Case-Control Studies: The Influence of Statistics on Culture and Health

Written by: Raphael Angelo Gonzalez Crisostomo

What is a Case Control Study and Why do we do it?

A Case-Control study design is a form of retrospective study to help **determine if an exposure is associated with an outcome**. It begins with a known outcome then retrospectively analyzes the data to investigate exposures.

From this description I am certain you can imagine the implications and utility of this type of study. From the top of your head, I am sure you can think of many clinical illnesses associated with certain risk factors

Let's have a look at some examples:

1. Smoking and lung cancer
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2038856/?page=4

2. Chronic alcohol consumption and liver disease
   https://www.sciencedirect.com/science/article/abs/pii/089543569390032V

3. Hormonal contraception and breast cancer
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4154499/

the list goes on…

I think the best way to understand a case control study is through an example of this design. A classic example of a case control study all the way back from the 1950's is the association of smoking and lung cancer which has had a global impact on health, well-being, and the ethics of advertising-- Smoking and Carcinoma of the Lung by Doll and Hill (1950).

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2038856/?page=4

When conducting a research design, the first step is to **define** the **cases** (the participants with the outcome of interest), as specifically as possible and select accordingly.

And, in the case of smoking and lung cancer?

It would be **patients with lung cancer** since this is the outcome we are trying to determine, but at times the definition of a disease may be based on multiple criteria; therefore, all these variables should be clearly defined within the case definition. For example, they may define lung cancer in terms of size, metastasis, location but in this case a histopathologic confirmation was sufficient.

This makes **smoking our exposure of interest**, but again we must clearly define what is a smoker. According to Doll and Hill, a smoker is anyone who had at least 1 cigarette a day for at least a 1-year time span.

Another important parameter to establish is the **matching criteria** – this is used to ensure that the cases and controls are similar in specific characteristics. Also, the control group must be at similar risk of developing the outcome of interest. Once a matching variable has been chosen, you should not analyze it as a risk factor.

When selecting the **controls**, the population from which the cases and controls to be included should be equal. In addition, exposure should be similar in both cases and controls.

Hospital controls
Patients admitted to or consulting with the hospital for illnesses apart from the outcome of interest. They are controls which are easy to recruit and have a similar standard of medical record keeping. However, certain co-morbities could be found within these patients and may be needed to be taken into consideration

Relative and friend controls
Cases may also recommend their friends or relatives a source of controls. They can be helpful if we want to make sure that measurable and non-measurable confounders are distributed at an equal approximation in cases and controls (such as lifestyle, socio-economic status, or genetic factors).

They are simple to find, and more likely to share socioeconomic status and other demographic characteristics with the patients. Although, these controls are also more prone to engage in similar behaviors (alcohol use, smoking, etc.).

Population controls
Having a list of people makes it simpler to carry out these controls. Lists derived from phone books, voter registration lists, local census etc.

Although, they may not be cost efficient and can be time-consuming. Additionally, many of these controls won't be motivated to take part in the study resulting in quite a low response rate.

So, let's go back to our study, for Doll and Hill. What is our control, control type and matching criteria?

They did a comparison of **709 non-cancer patients (control)** who **were matched by age, gender, and hospital from the general medical and surgical patients (control type).**

After that, the researcher evaluates the exposure in both of these groups. As a result, a case-control study's intended outcome must manifest itself in some of the study's participants.

Now that the two groups have been defined and identified, the cases (a group **WITH** the outcome of interest) and the controls (a group **WITHOUT** the outcome of interest).

Then, we retrospectively assesses the exposure in both these groups. If the exposure is more frequently observed in the cases than in the controls, you can formulate the hypothesis that the exposure is associated with the outcome of interest.

Now, let's summarize:

- Outcome: 709 patients with lung carcinoma, determined histopathologically (cases)
- Exposure: Smoking, defined as anyone who had at least 1 cigarette a day for at least a 1-year time span
- Control: 709 patients from general medical and surgical wards. (controls)
    - Matched according to age and sex.
    - Specifically: They included 649 males and 60 females in cases as well as controls.

Although, the controls and cases are equal this is not necessarily always the case, typically we would have more controls

How do we analyze this information?

In analyzing a case-control study we go to certain statistical tools, for example:

Odds Ratio (OR): A measure of association between an exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

This is used to compare the relative odds of the occurrence of the outcome of interest (Lung Cancer), given exposure to the variable of interest (Smoking). The odds ratio can also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome.

To process this information, we need to construct a contingency table to compute for the odds ratio:

	Outcome (+)	Outcome (-)
Exposure (+)	a	b
Exposure (-)	c	d

a = + exposure, + outcome
b = + exposure, - outcome
c = - exposure, + outcome
d = - exposure, - outcome

But, we need some additional information:
Of the 709 cases, 688 patients of the patients with lung cancer were smokers
Of the 709 controls, 650 non-lung cancer patients were smokers

Insert the data,

	Outcome: Lung Cancer (+)	Outcome: Lung Cancer (-)	
Exposure: Smoker (+)	688	650	1338
Exposure: Smoker (-)	21	59	80
	709	709	1418

And compute for the odds ratio,

OR = (A/C)/(B/D)

OR = (688 x 59)/(650x21) = 2.973
OR = 2.973

How do we interpret this result?

When interpreting the OR, we follow the following rules

OR=1 Exposure does not affect odds of outcome (no effect)
OR>1 Exposure associated with higher odds of outcome (risk factor)
OR<1 Exposure associated with lower odds of outcome (protective factor)

Our result of 2.973 indicates that our exposure (smoking) is associated with higher (2.973~3 times) the odds of (lung cancer)

An OR >1 would indicate that smoking is a risk factor for lung cancer

Since OR >1 we follow: OR -1

Your risk of developing lung cancer if you are a smoker is computed as:

OR-1 (x100%)

= 2.974-1 X 100% = **197.3% at risk of developing lung cancer if you are a smoker**

Although Doll and Hill did not utilize a **Confidence Interval (CI)** in their 1950s study, it can be another useful tool to further analyze its clinical relevance.

Typically, a 95% CI is used to estimate the precision of the OR. Since it gives a range for the true odds ratio to fall between. The rule is, if it includes 1 you cannot reject the hypothesis that the exposure and the outcome are not related

As you may recall: OR=1 means that the exposure does not affect odds of outcome (no effect). Therefore, if in the range wherein the true OR falls, the confidence interval, we observe that 1 is within that range it would indicate that the findings are not statistically significant.

Let's find out if our findings are statistically significant via the CI.

To compute for the confidence interval of an OR we use the following formula:

Confidence Interval = exp(log(or) ± Zα/2*√1/a + 1/b + 1/c + 1/d)

Z = 1.96

Now we plug in our computed values,

Confidence Interval = exp(log(2.973) ± (1.96)α/2*√1/688 + 1/650 + 1/21+ 1/59)

And we get the following result:

Confidence Interval **[1.787, 4.949]**

Therefore, we can conclude that indeed it is clinically significant

When should we perform a case-control study?

Since the data and outcomes are already established and our analysis is retrospective, we can enjoy the following **advantages**:

- Cost-effective and earlier publication of findings: The data is already there, no for recruitment, and experimentation
- (Data collection) Rare outcomes and outcomes with long latent periods: compared to a cohort study, there is no need to follow the progression of the outcome
- Multiple exposures in the same outcome: Since we clearly define the outcomes and variables beforehand this can be a useful tool to determine association between exposures and outcome
- The association of risk factors and outcomes in outbreak investigations.

However, Case-Control studies do have their **limitations**.

- Although they are useful for the data collection of rare exposures, a cohort study would be more ideal. Here you can find a detailed comparison of a case-control and cohort study (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2998589/)

- Susceptibility to biases – selection bias and recall bias. Selection bias because it is the investigator that retrospectively decides which subjects are placed into the control and case group (no randomization, no blinding). Recall Bias, since a subject who had mild outcome (symptoms) may not be allocated in the appropriate group.

- Another important limitation is, that we are **NOT** able to estimate the incidence or prevalence in a case-control study since we chose the number of cases and controls which may result in the proportion to be misrepresentative of the population

In summary, case-control studies are extremely important in establishing risk factors as we saw with smoking and lung cancer. When at that time, health care professionals

could not identify the reason for an outbreak in lung cancer. As we saw in the number of smokers in the sample population, a large portion of the population were smokers. I'm sure that this study has changed the lives of many people and helped them rethink their lifestyle choices.

References:

DOLL R, HILL AB. Smoking and carcinoma of the lung; preliminary report. Br Med J. 1950 Sep 30;2(4682):739-48. doi: 10.1136/bmj.2.4682.739. PMID: 14772469; PMCID:PMC2038856. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2038856/?page=4

Tenny S, Kerndt CC, Hoffman MR. Case Control Studies. 2022 Mar 28. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan–. PMID: 28846237. https://pubmed.ncbi.nlm.nih.gov/28846237/

Setia MS. Methodology Series Module 2: Case-control Studies. Indian J Dermatol. 2016 Mar-Apr;61(2):146-51. doi: 10.4103/0019-5154.177773. PMID: 27057012; PMCID: PMC4817437. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4817437/

Lewallen S, Courtright P. Epidemiology in practice: case-control studies. Community Eye Health. 1998;11(28):57-8. PMID: 17492047; PMCID: PMC1706071. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1706071/