Association and confounding

Introduction

To oversimplify, but not by much, you can think of science as a search for ***cause-and-effect*** relationships and for theories that unite them. In this module we illustrate two things to look for as part of using data to find causal relationships. The first is the ***association*** between two variables. Consider smoking and cancer. Initially, scientists found that smokers had higher rates of lung cancer than did nonsmokers. Smoking and cancer were *associated*. Did that *prove* that smoking *caused* cancer? No. For example, some scientists thought there might be a gene that made people both likely to smoke and likely to get lung cancer. The presence of the gene (yes/no) could be a *confounding* variable. The confounder could explain the association. To conclude that smoking causes lung cancer, you must be able to rule out the effect of possible confounders.

Is association evidence of possible causation? Yes. If smoking *causes* cancer, there has to be an association. Cancer rates will be higher for smokers. If the two are associated, one *might* cause the other. Association is necessary, but association alone is not enough to prove cause and effect.

Do COVID vaccines work?

On December 8, 2020, 90-year-old Margaret Keenan from England was the first person in the world to receive a COVID-19 vaccine outside of a clinical trial. This was the beginning of widespread vaccinations in England that started with residents in nursing homes and those who cared for them. This was followed by health care workers and then those aged 70 and older, then 65 and older, etc. By June 18, 2021, all adults in England were eligible to receive the first dose of the vaccine. Second doses were available anywhere from 3 to 12 weeks after receiving the first dose.

The vaccine was supposed to have a moderate effect on reducing the rate of infection but was to have a much stronger effect on reducing the severity of the infection and thus decrease the death rate for those infected. We will investigate this second point. To do this, we will look at results from a report from Public Health England specifically from all the reported cases of the Delta variant of the virus from February 1 to August 2, 2021. From these cases, we will look at whether the patient was vaccinated and whether the patient lived or died.

SARS-CoV-2 variants of concern and variants under investigation in England Technical briefing 20 published by Public Health England on August 6, 2021.
https://assets.publishing.service.gov.uk/media/610d031f8fa8f506aab17866/Technical_Briefing_20.pdf

1. Identify the observational units and variables in this study. Also classify each variable as categorical (also binary?) or quantitative.

Often, when a study involves two variables, it is natural to consider one the ***explanatory variable*** and the other the ***response variable***.

Definitions: The ***explanatory variable*** is the variable we think is "explaining" the change in the response variable and the ***response variable*** is the variable we think is being impacted or changed by the explanatory variable. The explanatory variable is sometimes called the independent variable and the response variable is sometimes called the dependent variable.

0. Which would you consider the explanatory variable in this study? Which is the response? (That is, what are the *roles* of these variables in this study?)

Table 1 shows all cases involving the Delta variant of SARS-CoV-2 (the virus that causes COVID-19) in England from February 1 to August 2, 2021 where the vaccination status of the patient was known. Vaccinated here means that the patient had received at least one dose of the vaccine.

	Unvaccinated	Vaccinated	Total
Died	253	481	734
Survived	150,799	116,633	267,432
Total	151,052	117,114	268,166

Table 1: The mortality and vaccination status for all COVID-19 cases in England involving the Delta variant from Feb to Aug 2021

0. Did a smaller proportion of the vaccinated patients die than unvaccinated? Support your answer by calculating the conditional proportions of deaths for those that were unvaccinated and also the conditional proportion of deaths for those that were vaccinated. (Write both these proportions as decimals.)

Surprisingly perhaps, you should have found that a larger proportion of the vaccinated group died than the unvaccinated group. Opposite of what you might expect. One way to compare these two conditional proportions is to calculate their ratio. This ratio is called a ***relative risk***. Relative risks are usually calculated by dividing the larger proportion by the smaller proportion.

Definition: ***Relative risk*** is the ratio of two conditional proportions. It indicates how many times greater the risk of an outcome is for one group compared to the risk for the other group.

0. Calculate the relative risk of death by dividing the proportion of deaths for those that were vaccinated by the proportion of deaths for those that were unvaccinated. Write a sentence interpreting this ratio value. Does the value of this ratio strike you as noteworthy?

Definition: Two variables are ***associated*** or related if the value of one variable gives you information about the value of the other variable. When comparing two groups, association can be seen when the proportions or means take different values in the two groups.

0. Do vaccination status and if someone died appear to be associated (albeit in the opposite direction of what we might expect to find)?

There are two possible explanations for this odd finding that those that were vaccinated were more likely to die than those that were unvaccinated:

- The vaccinations help *cause* more deaths to occur.
- The vaccinations did *not* cause more deaths to occur, and some other issue (variable) explains why there were a larger proportion of deaths among the vaccinated. In other words, a third variable is at play, which is related to both vaccination status and if someone died.

(Of course, another explanation is random chance though we can safely rule this out.)

6. Consider the second explanation. Suggest plausible alternative variables that would explain why those that were vaccinated were more likely to die than those that were unvaccinated. In other words, besides the vaccination status, were those that were vaccinated different from those that were not vaccinated in some ways that could make them more likely to die?

One possible alternative variable that might help explain why those that were vaccinated had a higher death rate is ***age***. While the vaccinated group and the unvaccinated group could be different in lots of ways that would also affect mortality rates, age is one that might have happened. Either by allowing older people to be vaccinated first and their greater desire to be vaccinated, the vaccinated group could, on average, be much older than the unvaccinated group. Let's see if that is the case. While exact ages were not given in the report, the patients were split into whether they were less than 50 years old (we'll call this group younger) or 50 years old and older (we'll call this group older). A two-way table showing vaccination status by age group is shown in Table 2.

	Unvaccinated	Vaccinated	Total
Older	3,440	27,307	30,747
Younger	147,612	89,807	237,419
Total	151,052	117,114	268,166

Table 2: The age category and vaccination status for all COVID-19 cases in England involving the Delta variant from Feb to Aug 2021

7. Which group, unvaccinated or vaccinated, included a larger proportion older patients? Calculate the relevant (conditional) proportions to support your answer.

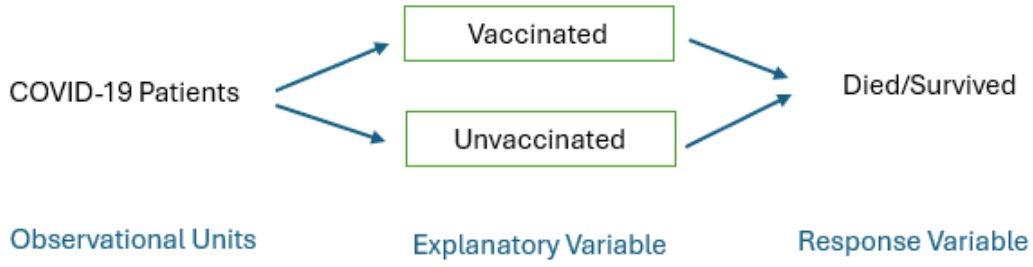
You should have seen in #7 that the vaccinated group had a much larger proportion of older patients. Thus, there seems to be an association between age and vaccination

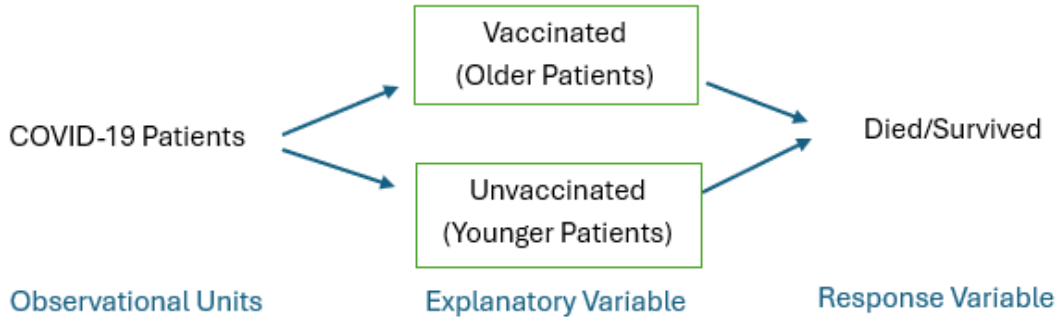
status. For age to matter or make a difference in death rates, it must also be associated with mortality. (If for some reason vaccinated people were much more likely to be left-handed but left-handedness had no effect on death rate, then left-handedness would not matter.) We call variables that are associated with both the explanatory variable and the response variable confounding variables.

Definition: A ***confounding variable*** is a variable that is related both to the explanatory and to the response variable in such a way that its effects on the response variable cannot be separated from the effects of the explanatory variable.

Confounding explains why you cannot draw a cause-and-effect conclusion from association alone: The groups defined by the explanatory variable could differ in more ways than just the explanatory variable when confounding is present. Figure 1 illustrates the confounding in the COVID-19 study. The top panel shows the study design: Observational units (COVID-19 patients) are sorted into groups according to the explanatory variable (whether or not they were vaccinated). Then the response (died/survived) was observed. The bottom panel shows the confounding: those vaccinated tended to be older while unvaccinated tended to be younger.

Figure 1: Confounding in the COVID-19 study





We already saw that age is associated with vaccination status. Now let's see if age and mortality are associated. Table 3 shows the number of patients that died and survived for each of our age groups.

	Younger	Older	Total
Died	69	665	734
Survived	237,350	30,082	267,432
Total	237,419	30,747	268,166

Table 3: The mortality and age category for all COVID-19 cases in England involving the Delta variant from Feb to Aug 2021

8. Which group, older or younger, had a higher proportion of deaths after contracting COVID? Again, calculate the relevant (conditional) proportions to support your answer.

9. Explain how your answers to #7 and #8 establish that age is a confounding variable that prevents drawing a cause-and-effect conclusion between vaccine status and death from the disease.

You should have initially found that the mortality rate was higher for those that were vaccinated from COVID-19 than those that weren't. You should have also seen that we can't determine any sort of cause and effect from this because of the presence of confounding variables. Namely, the vaccinated group and the unvaccinated group could be different in lots of ways that also could affect mortality rates, with age being one of them. Let's dig a little deeper into the data from the report and try to make the vaccinated group and unvaccinated group a bit more similar. To do this, we will focus on just the younger group and then just the older group.

Remember that the data back in Table 1 showed all cases involving the Delta variant of SARS-CoV-2 (the virus that causes COVID-19) in England from February 1 to August 2, 2021 where the vaccination status of the patient was known. Table 4 separate that data into just the younger group and just the older group.

Younger Group

	Unvaxed	Vaxed	Total
Died	48	21	69
Survived	147,564	89,786	237,350
Total	147,612	89,807	237,419

Older Group

	Unvaxed	Vaxed	Total
Died	205	460	665
Survived	3,235	26,847	30,382
Total	3,440	27,307	30,747

Table 4: The mortality and vaccination status for all COVID-19 cases in England involving the Delta variant from Feb to Aug 2021 separated into our two age groups

10. Let's find out which group, unvaccinated or vaccinated, had a smaller proportion deaths for each age group. Again, support your answers by calculating the conditional

proportion of deaths for those that were unvaccinated and also the conditional proportion of deaths for those that were vaccinated.

a. First compare those in the younger group.

a. Now compare those in the older group.

11. Initially, you should have found that there was a larger proportion of deaths in the vaccinated group. Is this also true when you just look at the younger patients? Just the older patients?

You should have seen that a larger proportion of patient died in the unvaccinated group than in the vaccinated group for both the younger and older groups. This is in the opposite direction from when both groups were combined together at the beginning. This phenomenon where the association reverses direction for the two groups compared to when all the data were merged together is called Simpson's paradox. With our data, the vast majority of deaths came from the older group and this group also had a larger proportion of vaccinations.

Definition: An association or comparison that holds for all of several groups that reverse direction when the data are merged to form a single group is called **Simpson's Paradox**.

12. Let's take another look at relative risk.

a. Calculate the relative risk of death for the younger group as well as for the older group. Remember to do this by dividing the larger conditional proportion be the smaller in each group.

a. Explain what these relative risks mean in context.

b. Which group does the vaccine seem to have the largest benefit? Explain.

13. Now based on what you found in #12 can you say that the vaccine is causing the reduced death rate for those with COVID-19? Why or why not?

When confounding variables could be present, we can't conclude cause-and-effect even though there is an association between two variables. So how can cause-and-effect be determined? More specifically, how do researchers determine that a certain drug causes a reduction in death or a decrease in symptoms? To do this, they need to create two very similar groups. One group is then given the drug and one is not. If they now see a difference in the response, they know it must have happened because of the drug. You will explore this idea in more detail in other modules.

Summary

Two variables are **associated** (related) if the values of one variable provide information about (help you predict) the values of the other variable.

Studies that involve two variables often distinguish between the roles played by the variables.

- The **explanatory variable** is the one that is suspected of possibly affecting the other.
- The **response variable is** the one that may be affected by the other.

A **confounding variable** is one whose effects on the response cannot be separated from those of the explanatory variable.

The possible presence of confounding variables is the reason that association alone does not justify a conclusion that differences in the explanatory variable *cause* differences in the response variable.

An association or comparison that holds for all of several groups that reverse direction when the data are merged to form a single group is called **Simpson's Paradox**.