In this lecture, I am going to discuss confounding. It is part of a series of lectures that looks at threats to epidemiology studies, things that can lead to incorrect assumptions.

So what exactly is a confounder?

In a research study, you are looking at the association of an exposure and an outcome but a third variable can get in the way. This third variable is something associated with the both exposure and the outcome. It is also a risk factor for the outcome, independently of its association with the exposure. It is also not on the causal pathway. What that means is this third variable isn’t something that is caused by the exposure and it then causes the disease. Let’s see how this works. Click on our confounder to see a demonstration.

Confounding layer: This picture illustrates what I am talking about. You have three variables and they are all related to each other. The exposure (X) is associated with the outcome (Y). The confounder (C) is also associated with the exposure (X), and independently of that association, it is associated with the outcome (Y).

So let’s look at an example. We are studying the association of alcohol and bladder cancer, and we find an association between alcohol use and bladder cancer. But before we can make any assumptions about this relationship, we need to consider other potential factors. One we would choose to look at is smoking. So we know that people who drink are more likely to smoke cigarettes. And we also know that cigarette smoking is more likely to cause bladder cancer. So it is possible that it is not the drinking that is causing the bladder cancer but rather it is the smoking that is causing it, and we saw an association that is not really true. I will show you ways to manage confounding in a future slide. We generally use the term control for confounding. But before I explain this, I want to describe one other possible
situation, in which the confounding variable is on the causal pathway. Click on the next button to see this example.

So sometimes, the potential confounding variable is actually caused by the exposure and it in turn causes the disease. If we try to eliminate the effect of the confounding variable, we will also eliminate the effect of the exposure. Let's consider smoking, elevated blood pressure, and heart disease. Smoking causes elevated blood pressure, and increased blood pressure causes heart disease. It is said to be on the causal pathway because it is directly between smoking and heart disease. We would not want to manage this confounding as we will lose the effect of smoking. So when you look at a variable that you think might be a confounder, one of the first things you want to do is see if it is on the causal pathway. Make a diagram as that often helps. Click on the next button to see how confounding might impact on our study.

A confounder can impact on the data in two different ways.

It can be the cause of an apparent association, which would not have occurred in the absence of a confounder. or
It can actually hide the effect of a relationship between an exposure and an outcome. This is sometimes called reverse confounding. Click on the buttons to see examples of these confounders.

Confounding example. We already looked at the example of smoking as a confounder for the association of alcohol and bladder cancer. Here is an example from maternal and child health. We know there is an association between Black race and preterm birth with Black women about twice as likely to have a preterm baby as white women. Even when we control for things like education and SES, this pattern persists. The reason for this association is not clear but there is evidence that it may be due to confounding by stress with black women more likely to experience stress than white women. Click on return to go back to the main slide.

Reverse confounding. An excellent example of this is infant feeding practices. The World Health Organization recommends breastfeeding for two years or more, because of evidence that breastfed children have a reduced risk of infectious agents and are less likely to die. However, some studies have found poorer outcomes among breastfed babies. A
possible explanations is that is that women in communities with very poor resources are the most likely to breastfeed and yet their children are at even greater risk than communities where children are not breastfed as long. So how does this work? As you can see in this picture the exposure, breastfeeding is associated with good health. However, poverty is also associated with breastfeeding and is negatively associated with healthy children. Poverty is a predominate cause of infant mortality. Thus it hides the relationship of breastfeeding with good health and makes it appear that breastfeeding leads to poor infant health and increased mortality. Click on return to go back to the main slide.

We refer to confounders as potential confounders because they may not necessarily be a confounder. We evaluate variables to see if they meet the criteria for confounding. A variable is not a confounder if it is only associated with the exposure but not the disease. It is also not a confounder if it is associated only with the disease but not the exposure. In this situation, the disease might possibly cause the confounder. The other situation we considered previously was when the potential confounder was on the causal pathway, and in that case the exposure might have caused the potential confounder which then caused the disease.

How does a researcher setting up a study identify confounders? The first step is to search the literature. Find other factors that are associated with the disease you are studying and be sure you have considered them. Biology is also useful. If a factor might be associated with a disease because it makes sense biologically, then that is a good one to include in the model. And lastly, there is an educated guess. Something may just seem important to you. One can almost always consider age, as it is often a confounder in many associations.

Here is a true story of my experience with confounding. Hit play to watch the video.
There are six ways to leave confounding, and they can be done in either the design or the analysis. I have identified the ones in the design phase in blue and the ones in the analysis phase in green. Click on the band to learn more about each of these. Also always click on return to return to this slide and on next to go to the next slide when you are done. I am sure many of these concepts will be familiar to you as we discussed many of them in prior lectures.

Randomization is done in the design phase: Individuals in different study arms are randomly assigned into treatment or control groups. This can be used in Randomized Clinical Trials or Community Intervention Trials. Randomization controls for confounding by equalizing exposed/unexposed groups (treatment/non-treatment) for many characteristics. Can control for both known and unknown factors. Needs a large number of participants. Randomization of 10 people into 2 groups may be unbalanced because of large variability in 10 people. Randomization of 1000 people into 2 groups increases the probability that characteristics will be similar between the two groups.

Restriction is done in the design phase: A method in which you limit your study to people without a certain characteristic. Often done when there is a strong association between one factor and an outcome, and you want to look at other factors. An example would be if you wanted to look at other risk factors for lung cancer than smoking. Eliminating smokers would allow you to see other associations, which you would likely miss if you had smokers in the study. If you restrict subjects to certain characteristics, you limit the variability on that factor and reduce (eliminate) the possibility of confounding by that variable. Advantages and disadvantages of restriction:

Advantages
- Cheap
- Straightforward
- Convenient
- Offers complete control if confounder is nominal or simple dichotomous
Disadvantages

- Can substantially reduce N
- Residual confounding if category is insufficiently narrow (e.g., age)
- Cannot study restricted factor
- Association is not generalizable to all people, only subgroups being studied

Matching is done in the design phase, but it has a strong impact on the analysis. Matching is done when you match people in your study groups on different characteristics.

We covered this in an earlier lesson on case control studies where this is commonly used.

If you are unsure of this, please review the earlier slide and complete the matching exercise.

There are different types of matching:

- Individual in which each case is matched to a specific control
- Frequency in which you make sure you have the same frequency of a certain factor, but you do not match individuals

Limitation is once you match of a factor you cannot identify its impact

Limits the number of people available for your study. For example if you match on age and identify a case who is age 25-30, you might have to skip enrolling the next 5 possible controls if they are the wrong age.

Again review the past material if you are not sure how to do the matched analysis:

- You will need to calculate a Matched OR
- You do that by only considering the discordant pairs, that is those in which the case and control differ in terms of exposure

Multivariate analysis is done in the analysis phase. It is the most common means of controlling for confounding. You create a statistical model that allows you to put a number of variables into it and identify the effect of each variable if we controlled for the effect of the others.

Used when we have too many variables to consider in stratifying

We can adjust for multiple confounders at one time.

For example:

Diabetes = age + gender + genetics + diet + exercise

Diabetes is the dependent variable

We get an odds ratio for each independent factor which has taken into account the effect of every other independent factor.

Some types of multivariate analysis you will hear about are linear and logistic regression. We will not teach you how to do this in this class but wanted you to know...
that it exists. This will be covered in Biostatistics classes.

**Standardization:** This is the age adjustment we discussed in a prior lecture. We create a standardized model which allows us to control populations if we adjusted for another factor, such as age. We might use the distribution of a standard population to compare mortality between two populations with different age distributions. Please feel free to review the materials on this that we covered before. I listed the two types of age adjustment and the links to the videos by Dr. Patwari if you want to review this. Just click on the blue arrows to see the links.

**Stratification:** An analytic technique to control for confounding, involving the evaluation of the association within homogeneous categories, or strata, of the confounding variable. When you have two variables and you want to know if one is a confounder, you can test this in a stratified analysis. It is easier to understand in the context of an example. Let’s say you are looking at the relationship between smoking and heart disease but you think there may be some confounding by age. You can divide your subjects into different groups by age. Although you can do multiple ages, for the sake of this example, let’s just do two age groups. Start with the base table first which looks at the relationship between high smoking and heart disease. So you create a 2 by 2 tables with disease (yes, no) across the top and exposure (yes/no) along the side. Multiply a times d and divide this by b times c to get the odds ratio which is 2.86. This is called the crude odds ratio as it is not adjusted for any factors. Now we are going to obtain odds ratios stratified by age. Click on the next button to go to this slide.

Now, the people in the original study have been divided into old (older than 60) and young (45-59). Age is a relative factor for those of you who think 45 is pretty old. Once the tables are created, you calculate the odds ratios for the two strata: old and young. So the next step is to determine if there is evidence for confounding. This can be done fairly simply by creating a line and plotting the odds ratios on it. If the crude odds ratio is between the odds ratio for the stratified analysis then there is evidence of confounding. Click on the arrow to see this line.

This line shows the three odds ratios in numerical order. The first question is the crude odds ratio between the adjusted...
odds ratios? The answer is clearly no. The first adjusted rate for the old participants is 1.84. The second adjusted odds ratio for the younger people is 2.17 and the crude odds ratio falls higher than those at 2.86. It is not between the adjusted rates.

Thus the second question is “Is the evidence of confounding?” And the answer is no. In this case, we did not find confounding but if we did, we could adjust for this using a statistical technique called the Mantel-Haenzal test. Click on the arrow to see how this works.

Basically, you create a pooled odds ratio. In the formula, the Greek letter E (Epsilum) stands for sum. The i’s stand for the fact that this includes all of these items in the calculations. So this formula becomes the following formula when you have 2 strata. If you take a minute to look at it, it really is pretty simple. On the top of the formula, you are multiplying a times d for each 2 by 2 table and then adding them together and on the bottom you are multiplying b times c for each 2 by 2 table and adding them together, and then dividing the top line by the bottom line. That gives you one odds ratio for the two strata but you can only do this if they are similar to each other. If they are not, then you have something called effect modification which we will cover shortly. But even better, you can get SPSS to calculate this for you and you do not need to do it by hand. Click on the video button to see a demonstration on how to do this calculation in SPSS.

Here is a quiz for you in which different measures for controlling confounding are described. Match each to its definition.

Identify the type of confounding for each description

- Not allowing smokers into the study
- Using logistic regression
- Creating sub-analyses for different groups
- Identifying a control with the same characteristic as a case
- Assigning subjects into different treatment groups by chance
- Performing age adjustment

Restriction  Multivariate analysis  Stratification  Matching  Randomization  Standardization

Try this drag and drop exercise to see if you can correctly place each means of controlling for confounding into the correct oval: ways of controlling in the design phase and ways of controlling in the analysis

Try again. Remember there are three methods for design and three for analysis.

Here is the correct answer. Please review the slides to make sure you understand all of these.
Summary

- Defined a confounder
- How to identify confounding
- My personal confounding story
- Learned 6 ways of controlling confounders