This lesson is going to focus on a very important and commonly used study design, the case control study.

Let's say we wanted to know more about the causes of brain cancer. Brain cancer is relatively rare. Overall, the chance that a person will develop a malignant tumor of the brain or spinal cord in his or her lifetime is less than 1% (about 1 in 140 for a man and 1 in 180 for a woman). There are also different types of brain cancer and researchers often just focus on one type, which increases the rarity of the disease. Causes of brain cancer are not fully known but radiation exposure has been found to be a risk factor. Brain cancer also takes many years to develop. We don't know exactly how long but we do know that people who receive radiation therapy to the head during cancer treatment have an increased risk of developing brain tumors 10 - 15 years later. So we can assume that there could be 10-15 years of even more between exposure and disease. One hypothesized cause of brain cancer is cell phone use. Another hypothesized risk is electromagnetic field exposure from power lines or household appliances. Research to date has not clearly shown an association with these but there is interest in this possibility. There is also evidence of occupational exposure to certain metals and chemicals. So let's say we want to study the association between potential exposures in adolescence and brain cancer for people 30-40 years of age. Here are some factors that will impact on choosing a study design.

1. The disease is rare.
2. We need to consider a long time between exposure and disease, around 10-20 years.
3. We want to look at multiple exposures
4. We only have one outcome.

Remember, when selecting a study design, one of the first things we consider is how to select people to be in the study. We can select people based on being representative of a certain group (Cross sectional study), people could be selected by exposure (Cohort study) or people selected by outcome (Case-control study).

If we wanted to look at a representative sample of people without regard to exposure or disease, as in a cross-sectional study, we would need to enroll many thousands of people into the study because brain cancer is rare, especially between the ages of 30 and 40. It is also a disease with fairly low survival so people with brain cancer may not be selected for our study because they may have died before we could enroll subjects into the study.

If we wanted to look at people exposed and follow them over time, we would also have some serious problems. We would need a very long study, since our exposure is during adolescence and our outcomes is age 30-40 years. We wanted to look at many different exposures so we would have to include many people in the study to cover all these different exposures. And again because of the rarity of the disease, we would need many thousands of people in the study to be sure we had enough people to obtain individuals with the disease. It is clear that these two approaches are not very feasible.

An alternative would be to identify people by disease using a Case-control study. One can find people with brain cancer in Cancer hospitals, or Cancer registries. It is possible to find people without cancer to use as controls although there are challenges there, which we will see later. We can ask people about past events so we would not need a long study, and we could ask them about many exposures. Overall this approach is much more feasible for a rare disease with a long time period between exposure and disease. It is not without challenges as we want to accurately obtain information which occurred in the past. But this
study design becomes possible. So let’s take some time to review the different characteristics of the case-control study.

Case control studies are a type of Analytical study design with the purpose of identifying the cause of a disease. Essentially a group of individuals with a disease are compared to a group of individuals without the disease. They are also called a retrospective study because case control studies start with people with disease and look backward for previous exposures which might be relevant to the development of the disease. The overall goal is to compare differences in exposures between the two groups.

In this example, we are going to consider the exposure as having heart disease. People with heart disease are identified by having a red X on their chest. People without disease do not have a red X. Exposure will be having smoked cigarettes in the past. So we want to know if people who have heart disease were more likely to have smoked cigarettes in the past. Exposure is identified by a cigarette package and non-exposure is a cigarette package with a blue X.

Start by identifying your cases. These are the people with heart disease. Then identify a group of controls. These are people without the disease. Identify the numbers of cases who were exposed and those who were not exposed. Identify the number of controls who were exposed and those not exposed. Compare the rate of exposure between the two groups.
As I mentioned before we will see 2 by 2 tables many times in epidemiology. I try to always set them up the same with cases yes/no across the top and exposure yes/no down the side. But each time you see a 2x2 table be sure to check and make sure that is how it is set up.

Box A includes the people with disease who are exposed
Box B includes the people without disease who are exposed
Box C includes the people with disease who are not exposed
Box D includes the people without disease who are not exposed.

Let’s look at this picture to review how this study works. We begin by finding people with and without the disease. Those with the disease are called cases and those without the disease are called controls. We look to the past to identify exposures. We compare the rate of exposure between the two groups.
When we conduct a study, we first need to first consider your population. When you conduct a research study, you think about who you want your study to represent. So if you wanted to do a study on the association of smoking and asthma, you would have to consider your source population. If you wanted to do a study that could be generalized to all Florida residents, you would have to sample from people across the Florida (purple box). If you only wanted to generalize to the Tampa population, you would need to include people from the Tampa area. If you only wanted to generalize to USF students then you would only include students from USF. And you would want your cases and controls to come from the same source population. While this seems a bit simplistic, I will tell you that many early research studies on heart disease were conducted on samples of high income white men. At the time of these studies, these findings were considered representative of the US population but it did not take into account hormonal differences in heart disease among women and racially and economically diverse populations.

Many studies are done on relatively small source populations because they are available and it is more feasible. There are national studies which include individuals across the United States, but these are expensive. You just need to think about the meaning of your study results. It can indeed be that people across the country would have the same risk between cigarette smoking and asthma that you found, but generally we learn that by repeating important studies in different populations. The main thing is to think about it so when you finish a research study, you can accurately describe the source populations and who your study applies to.
You need to think about what differences there might be between your study subjects and the larger population.

Just to restate, cases and controls should come from the same population. Controls ideally are those individuals who would have been cases if they had gotten the disease.

Just to remind you how this is set up:

Box A includes the people with disease who are exposed.
Box B includes the people without disease who are not exposed.
Box C includes the people with disease who are not exposed.
Box D includes the people without disease who are not exposed.

When we analyze case-control studies, we need to do so in a special way because we do not have incident cases. We essentially want to determine the odds of exposure among the diseased as compared to the odds of exposure among the non-diseased. We use the following formula to calculate this:

\[ \frac{a \times d}{b \times c} \]

This is called the odds ratio.

Think about it a minute as this makes sense. If we hypothesize that the exposure is associated with the disease then box a (people with the disease who are exposed) and box d (people without the disease who are not exposed) both support our hypothesis. Likewise box b (people without the disease who are exposed) and box c (people with the disease who are not exposed) do not support our hypothesis.

This is one of the formulas that you are expected to know. When we look at incident cases, we...
calculate something known as the relative risk which is a measure of risk in new cases of disease. You will learn this when we do a cohort study. The important thing to know is the odds ratio approximates the relative risk as long as two things are true:

So let’s look at an example calculating an odds ratio. We have the usual 2 by 2 table with a, b, c, and d labelled. If you are unsure what the cells represent click on the letters for clarification. Cases go across the top and exposure goes down the side. First, we need to fill in the table. We know there are a total of 120 cases and 120 controls so put those numbers in first along the bottom of the table. Look at the case side. We know that 53 cases had the exposure, attending daycare. Put 53 into Cell A (the number of cases who were exposed). 120 - 53 = 67 so put 67 in cell C. We then do the same for the controls. We know 26 controls were exposed (attending daycare) so put 26 in cell C. Subtract 120 - 26 = 94 so 94 goes in Cell D.

Now we have all the information we need to calculate the odds ratio. This is a times d divided by b times C / 53 times 94 divided by 67 times 26 which equals 4983 divided by 1742 which equals 2.86. The way you interpret this result is that children with pediatric pneumococcal disease were 2.86 times more likely to attend day care than children without pediatric pneumococcal disease. Or children attending day care had almost at 3 times the risk.

Interaction

Now you try to see if you can fill in the 2 by 2 table and calculate the formula. Use the numbers in green boxes for the table and the numbers in blue boxes for the formula. If you get it wrong, just try again. I put the formula on top as a reminder for you. This type of calculation will be on the test so you want to be sure you can do it. Good luck.
Sometimes in research studies we do not find as many cases of disease as we wish. We will learn about determining sample size later but it is important to know that you can have multiple controls for every case. You might remember that was done in the case control study of the 8 girls with vaginal cancer. The researchers selected 32 controls or 4 per case. In general, around 4-5 controls per case maximizes the sample size with not much benefit if we go beyond that.

Another technique that is sometimes used in case-control studies (and other studies as well) is matching. There are two types of matching: group matching and pair matching. In group matching for example, you would select controls that have the same distribution of a characteristic as cases. If 20% of your cases were older than 65, then you would want 20% of the controls to be in the same age group. In pair matching, you match each individual case to its control. In this instance you select a specific case for each control and match on some important factor, like disease severity or age. When you choose to match on a factor, there are two important things to consider. You can never evaluate the effect of the variable you matched on as you artificially set its rate. Also your analysis is then limited to the discordant pairs, that is the matched pairs in which the case and control differ in terms of exposure.

We actually covered quite a bit in this lecture. We reviewed:

- Definition of a case control study
- Times the case-control study is a good choice
- The construction of a case-control study
- Advantages and disadvantages of case control studies
- Understanding the source population
- How to analyze case-control studies
- Calculating the number of controls

In the next lecture on case control studies, we are going to focus on identifying cases and controls, and review matching in case control studies as well as go over some different types of designs. If you had difficulty with the exercises in this section, you should review them so you will be more comfortable working with this design.