

CASE-CONTROL STUDIES: DEFINITION AND STATISTICAL ANALYSIS

By Isabel Herrero del Real

Regarding clinical research, we can distinguish two major approaches for medical investigation: experimental studies and observational studies.

Experimental studies are studies where the researchers introduce an intervention and will study its effects. The clearest example of an experimental study is a **randomised clinical trial**. For example, a randomised clinical trial of *smoking reduction using oral nicotine inhalers*: eligible volunteers will be randomly assigned to different groups; one group will receive the intervention (drug being studied, in this case, the oral nicotine inhaler) and the control group will either receive placebo or nothing. The researchers will study what happens to the individuals constituting each group and will test the efficacy, effectiveness, efficiency, and safety of the drug.

On the other hand, **observational studies** are studies where the researchers observe the effect of a specific variable (risk factor, diagnostic test, etc.) without interfering or manipulating the research subjects. They can be classified as cohort studies, cross-sectional studies, and case-control studies. **Cohort studies** are those that allow to follow research participants over a period of time to see what outcomes emerge as a result to an exposure. For example, the *Nurses' Health Study* followed the potential long-term consequences of the use of oral contraceptives. Additionally, we can find **cross-sectional studies**, which allow the analysis of the study population at a specific point in time. An example of a cross-sectional study would be evaluating COVID-19 positive infections among unvaccinated and vaccinated teenagers during Spring of 2022.

The last type of observational studies are the case-control studies, which we are going to review in depth.

CASE-CONTROL STUDIES

Case-control studies are studies used to investigate the relationship between an exposure, which can be a characteristic of the environment or of people, and a health outcome. There are two clearly defined groups: one with the disease/outcome (cases) and one without it (controls). It is important to mention that they are **retrospective studies**, meaning that there is existing data that will allow the comparison of both groups. In other words, the study will look back in time to analyse whether there are statistically significant differences between the groups in the rates of exposure to a specific risk.

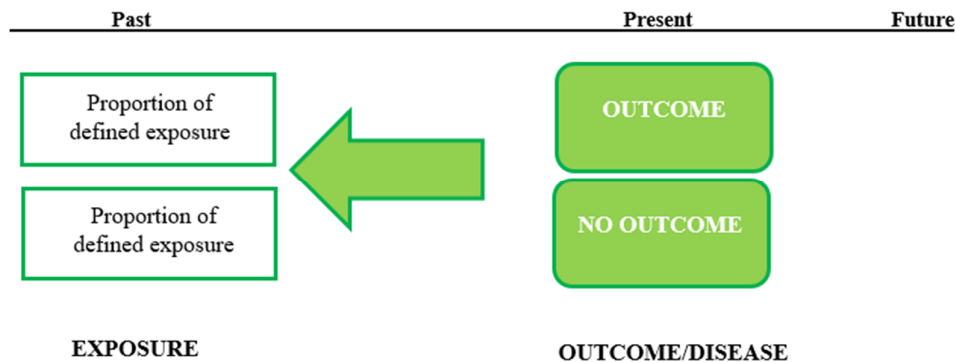


Figure 1. Design of a case-control study

The **benefits** of case-control studies include:

- **Simple, quick, and inexpensive.** The data is already existent, and experimenters establish the groups after the outcomes are known. No experiment is going to be conducted.
- **Good for studying rare conditions or diseases.** We start with a study population who already have the disease, as opposed to following a population and waiting for the development of the disease. Therefore, it eases the enrolment of patients.

- **Useful for preliminary research and to assess multiple risk factors.**

However, case-control studies also have certain **drawbacks**:

- **Do not demonstrate causation.** They can be used to establish correlations between exposures and outcomes but cannot demonstrate causation.
- **Potential for recall bias.** People can be more motivated to recall risk factors. Recall bias can lead to associations between exposures and outcomes which, in fact, do not exist.
- **Difficult to find a suitable control group.**
- **Confounding variables and bias.** There can be a distortion of the measure of correlation between the exposure and the outcome.

STATISTICAL ANALYSIS OF CASE-CONTROL STUDIES

In clinical studies, it is important to figure out how the exposure influences in the health outcome. The **relative risk (RR)** and **odds ratio (OR)** are two measures that allow to evaluate the association between the exposure and the outcome.

Data is going to be reflected in a **contingency table**, which will have two entries: a row entry, which will reflect the exposure (i.e., a treatment), and a column entry, which will reflect the outcome. An example of a contingency table is represented in **Figure 2**:

	Cases	Controls
Exposed	a	b
Unexposed	c	d
Total	a + c	b + d

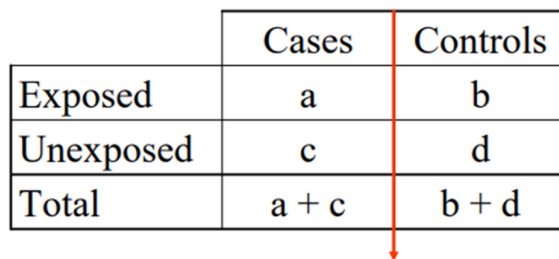


Figure 2. Example of a contingency table.

Relative risk (RR)

The **relative risk (RR)** is the number of times a patient is more likely to improve with the treatment compared with placebo. It reflects the ratio of probability of an outcome in an exposure group divided its likelihood in an unexposed group.

It can be calculated using the following formula:

$$Relative\ risk\ (RR) = \frac{Probability\ of\ outcome\ in\ exposed\ group}{Probability\ of\ outcome\ in\ unexposed\ group} = \frac{\left(\frac{a}{(a + b)}\right)}{\left(\frac{c}{(c + d)}\right)}$$

The values of relative risk (RR) can be interpreted as follows:

- If the **RR is equal to 1 (RR = 1)**, the exposure does not affect the outcome.
- If the **RR is lower than 1 (RR < 1)**, the risk of the outcome is decreased by the exposure, so it is considered a “protective factor”.

- If the **RR is higher than 1 (RR > 1)**, the risk of the outcome is increased by the exposure, so it is considered a “risk factor”.

To understand the use of this formula and the interpretation of the results, we will use the following example:

Example. Unvaccinated and vaccinated people and measles. The exposure is the vaccine, and the outcome would be measles infection.

	Measles infection	No measles infection
Vaccinated	15	110
Unvaccinated	40	90
Total	55	200

If we calculate the RR using the previous formula:

$$Relative\ risk\ (RR) = \frac{\left(\frac{a}{(a+b)}\right)}{\left(\frac{c}{(c+d)}\right)} = \frac{\left(\frac{15}{(15+110)}\right)}{\left(\frac{40}{(40+90)}\right)} = \frac{\left(\frac{15}{125}\right)}{\left(\frac{40}{130}\right)} = \frac{0.12}{0.31} = \mathbf{0.387}$$

As we can see, the relative risk is lower than 1 ($0.387 < 1$), which means that vaccination decreases the risk of measles infection.

In case-control studies the **relative risk cannot be estimated** because the overall prevalence and incidence of the outcome is unknown. Nonetheless, researchers can calculate the odds that a person with the outcome was exposed to the risk factor and the odds that a person without the outcome was exposed to the risk factor. Thus, case-control studies use the **odds ratio (OR)** as a measure of the association between the exposure and the outcome.

Odds Ratio (OR)

The **odds ratio (OR)** quantifies the relationship between the exposure and the outcome in a case-control study and it tells us how much higher the odds of exposure in case-patients are in comparison to those in control-patients. Statistically speaking, odds represent the probability of an outcome occurring divided by the probability of the outcome not occurring.

It can be calculated using the following formula:

$$Odds\ ratio\ (OR) = \frac{Probability\ of\ outcome\ occurring}{Probability\ of\ outcome\ not\ occurring} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a * d}{b * c}$$

The values of the odds ratio (OR) can be interpreted as follows:

- If the **OR is equal or close to 1 (OR = 1)**, the odds of exposure among case-patients are the same or similar to the odds of exposure among control patients. This means that the exposure is not associated with the disease.
- If the **OR is lower than 1 (OR < 1)**, the odds of exposure among case-patients are lower than the odds of exposure among control-patients. This means that the exposure might be a “protective factor” against the disease.
- If the **OR is higher than 1 (OR > 1)**, the odds of exposure among case-patients are greater than the odds of exposure among control-patients. This means that the exposure might be a “risk factor” for the disease.

The further away the OR is from 1, the more likely that the association between the exposure and the outcome is causal. To understand the use of this formula and the interpretation of the results, we will use the following example:

Example. Smoking and lung cancer. The exposure is smoking, and the outcome would be lung cancer.

	Lung cancer	No lung cancer
Smokers	120	30
Non-smokers	55	150
Total	175	180

If we calculate the OR using the previous formula:

$$\text{Odds ratio (OR)} = \frac{a * d}{b * c} = \frac{120 * 150}{30 * 55} = \frac{18000}{1650} = \mathbf{10.91}$$

As we can see, the odds ratio is higher than 1 (10.91 > 1), which means that the exposure of tobacco smoke is a risk factor for lung cancer.

The odds ratio can also be linked to confidence intervals and p-values. When interpreting odds ratios, OR = 1 represents no effect, therefore:

- **Null hypothesis (H₀).** The OR will be 1, there is no relationship between the exposure and the outcome.
- **Alternative hypothesis (H₁).** The OR is different to 1, so there is a relationship between the exposure and the outcome.

REMEMBER:

The **p-value** is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming H₀ is true.

Confidence intervals give an expected range of where the parameter of study will fall within.

If the **p-value** of the OR is lower than the significance threshold (i.e., 0.05), we reject the null hypothesis because the difference between the OR and 1 is statistically significant and our data display enough evidence to conclude that there is an existing association between the exposure and the outcome and that it is not due to chance.

Alternately, we can use **confidence intervals (CI)** to evaluate odds ratios. If our confidence interval includes 1, the results are not statistically significant, whereas if it excludes 1, the results would be statistically significant.

The upper and lower 95% CI can be calculated using the formulae shown below:

$$\text{Upper 95\% CI} = e^{\wedge} [\ln(\text{OR}) + 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}]$$

$$\text{Lower 95\% CI} = e^{\wedge} [\ln(\text{OR}) - 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}]$$

To understand the use of this formulae and the interpretation of the results, we will use the previous example of smoking and lung cancer:

	Lung cancer	No lung cancer
Smokers	120	30
Non-smokers	55	150
Total	175	180

The OR was of **10.91**.

To calculate the upper and lower 95% CI, we apply the formulae:

$$\text{Upper 95\% CI} = e^{\ln(10.91) + 1.96 \sqrt{\frac{1}{120} + \frac{1}{30} + \frac{1}{55} + \frac{1}{150}}} = e^{2.895} = 18.07$$

$$\text{Lower 95\% CI} = e^{\ln(10.91) - 1.96 \sqrt{\frac{1}{120} + \frac{1}{30} + \frac{1}{55} + \frac{1}{150}}} = e^{1.884} = 6.58$$

As we can see, the 95% CI corresponding to an OR of 10.91 would range from 6.58 to 18.07 → 95% CI = [6.58, 18.07]

As it excludes 1, we can say that the results are statistically significant and that there is an association between the exposure (smoking) and the outcome (lung cancer).

Now, we are going to apply all the addressed concepts into a real-world case-control study: **Bohlken, J., Jacob, L. and Kostev, K. (2018). Association Between the Use of Antihyperglycemic Drugs and Dementia Risk: A Case-Control Study. *Journal of Alzheimer’s Disease*, 66(2), pp.725–732.**

The case group was formed by type 2 diabetes mellitus patients who had received a dementia diagnosis, whereas the controls included T2DM patients without dementia. Two multivariate regression models were used to study the association between the use of antihyperglycemic drugs and dementia risk:

The **results of the first multivariate regression model** were the following:

- Glitazones. OR = 0.80; 95% CI = [0.68 – 0.95]
- Insulin. OR = 1.34; 95% CI = [1.24 – 1.44]

In this case, we can see that glitazones were associated with a decrease in the risk of developing dementia, because the OR was lower than 1 (0.80 < 1), so these drugs act as a “protective factor”. As the 95% CI excludes 1, we can say that the results are statistically significant and that there is an association between glitazones and the reduction of dementia.

On the other hand, insulin show an OR higher than 1 (1.34 > 1), indicating that insulin administration acted as a risk factor for the development of dementia. Again, as the 95% CI excludes 1, we can say that the results are statistically significant and that there is an association between glitazones and the reduction of dementia.

Among the different types of insulin, basal insulin (OR = 1.18; 95% CI: 1.07–1.29) 163 and premix insulins (OR = 1.31; 95% CI: 1.19-1.44), both were considered risk factors for the development of dementia.

The **results of the second multivariate regression model** were the following:

- Metformin monotherapy. OR = 0.71; 95% CI = [0.66 – 0.76]

- Metformin combined with sulphonylureas. OR = 0.90; 95% CI = [0.89 – 0.92]
- Combination of basal insulin + bolus insulin. OR = 1.47; 95% CI = [1.32 – 1.63]
- Combination of basal insulin + premix insulin. OR = 1.33; 95% CI = [1.14 – 1.56]

In this case, the administration of metformin both in monotherapy or in combination with sulphonylureas was considered a “protective factor” against the development of dementia due to the OR being lower than 1 and the 95% CI excluding 1.

Conversely, the administration of insulin resulted to be a risk factor in the development of dementia, as the OR is greater than 1 and the 95% CI excludes 1.

Useful information:

- Relative Risk & Odds Ratios: <https://www.youtube.com/watch?v=Sec4fewyUig>
- Medical Statistics – Part 7: OR and RR in Observational Studies: <https://www.youtube.com/watch?v=7ymCiLPP9os>
- Mann, C.J. (2003). Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emergency Medicine Journal*, 20(1), pp.54–60.
- Statistics by Jim - Case-control study: <https://statisticsbyjim.com/basics/case-control-study/>
- Statistics by Jim – Relative Risk: <https://statisticsbyjim.com/probability/relative-risk/>
- Statistics by Jim – Odds Ratio: <https://statisticsbyjim.com/probability/odds-ratio/>