This lesson is going to focus on a common in creating case-control studies, beginning with identifying cases and controls and then moving onto analysis and study design issues.

Since cases have the disease of interest, they can be obtained from different places where people are sick. Hospitals, HMOs, and physician practices are common sources. Disease registries are often used, especially cancer registries. But there are registries for a number of diseases. The other source might be a sample of the general populations in a community. The source of cases will depend on a number of factors, including the rarity of the disease. The more rare a disease is, the less likely you could use a community sample.

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.
Before starting to identify cases, we need to develop a case definition. Case definitions include criteria that identify the characteristics of a case. Ideally they are very sensitive, i.e., if someone truly has the disease of interest he/she will be correctly identified. We also want them to be specific, in that if someone does not have the disease they will be not be considered a case. We will learn more about sensitivity and specificity in future lessons.

So what do we use to identify cases? We may use a list of symptoms, e.g., fever over 100 degrees, vomiting. We might base it on a positive laboratory test for a given infectious disease or platelet count. Sometimes, we simply use a physician diagnosis as definitive. Other times cases may not actually be identified until an autopsy is done after death. On the other side, we may identify cases through self-reported symptoms, a medical record review, or even based on the results of a survey. The method we use is dependent upon the given situation as well as the availability of laboratory testing. Early in the AIDS epidemic, the US would use HIV testing to identify cases but in countries where these tests were not available, symptoms were used. It was often very difficult to determine the difference between AIDS and tuberculosis in people in many developing countries. And to make it more difficult, sometimes people had both illnesses.
In some studies, cases can be classified as definite and probable, depending upon the number of positive findings. For example, ebola may be defined as definite if there is a positive blood test, and probable if there is fever and bleeding but no test is available. Sometimes, it takes time to determine who is a true case.

Cases can also be classified as new cases of disease (incident cases) or prevalent cases of disease (existing cases). Studies relying on incident cases can take longer to conduct as you need to wait for new cases. At times this may not be possible.

Controls allow us to identify the history of exposure among a population that is similar to the cases except that they do not have the disease. If I were to say to you that 30% of people with Disease A have exposure B, it would be hard for you to know if exposure B was a possible cause of Disease A unless you knew how many people without the disease had the exposure. So the goal of a control group is to provide an estimate of exposure in the source population in the absence of disease. If people with a disease have a higher prevalence of exposure than those without the disease, it is possible the exposure causes the disease. If people with the disease have a lower prevalence of exposure than those without the disease, then the exposure may be protective. If there is no difference in the
The main characteristic of controls then is that ideally controls are people who would have been selected as cases if they had the disease. In addition, there should be no association between the selection of controls and the exposure. The reason for this is if controls have an increased risk of exposure from some other factor, we may miss an association between cases and controls. For example if you were studying the relationship between heart disease and cigarette smoking, you would not use lung cancer patients as your controls, as they would be even more likely to be smokers than your cases with heart disease. So we want our controls to have the “normal” rate of smoking. One common mistake that students make is they think they should pick controls who do not have the disease and the exposure. But if we did that, we would get inaccurate results as well. Controls serve to give us an estimate of what the usual exposure is.

So where do we find controls? Remember that controls should be those individuals who if they had the disease would likely have been selected as cases.

General population controls are those we obtain through getting a random sample of the general population from which the cases came.
There are two main types of matching: group matching and individual matching. In group matching, the final control group will match the final case group in terms of some characteristic. For example, if you match on gender and one third of your cases are female than one third of your controls are female. You do this by selecting your cases first and then selecting the correct proportion of controls from your source population. You end up with the same proportion of genders.

In pair case matching, every time you identify a case, you identify a control that is matched to that specific case. Each time you enroll a female case, you need to enroll a female control. Again, you end up with the same proportion in each group but the controls are specifically assigned to a case.

It is in the analysis that this becomes challenging.
This is an example of the matched case control analysis. In this study cases were people with colon cancer and controls did not have cancer. Cases and controls were individually matched by age. The exposure of interest was eating a high fat diet. The study groups were as follows:

30 cases and their controls both ate high fat foods. This is a concordant pair as the case and control had the same exposure. They are in cell A and will not be in the analysis.

60 colon cancer cases ate high fat food while their controls did not. This is a discordant pair. They are in cell B and they will be in the analysis.

25 cases did not eat high fat foods while their matched controls did. This is a discordant pair. They are in cell C and they will be in the analysis.

85 cases and their matched controls did not eat high fat foods. These are concordant pairs and they will not be in the analysis.

Thus we divide the two discordant pairs b/c 60/25 which equals 2.4
Thus cases were almost 2 1/2 times more likely to eat high fat foods than controls.

Interaction
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.

"From January 21, 2000 until December 1, 2000, we enrolled 16 patients in a randomized trial of clonazepam versus placebo in fibromyalgia. This was designed as a crossover study and each patient received 0.5 mg of clonazepam and an identical looking placebo-each for a 12-week period. There was a washout period during which no medications were taken between the two 12-week sessions. In this crossover design, each patient functioned as their own control. The patients were examined at six points during the study. We evaluated each patient's tender points.
as well as each patient's assessment of global well-being, pain, sleep, and fatigue at six different time intervals.

"For each of these parameters, we found no significant difference in clonazepam versus the placebo. Although clonazepam improved sleep and fatigue significantly, the placebo response was just as robust. There was more of a homogeneity (consistency) in the response to clonazepam, whereas the placebo response varied dramatically. Some patients got much worse on placebo, but some surprisingly got much better.

So take this small test to see if you know who belongs in the two groups of a case-control study. Our next lecture will review another common study type, the cohort study. You will see as we go through it that in many ways it is the opposite of a case-control study.