This module is going to present important information about cohort studies. Since we already reviewed case-control studies, I will also use this opportunity to contrast cohort studies with this other design.

The definition of a cohort study is that it is a study in which two or more groups of people that are free of disease and that differ according to the extent of exposure (e.g. exposed and unexposed) are compared with respect to disease incidence. While this definition says two or more groups, usually we just look at two groups for simplicity in presenting the concept. So while we will talk about smokers vs. non-smokers, you might have a third group, for example former smokers. The principles are the same. It is just easier to present only two groups.

Cohort studies are the observational equivalent of experimental studies but the researcher cannot allocate exposure; he must locate a natural experiment to observe the relationship between the exposure and disease.
In this example, we are going to consider the exposure as smoking cigarettes and the disease as heart disease. Exposure is identified by a picture of a cigarette case and non-exposure is identified by having a cigarette case with a blue X through it. People with heart disease have a red X and those without heart disease do not have a red X. So we want to know if people who smoke are more likely to develop heart disease than those who did not smoke.

Start by identifying your exposed subjects. These are the people who smoke cigarettes (cigarette case). Then identify your non exposed subjects. These are people who do not smoke (cigarette cases with a blue X).

Identify the numbers of smokers who developed heart disease and the number of smokers who did not.

Identify the number of nonsmokers who developed heart disease and those who did not.

Compare the rate of disease between the two groups.

You can see there are some similarities between this study and the case-control studies we learned about previously. But the biggest difference is that in the case control study we started by looking at sick people (cases) and compared them to healthy people (controls) and then looked back to exposure. In the cohort study, we start by looking at people who are exposed (smokers) and those who are not exposed (non-smokers) and then we follow them over time to identify heart disease.

So essentially cohort studies want to identify the association between an exposure and a disease. The ultimate goal is to identify the cause of a disease and ultimately be able to prevent the disease. It is an observational study in that we observe what people choose to do. Researchers do not influence what the participants do but simply monitors and records it. The key factor in identifying the type of study is to determine how people are selected for the study, and in a cohort study, people are selected based on exposure. We will talk a bit more about how that is done. A key factor of cohort studies are that people who are enrolled are free of the disease and they are followed over time to see if they develop disease. This may be from the present...
to the future but it can also be at different times. I will explain that in the lecture. Since we are identifying new cases of disease, we will be looking at incidence. Remember I told you that when we studies incidence and prevalence it would come back to haunt us again. Since we are looking at incidence data, we will compare the incidence in the exposed with the non-exposed by calculating a measure called relative risk. I will show you this later in this lecture.

See if you can correctly identify who belongs in the different groups at the start of a cohort study. The red X indicates disease and exposure is working in a medical care position (white coats). The red button provides more on the directions.

Again, since nothing is as simple as it should be, we have a number of decisions to make when selecting the exposure group. If I did not tell you before, the correct answer to almost every epidemiology question is “It depends”. So when I ask what is the best way to identify the exposure group, the answer is it depends on the hypothesis and the feasibility of identifying people. Sometimes the choice of an exposed group is obvious. If we want to look at the impact of a certain chemical or environmental exposure, we might want to look at people with jobs that expose them to that exposure.

There have been a number of studies among car radiator repair workers looking at the health risks of environmental exposures in the job. If you click on the icon, you can see a graph from an Australian study that looked at lead exposure by different occupational groups. Push return to go back to the main slide.

Groups undergoing a particular medical treatment may be studies. We already discussed the study of DES exposure in which the children of women who were given DES during pregnancy were followed to evaluate other health outcomes. And I mentioned a study that found people with radiation treatment for
cancer were found to have higher risk of brain cancer later.

We sometimes study groups with unusual dietary or lifestyle factors but these can be challenging as these groups may differ on a number of factors, not only on what we are looking at. For example, vegetarians may have lower smoking rates than meat eating individuals, and there may be other differences as well.

There have been a large number of studies on certain professional groups. One of the best known ones is The Nurses Health Study. This study has been conducted in several waves, and has more than 200,000 participants to date. I put the link into the study. If you follow the link and click on Findings: Some Highlights, you can answer the trivia question. Click on return to go back to the main slide.

Students are common groups to study as are alumni of colleges. And one of the most famous cohort studies is Framingham Heart Study, conducted in Framingham Massachusetts. This study was begun in 1948 and is still providing important health information today. I posted a link to their homepage if you wanted to know more about it.

When you identify the exposed group, it can be challenging to find a group with a rare exposure. As I said before since occupational groups generally have higher levels of exposure to certain chemical and environmental risk factors, they may be a better choice for studies of rare exposures. You may be limited in the information available to you, as you might be using work records. Sometimes people use job title as a proxy for exposure. You have more options with common exposures as there are more places you can find exposed people. You are able to look for the population to study that has more accurate and complete exposure data. You just have more options as to who you can include in the study. The Framingham Hearst Study that I gave you a link to earlier followed a general population of residents of the Framingham, MA community.
Just like finding controls in case control studies, the choice of the comparison group can be fairly challenging. Basically, you want your comparison group to be a similar to the exposed group as possible with respect to all other factors except the exposure. This is easier said than done. So the ideal comparison group is essentially the exact individuals who were in the exposed group if they had not been exposed. You might remember we said a similar thing about controls. We wanted them to be people we would have selected as cases if they had the disease. Since we cannot have the same person as exposed and unexposed, then the goal of epidemiologists is to select different sets of people who are as similar as possible. In some studies this is easier than others.

Basiclly there are two types of comparison groups: Internal and external. Internal comparison groups come from the same populations. People are categorized as exposed and non-exposed. Internal groups are likely to be fairly similar as both the exposed and non-exposed subjects came from the same population. There may be other differences between them though as sometimes people with different exposures share other characteristics.

Examples of internal comparison groups include people in the Framingham Heart Study who did not have the exposure of interest, e.g., non-smokers or nurses who did not use oral contraceptives in the Nurses Health Studies. These larger studies collect a lot of information on the study subjects and depending on the analysis, a person may be counted as opposed and non-exposed. For example, a nurse in the Nurses Health Study who smokes but also exercises regularly, may be considered exposed if the study is looking at smoking but non-exposed if the study is looking at sedentary lifestyle. Smaller cohort studies tend to include fewer different exposure groups, and in fact, many cohort studies only look at one exposure. When you use external comparison groups, it can be more challenging to find a similar non-exposed group. Again, occupational studies illustrate some of the problems. You might choose as a non-exposed group people working in a different industry, asbestos textile workers versus cotton textile workers for example. But often it can be hard to find another occupational group that is similar. In this case, researchers might use people with a different
job title at the same company. But the problem with this approach is secretaries or office staff may have very different lifestyle experiences than line workers in a factory.

A third option is to use general population data. For example, we have national and state rates of many cancers, and we might compare these to the people in our study to see if our subjects have a higher rate of these diseases than the general population. Use pre-existing data from the general population as the basis for comparison. General population is commonly used in occupational studies. We might compare rates of certain cancers between our population and general population statistics. The problem is there is a bias called the healthy worker effect, in which people who work are generally healthier than the general population, as less healthy people may not have jobs. Thus, general statistics are often based on a less healthy population than a population of workers.

This mini test will give you a chance to be sure you are comfortable with the material we covered. Select the boxes that accurately describe cohort studies. If you get it wrong review the slides and try again.

There are three different types of cohort studies. They are differentiated by timing but they all select people based on exposure. Click on each one to learn more about its features.

In the prospective cohort study, the researcher enrolls subjects from the present moving forward. People who currently do not have a disease are identified by exposure and followed into the future to identify the incidence of disease. Researchers can collect information on exposure through many different means as they can actively collect this information from the participants. The advantage of this study design is that researchers have excellent control over data collection but the study takes time to conduct as you have to wait for the disease to
occur. This may not be feasible for diseases like cancer in which the latency period (time from exposure to disease) is very long.

In the retrospective cohort study (also known as a historical cohort study), the exposure and disease has already occurred. This study differs from a case-control study because in the cohort study people are identified based on exposure, and after they are identified into the study, the outcomes is determined. In a case-control study remember that people are selected based on their disease status and then we look back to determine exposure. Retrospective cohort studies are frequently done in occupational studies. For example, researchers conducted a retrospective conduct study to identify the risk of cosmic radiation among female flight attendants as female flight attendants may have a higher risk of breast and other cancers than the general population because of routine exposure to cosmic radiation. The data on exposure could be obtained through company records or interviews with their flight attendants.

The ambi-directional cohort studies contain aspects of prospective and retrospective studies. People are selected by exposure which has already occurred. Some of the study outcomes have occurred prior to the start of the study (like a prospective cohort study) but some outcomes have not yet occurred and subjects will be followed into the future.

The common features of all these designs are that people are selected by exposure and followed over time (real or past, depending on the design) to identify incidence of disease in the future.

We can obtain data on our exposures and outcomes from many different sources. There are advantages and disadvantages to each. And in some instances you may be limited in what you can do. Pre-existing records are often the only options for retrospective cohort studies. The greatest challenge is they may not have enough information on other factors, called confounders, which may be associated with disease and exposure. For example, if we are looking at the association of radiation exposure and lung cancer, work records may have excellent data on exposure but not have any information on smoking. Pre-existing data that is commonly used includes birth and death certificates.
When we use self-reported data, we are able to design data collection instruments that ask exactly what we want to know, but we are subject to the accuracy of people’s responses. It is also more costly to administer surveys than to relay on pre-existing data. Another challenge includes identifying people for your study and being able to obtain their information. Phone surveys are much less useful than they were in the past.

Physical exams and biological testing can provide the most accurate data but there is usually much more expense involved. Plus people may not want their blood drawn or feel what is asked of them is invasive or uncomfortable. You would be more likely to have a biased sample of participants as there may be differences between those who agree and disagree to participate.

There are a number of strengths and weaknesses in cohort studies. As you review them, you may notice many are the opposite of those we listed for case-control studies.

**Strengths**
- Efficient for rare exposures, diseases with long induction and latent period
- Can evaluate multiple effects of an exposure
- If prospective, good information on exposures, less vulnerable to bias, and clear temporal relationship between exposure and disease

**Weaknesses**
- Inefficient for rare outcomes
- If retrospective, poor information on exposure and other key variables, more vulnerable to bias
- If prospective, expensive and time consuming, inefficient for diseases with long induction and latent period
In a cohort study, you compare the risk of disease among those who are exposed to the risk of disease among those who are not exposed. It is called the relative risk, and it is actually pretty simple to calculate. Think back to Module 2 when you learned incidence. Remember the incidence of disease is the number of new cases of disease in a population. The relative risk essentially compares the incidence of disease among the exposed to the incidence of disease among the non-exposed. Let’s look at our very familiar 2 by 2 table. If you need an explanation of the parts of the 2 by 2 table, click on the information button.

Audio, Cell A is where people have the exposure and the disease. Cell B is where people have the exposure but not the disease. Cell C is where people do not have the exposure but have the disease. Cell D is where people do not have the exposure and do not have the disease.

If you look across the table, we want to first identify the risk of disease among everyone who was exposed (\( \frac{a}{a+b} \))

We then identify the risk of disease, among everyone who was not exposed (\( \frac{c}{c+d} \))

We then divide the risk of disease among the exposed by the risk of disease among the unexposed to obtain the relative risk.

Now let’s try an example.

We do a hypothetical cohort study that compares the risk of low birth weight among women who smoke with those who do not smoke. We obtain 100 smokers and 200 non-smokers from women attending Tampa prenatal clinics. We followed them over pregnancy to see if they had a low birth weight baby. Of the 100 smokers, 20 had a low birth weight baby. Of the 200 non-smokers 10 had a low birth weight baby. Put 10 in cell C. 200-10=190 which goes in cell D.

Let’s fill in the table. We put 20 in cell A, the women who smoke who have a low birth weight baby. 100-20= 80 which is the number of exposed people with a normal weight baby. Of the 200 non-smokes 10 had a low birth weight baby. Put 10 in cell C. 200-10=190 which goes in cell D. Now we are ready to do the math.
If you look across the table, we want to first identify the risk of disease among everyone who was exposed \((a/a+b)\).

We then identify the risk of disease among everyone who was not exposed \((c/c=d)\).

We then divide the risk of disease among the exposed by the risk of disease among the unexposed to obtain the relative risk.

So we have 20 divided by 100 over 10 divided by 200 which equals 0.2 divided by 0.05 which equals 4.

This means that women who smoke during pregnancy have 4 times the risk of having a low birth weight baby than women who do not smoke during pregnancy. We will discuss the interpretation shortly but first I want you to do the analysis.

Now you try it. Fill in the 2 by 2 table and calculate the relative risk for this study. If you have trouble take the time to review the notes and try again.

Studies measuring person-time have some differences in the calculations but the general formula is the same. Instead of number of people in the denominator, we use person time in its place.

Person time is obtained by adding up all the time each person is in a study (in days, weeks, years, etc.). This is done for the exposed and non-exposed and the calculations are based upon a more accurate measure of the study population, especially for studies in which there is differential follow up of the participants. This is often true of long studies where loss to follow up is a real concern. You can see the video on person time as it explains this concept. We also covered this concept in Module 2.

This slide shows the calculation of incidence rate for a population. X indicates the person has the disease and the lines show how long each subject was followed in the study. There were two people who developed the disease during 37.5 person years of follow-up, resulting in an incidence rate of 0.053.
So when you calculate the relative risk, if you have incidence, you use the formulas written in purple and if you have incidence rate, you use the formulas in light blue. The numerators are the same for each calculations and the denominators are number of person (incidence) and person time (incidence rate).

http://www.slideshare.net/sumizin/2epidemilogic-measures

Patwari video odds ratio vs relative risk
https://www.youtube.com/watch?v=hOtoV2Kjb0o

Patwari incidence rates
https://www.youtube.com/watch?v=bbIl7jyD1KA

So how do we interpret the relative risk? Remember in the relative risk, we are comparing new cases of disease, so we are looking at incidence. One study shows association but cannot prove cause and effect. In this slide I am talking about the exposure increasing or decreasing the rate of disease but in reality, I may just be identifying an association and there is no causality. But for the purpose of this slide, let's assume causation. One reason cohort studies are so strong in identifying causation is that we know the exposure occurred before the disease because no one had the disease when we began the study.

- Relative Risk > 1, then exposure increases the risk of disease
  If the exposure increases disease occurrence, then the rate of disease in the exposed group will be higher than the rate of disease in the comparison group.
- Relative Risk<1, then exposure decreases the risk of disease
  If the exposure decreases disease occurrence, then the rate of disease in the exposed group will be lower than the rate of disease in the comparison group.
- Relative risk =1, then the exposure has no effect on disease
  If the exposure has no effect on disease occurrence, then the rate of disease in the exposed and comparison groups will be the same.
This ends our discussion of cohort studies. In the last two lectures we reviewed case-control studies. I want to give you a chance to compare and contrast these two studies so I set up a quiz in which you can drag attributes of the study designs into each area. I trust you will succeed but if you have any difficulty review the materials and try again. If there is anything you don’t understand, please let I or the TAs know.