>
<th>Control group</th>
<th>Odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7</td>
<td>61</td>
<td>1</td>
</tr>
<tr>
<td>1-4</td>
<td>55</td>
<td>129</td>
<td>3.7</td>
</tr>
<tr>
<td>5-14</td>
<td>489</td>
<td>570</td>
<td>7.5</td>
</tr>
<tr>
<td>15-24</td>
<td>475</td>
<td>431</td>
<td>9.6</td>
</tr>
<tr>
<td>25-49</td>
<td>293</td>
<td>154</td>
<td>16.6</td>
</tr>
<tr>
<td>50+</td>
<td>38</td>
<td>12</td>
<td>27.6</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>1,357</strong></td>
<td><strong>1,357</strong></td>
<td></td>
</tr>
</tbody>
</table>
Types of case-control study:
Case-control study can be of

a) **Unmatched case-control study**: where controls are selected without matching with the cases;
b) **Matched case-control study**: where controls are matched with one or more factors with the cases.
c) **Nested case-control study**: It is a case-control study within a cohort study. For example, researcher is interested to measure the effect of certain micronutrient (eg, serum vitamin A and colonic cancer) on development of cancer. At the beginning of a cohort study blood samples of all the subjects may be collected and preserved. After a certain period of follow-up, cases of cancer are identified from the cohort. Once sufficient numbers of cases are available for a case-control study, a comparison group (who has not developed cancer) is selected from the cohort to analyse data. The blood samples (of only the cases and controls) are then be analysed. With this design the risk factor can be evaluated with much smaller cost, as fewer blood samples need to be analysed.

Definition and selection of cases and controls:
First issue for a case-control study is to define the case or disease of interest. Criteria for defining a case should be set in such a way that it represents a disease as homogeneous as possible, since very often similar manifestations may be found in different diseases with different aetiologies (e.g., all congenital manifestation vs cleft palate). Therefore, strict diagnostic criteria for disease need to be established before conducting the study.

Once the diagnostic criteria are fixed, cases can be selected from a number of sources, such as hospital, general population, community etc. at a single point in time or over a specified period of time. If cases are selected from a hospital it is called hospital-based case-control study, while if the cases are selected from the population it is called population-based case-control study. Specific advantage of the population-based case-control study is that it avoids selection bias arising from hospitalisation. However, as the cost for population-based case-control study is prohibitively high, population-based case-control studies are not usually done.

Cases to be selected for a case-control study can be:

a) **Prevalent cases (existing cases)**: If disease is rare, inclusion of prevalence cases (for a specific period of time) will make it easier to find the cases for the study;
b) **Incident (newly diagnosed) cases**: It is always preferable to select newly diagnosed cases over a specific period of time.

Selection of appropriate controls is perhaps one of the most difficult issues in the case-control design. However, the controls selected should be comparable with the cases as regards to other factors except the exposure of interest. Controls should represent the population of non-diseased persons who would have been included as cases had they developed the disease. Such a case-control study will thus provide a valid estimation of association between exposure and disease.

To enable the investigator to determine the true effect of exposure, given the same underlying condition, cases and controls are selected from the same population group. Criteria for selection of controls for each type of cases are given in the following table.

Table 1. Guideline for selection of cases and controls

<table>
<thead>
<tr>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>All cases diagnosed in the community</td>
<td>Controls should be from general population of that community</td>
</tr>
<tr>
<td>All cases diagnosed in a sample of general population</td>
<td>Controls should also be from the sample of general population</td>
</tr>
<tr>
<td>Cases diagnosed at all the hospitals of</td>
<td>Controls should be from sample of patients in all hospitals in</td>
</tr>
</tbody>
</table>
Cases | Controls
---|---
the community | the community without the disease of interest
Cases diagnosed in a hospital | Sample of patients in the same hospital where cases were selected
Cases diagnosed at one or more hospitals | Samples of individuals who are residents in the same block or neighbourhood of cases
Cases selected by any of the above methods | Spouse, siblings or associates of cases, accident victims

Ideally there should have a single control group for cases. However, multiple control groups may be necessary under certain circumstances. For instance, to study association between coffee drinking and pancreatic cancer, it is found that the hospital controls differ from general population in coffee drinking. In such a situation it is preferable to select two control groups - one from the hospital from where cases are selected and another from the community. Once the source and number of controls are decided, it is necessary to decide how many controls for each case should be selected. When the number of cases and controls available are large, a ratio of 1:1 is optimum. When cases are limited, only a small number is available for the study, the ratio can be altered. As number of controls per case increases, the power of the test also increases. However, it is not recommended having a case-control ratio beyond 1:4. Finally, when cases and controls are selected, their previous exposure status for the factor of interest is determined for the purpose of analysis.

**Matched case-control study:**

Matching means “pairing” of cases with one or more controls based on their similarity. Cases may be matched with controls as regards to some characteristics, such as age, sex, socio-economic condition, blood group, ethnicity, occupation etc. Matching is usually done during the sampling phase of the study, when controls are selected matching with the cases. Matching can also be done during data analysis (post-matching). In this case (post-matching), cases are paired with controls after the unmatched controls are already selected for the study.

Primary objective of matching is to eliminate biased comparison between cases and controls. However, this objective can only be achieved if data analysis is in line with the matched design, i.e., matched design should follow matched analysis. Another objective is to achieve the balance between number of cases and controls at each level of matching variable (for example, if the study design matches with gender, then equal number of cases and controls will occur for males and females).

**Criteria for matching:**

Note that primary objective of matching is to control the extraneous factor(s) that may influence the exposure and outcome relationship. Therefore, matching should be done for the variables which are associated with outcome and/or exposure of interest. Matching cases with controls is appropriate in the following situations (E: exposure of interest; D: disease or outcome of interest and F: factor for matching):

- When there is interaction between F and E, and F is associated with the outcome. In this situation, without controlling for F, the data may show association (spurious association) between E and D, which is actually due to interaction between E and F.
- When F is a confounding factor, i.e., F is associated both with E and D, and D itself is a risk factor for the disease. In this situation, failure to control factor F would provide biased estimate of the individual effect between E and D.

On the other hand, situations when matching for a third factor is unnecessary are:

- When both the factors E and F are associated with the outcome (D) independently without any interaction between E and F. Here it may be logical to match for F as it is associated with D. However, as there is no interaction between E and F, it may not distort the association between E and D.

Dr Md Nazrul Islam
• Another situation when matching is unnecessary is when F is associated with E but not a risk factor for the disease independently.

<table>
<thead>
<tr>
<th>Situation when matching is appropriate</th>
<th>Situation when matching is unnecessary</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Diagram" /></td>
<td><img src="image" alt="Diagram" /></td>
</tr>
</tbody>
</table>

**Advantages and disadvantages of matching:**

Matching has several advantages.

- The process of matching achieves balance between cases and controls.
- Matching makes the cases and controls comparable to the factors matched.
- Matching eliminates the influence of extraneous factor (matching variable) in the relationship between exposure and disease. Thus, it is not necessary to adjust for the matched factor during analysis.
- Matching is the best way to test a hypothesis between exposure and outcome, by allowing one to conduct a comparatively small study. Matching is thus, very useful in small study.

However, there are a number of disadvantages for matching.

- As the controls are matched, it is not possible to assess the individual contribution of the factor matched in altering the risk of disease.
- It increases the cost for time and labor to find matches.
- There is loss of information on unmatched individuals.
- Matching increases the complexity of control selection process (e.g., you may look for a control who is female, age 45-46 years and blood group B, but you may not get this match during desired time period. Therefore, post-matching is sometimes useful for matching).
- Extensive matching increases data collection time.
- A certain fraction of cases are often discarded due to lack of a suitable control within the time frame.

**Strengths and limitations of a case-control study:**

**Strengths:**

- Case-control study is relatively quick and inexpensive compared with other analytic studies
- Suitable for evaluation of risk factor of disease of long latent period
- Suitable for evaluation of rare diseases
- Multiple risk (aetiology) factors can be studied for a single disease
Limitations:
- Not suitable for evaluation of diseases with rare exposure, unless exposure rate among cases is high
- Cannot directly compute the relative risk
- Prone to bias (such as selection bias, misclassification bias and recall bias) compared with other analytic studies,
- In some situation temporal relationship between the exposure and outcome is difficult to ascertain

Data analysis of a case-control study:

Unmatched case-control design:

Example: A case-control study is designed to test the hypothesis whether maternal malnutrition is a risk factor for low-birth weight (birth weight less than 2,500gm). To test the hypothesis, 200 new-born babies weighing less than 2,500gm were selected from a maternity hospital and an equal number of normal weight (weighing more than 2,500gm) new-born babies were selected from the same maternity hospital. Data obtained from the study is given in the following table.

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malnourished</td>
<td>95 (a)</td>
<td>50 (b)</td>
</tr>
<tr>
<td>Well nourished</td>
<td>105 (c)</td>
<td>150 (d)</td>
</tr>
<tr>
<td></td>
<td>200 (a+c)</td>
<td>200 (b+d)</td>
</tr>
</tbody>
</table>

As mentioned earlier, we cannot calculate the relative risk (RR), as it is not possible to compute the incidence rates from the case-control design. Rather, we calculate the Odds Ration (OR). The odds is the probability of success (exposed) by probability of failure (un-exposed) and is given by the formula:

\[
\text{Odds of exposure among cases} = \frac{a}{c} = \frac{a/(a + c)}{c/(a + c)} \quad \text{Or,} \quad \frac{a}{c} \quad \text{Or,} \quad \frac{\text{Probability of exposure}}{(1 - \text{Probability of exposure})}
\]

\[
\text{Odds of exposure among controls} = \frac{b}{d} = \frac{b/(b + d)}{d/(b + d)} \quad \text{Or,} \quad \frac{b}{d}
\]

Therefore,

\[
\text{Odds Ratio, OR} = \frac{a/c}{b/d} = \frac{ad}{bc}
\]

Note that odds can be converted into probability by using the following formula:

\[
\text{Probability} = \frac{\text{Odds}}{(1 + \text{Odds})}
\]

However, in our example, OR = 2.71

Interpretation: OR greater than 1 indicates that the exposure (here it is maternal malnutrition) is a risk factor for the disease (or outcome, here it is the low birth weight). If OR is less than 1, then the exposure is a protective factor (i.e., exposure is associated with lower risk for the disease), while OR equal to 1 indicates no association between the exposure and outcome. In our example OR 2.7 indicates that there is 2.7 times greater chance of having a low birth weight baby if the mother is malnourished compared with the well-nourished mother.

Dr Md Nazrul Islam
As we are not sure whether this OR (OR = 2.7) is by chance alone or not, we need to compute the 95% confidence interval (CI) for OR. The 95% CI for OR is given by the following formula.

\[
\text{95\% CI of } \ln \text{OR} = \ln \text{OR} \pm 1.96 \times (\text{SE of } \ln \text{OR})
\]

\[
\text{SE (standard error) of } \ln \text{OR} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}
\]

In our example, OR is 2.7; and

\[
\text{SE of } \ln \text{OR} = \sqrt{\frac{1}{95} + \frac{1}{50} + \frac{1}{105} + \frac{1}{150}} = 0.22
\]

Therefore,

\[
\text{95\% CI for } \ln \text{OR} = \ln 2.7 \pm (1.96 \times 0.22), \text{ or } 0.56 \text{ and } 1.42
\]

We can have the 95\% CI for the actual OR by taking the exponential of the calculated 95\% CI for lnOR, which is 1.75 and 4.14.

The 95\% CI for OR can also be calculated as:

\[
\text{95\% CI of OR} = \text{OR} \times e \pm (1.96 \times \text{SE of } \ln \text{OR})
\]

If, the 95\% CI includes 1, then the OR is not significant, that is there is no association between exposure and outcome. As, our calculated 95\% CI does not include 1, therefore there is significant association between the maternal malnutrition and low birth weight of the baby. We can also find the p-value by doing the Chi-square test as given by the following formula.

\[
\chi^2 = \frac{n (ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}
\]

In our example, \(\chi^2 = 21.9\), which is much higher than the tabulated value (3.841 with df 1). The p-value for this test is therefore < 0.001.

**Note:** It is the OR, which indicates the strength of association, rather than the p-value. The p-value indicates type I error that may have been committed in this study. Lower p-value indicates stronger evidence of association between exposure and outcome.

**Matched pair case-control design:**

In this design, cases are matched with controls one by one and the status of exposure is noted. The data generated in this design are expressed as per following table. Please note that data in each of the cells are coming as pairs.

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exposed</td>
</tr>
<tr>
<td><strong>Case</strong></td>
<td></td>
</tr>
<tr>
<td>Exposed</td>
<td>a</td>
</tr>
<tr>
<td>Not exposed</td>
<td>c</td>
</tr>
</tbody>
</table>
Cell “a”: indicates both the cases and controls are exposed;  
Cell “b”: indicates cases are exposed while the controls are not exposed;  
Cell “c”: indicates cases are not exposed while the controls are exposed  
Cell “d”: indicates both the cases and controls are not exposed.

Cells “a” and “d” are called concordant pair, as they are similar in terms of exposure. These cells do not therefore contribute in the analysis of data. On the other hand, cells “b” and “c” are called discordant pair, as they are dissimilar in terms of exposure of interest. For a matched pair case-control design, OR is calculated by the following formula:

$$OR = \frac{\text{Cell in which cases are exposed}}{\text{Cell in which controls are exposed}} \quad \text{or,} \quad \frac{b}{c}$$

Note that, if cases are placed on the column and controls on the row of the 2X2 table, then the formula will be:

$$OR = \frac{c}{b}$$

SE of ln OR for the matched pair case-control study is given by the following formula:

$$\text{SE of ln OR} = \sqrt{\frac{1}{b} + \frac{1}{c}}$$

We can therefore compute the 95% CI for OR by using the formula as given earlier. We can also determine the p-value by using the Chi-square test, the formula for which is as follows:

$$\chi^2 = \frac{(|b - c| - 1)^2}{(b + c)}$$

Two bars (before “b” and after “c”) indicate the absolute difference between b and c, ignoring the negative sign. In some book you may find the formula for chi-square test as $$\chi^2 = \frac{(b - c)^2}{(b + c)}$$.

Example: Data of a matched pair case-control study is given in the following table. Calculate the OR including the 95% CI.

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Not exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>55</td>
<td>130</td>
</tr>
<tr>
<td>Case</td>
<td>75</td>
<td>100</td>
</tr>
</tbody>
</table>

$$OR = \frac{c}{b} = \frac{130}{75} = 1.73$$

SE of ln OR = 0.15
95% CI of OR: 1.30; 2.30
$$\chi^2 = 14.22; p\text{-value} < 0.01$$
**Interpretation:** Compared to the unexposed, there is 73% (1.73 – 1) higher chance of developing the disease if a person is exposed to the risk factor of interest, which is statistically significant at 5% level of significance (95% CI: 1.30, 2.30; \( p < 0.01 \)).

**Relationship between OR and RR:**

OR is a good estimate of RR provided
- The incidence of the disease is low. OR tends to over estimate the RR unless the disease incidence is low (less than 10%)
- The cases are representative of all cases as regards to the exposure
- Controls are representative of general population (or non-diseased population) as regards to exposure

However, if it is felt that OR exaggerates the RR, then using the following formula one can easily estimate the RR, provided the incidence among unexposed (CI\(_0\)) is known.

\[
RR = \frac{OR}{(1 - CI_0) + (CI_0 \times OR)}
\]

**Proportional attributable risk and population attributable risk:**

Proportional attributable and population attributable risks can be calculated as discussed in the cohort study, just by replacing RR by OR.

\[
Proportional \ AR = \frac{OR - 1}{OR}
\]

\[
PAR = \frac{[P(E+)] \times [OR - 1]}{[P(E+)] \times [OR - 1] + 1}
\]

Where,

AR: Attributable risk
PAR: Population attributable risk
P(E+) indicates proportion of population exposed, and
OR is the odds ration

**Exercise:**

**Problem 1:** Dietary habits, alcohol and tobacco consumptions were studied in men with cancer of the larynx and a group of male controls. The results are shown in the following tables:

<table>
<thead>
<tr>
<th>Cigarette smoking (packs/day)</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>43</td>
<td>163</td>
</tr>
<tr>
<td>&lt;½</td>
<td>25</td>
<td>41</td>
</tr>
<tr>
<td>½ - 1</td>
<td>171</td>
<td>126</td>
</tr>
<tr>
<td>&gt; 1</td>
<td>135</td>
<td>51</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>374</strong></td>
<td><strong>381</strong></td>
</tr>
</tbody>
</table>

---

Dr Md Nazrul Islam
### Alcohol intake (units/month)

<table>
<thead>
<tr>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>38</td>
</tr>
<tr>
<td>&lt; 3</td>
<td>41</td>
</tr>
<tr>
<td>3 – 30</td>
<td>111</td>
</tr>
<tr>
<td>&gt; 30</td>
<td>179</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>369</strong></td>
</tr>
</tbody>
</table>

### Vitamin A (IU/month)

<table>
<thead>
<tr>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 50,000</td>
<td>98</td>
</tr>
<tr>
<td>50,000 – 150,000</td>
<td>201</td>
</tr>
<tr>
<td>&gt; 150,000</td>
<td>39</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>338</strong></td>
</tr>
</tbody>
</table>

**Questions:** Calculate the relative risk (in terms of OR) including the 95% CI of cancer of the larynx:
1. At every level of cigarette smoking compared with non-smokers.
2. At every level of alcohol consumption compared with no alcohol consumption.
3. At every level of vitamin A consumption compared with the highest.
4. How should the results be interpreted?

**Problem 2:** In order to find out whether the use of oral contraceptives (OC) affects the risk of developing myocardial infarction (MI), a case-control study was carried out among married nurses in Bangkok, where 159 hospitalized cases of MI were compared with 3,180 non-MI cases. The results showed that 21 of the cases and 273 of the controls had used oral contraceptives.

**Questions:**
1. Calculate the relative risk (OR) with 95% CI and interpret your results
2. If the oral contraceptive prevalence rate among the nurses is 56%, what amount of disease incidence could be prevented if nurses stop using OC?