Slide 1: Introduction
Hello and welcome to Module 2: Fundamentals of Epidemiology. This presentation will focus on observational studies. My name is Jeffrey Bethel, assistant professor of epidemiology at East Carolina University, Brody School of Medicine.

Slide 2: Acknowledgements

Acknowledgments
APTR wishes to acknowledge the following individual that developed this module:

- Jeffrey Bethel, PhD
  Department of Public Health
  Brody School of Medicine at East Carolina University

This education module is made possible through the Centers for Disease Control and Prevention (CDC) and the Association for Prevention Teaching and Research (APTR) Cooperative Agreement, No. 5U50CD300860. The module represents the opinions of the author(s) and does not necessarily represent the views of the Centers for Disease Control and Prevention or the Association for Prevention Teaching and Research.

Slide 3: Presentation Objectives

Presentation Objectives
There are four objectives for this presentation. They are to recognize criteria for initiating various observational studies; identify design components of various observational studies; to calculate and interpret outcome measures in various observational studies; and recognize advantages and disadvantages to various observational studies.
Slide 4: How did we get from here...

So how did we move from a culture of healthcare professionals promoting cigarette smoking as you see here in these ads where nurses, dentists and physicians are promoting smoking,

Slide 5: and here...

And similarly, Santa Claus, babies and more physicians promoting smoking,

Slide 6: to here?

To here, in which the Surgeon General’s Warning appears on all packages of cigarettes.
One of the principal driving factors of the move from a culture where smoking was openly promoted in the media to a culture in which the Surgeon General’s Warning now appears on all packages of cigarettes was through a series of epidemiological studies, specifically observational studies, linking smoking to lung cancer. So in this presentation I will cover observational studies, which are different from experimental studies in that they are natural studies in which participants self-select exposure such as smoking. The researcher does not allocate treatment or exposure to study participants, which is what is done in experimental studies. Experimental studies are covered in another presentation.

Observational studies are used to study a wider range of exposures than experimental studies. These “natural” studies or experiments get around aspects of experimental studies which are not feasible. For example, an experimental study could not examine the effect of smoking during pregnancy on birth outcomes. However, observational studies can.

Observational studies can also provide information about the etiology of a disease, such as risk factors by examining occupational and environmental exposures. They also provide information about prognosis. That is, what factors predict mortality or disability? And in the realm of healthcare delivery, what factors predict health related outcomes such as increased quality of life.
Module 2: Observational Studies

Slide 10: Cohort Studies

So as previously mentioned, this presentation will address three types of observational studies. The first type we will cover is cohort studies.

Slide 11: Epidemiological Study Designs

Here is an informative figure showing the hierarchy of epidemiological study designs. We have validity on the y-axis and cost on the x-axis. You’ll see that intervention trials or experimental studies sit at the top of the hierarchy with the highest validity but also with the highest cost. The three observational studies discussed in this presentation are next in the hierarchy. To sum this figure up: you get what you pay for by choosing an epi study design. However, choice of study design type is dependent on several factors. Including resources available, level of knowledge on the topic, etc.

Slide 12: Cohort Studies

Again, first up are cohort studies. The word cohort is derived from Latin and means warriors or a group of persons proceeding together in time. In ancient Rome, a cohort was one of ten divisions of a legion, the major unit of the Roman army. Once a cohort was formed, no new soldiers could join and they stayed together throughout their service or until death. In general in cohort studies, investigators recruit exposed and unexposed participants who are followed longitudinally through time and observed for the disease or health outcome of interest. The incidence of disease is then compared between the two groups, the exposed and the unexposed. Cohort studies are also called incidence or follow-up studies.
Slide 13: Cohort Study Design

Here is a figure depicting what was described in a previous slide. Study participants are classified or selected by exposure status through a non-random assignment. That is, participants self-select their exposure status such as smoking cigarettes or drinking alcohol. And they are followed over time to determine disease incidence. The incidence of disease is then compared between exposed and unexposed groups.

Slide 14: Prospective Cohort Study Design

There are two types of cohort studies that are classified by when the study population is identified. In prospective cohort studies, the study population is identified at the beginning of the study and then followed prospectively through time.

Slide 15: Retrospective Cohort Study Design

In retrospective cohort studies, the study population is identified in the past and then typically through records, exposure and disease status determined. Retrospective studies are used for diseases with a long latent period such as cancer.
Slide 16: Cohort Studies – When to Initiate?

Since cohort studies are often costly due to the sample size required and the long follow-up period needed, careful consideration should be given to determine when they should be initiated. First, there should be evidence from less expensive studies down further on the hierarchy list, linking the exposure to the disease or the health outcome of interest. Another reason a cohort study could be initiated is when there is a new agent requiring monitoring for possible association with several diseases, such as diseases associated with oral contraceptives or hormone replacement therapy.

Slide 17: Cohort Studies – Assembling Cohort

Once a researcher decides to conduct a cohort study, the next step is to assemble the cohort. There are two ways of doing this. First, researchers can select a cohort based on the exposure status, such as exposed and unexposed. For example researchers could identify vaccinated and unvaccinated children to examine the link between certain vaccines and autism. And follow this group of children over time to compare the incidence of autism between the exposed and unexposed group. More commonly used method is to select a defined population such as participants in a given occupation. For example nurses or physicians. Or in a given geographic area such as a county or city. Exposure status can then be determined among the participants as they are followed longitudinally. Then the outcomes are assessed and compared between the various exposure groups. The study design is fundamentally the same. That is, to compare the outcome between the exposed and unexposed groups.
Slide 18: Cohort Studies – Defined Populations

Here are examples of cohort studies that used defined populations. That is the Framingham Study, the Nurse’s Health Study, Women’s Health Initiative, and the Health Professional’s Follow-up Study. They use occupations as well as geographic areas; for example, the Framingham Study recruited participants from Framingham, Massachusetts. I urge you to click on the links on this slide to explore the studies a little further.

Slide 19: Cohort Studies – Inclusion Criteria

Regardless of how the cohort is formed, researchers must develop specific inclusion criteria to determine who is eligible to be in the study. For example in the Women’s Health Study, which was a randomized trial of low dose aspirin in the primary prevention of cardiovascular disease in women, researchers identified these various inclusion criteria: Participants must be greater than or equal to 45 years of age; female; no history of coronary heart disease, cerebrovascular disease, cancer, or other major chronic illness; no history of side effects to any of the study medications; were not taking any of the meds listed on the slide more than once per week, and not taking anticoagulants or corticosteroids. Developing these specific inclusion criteria before the study begins removes subjectivity and it allows the study to be reproducible by other researchers.
Slide 20: Cohort Studies – Assembling Cohort

So here are some potential sources when assembling a cohort. Again the occupational cohorts, like the Nurse’s Health Study, the advantage of this source is that they are easy to identify and you can assume there is an adequate number of exposed individuals. You could use prepaid health plan members. Again, gives ease of identification and potential access to health records. Other sources include schools or military. Again, ease of identification and follow-up.

Slide 21: Cohort Studies – Determining Exposure

Once the cohort has been formed, the investigator needs to determine exposure status of the participants. Particularly if a defined population has been recruited. There are several ways to assess exposure; these sources can be used in combination with each other. Questionnaires could be used at baseline to determine exposure status such as age, race/ethnicity, smoking status, alcohol use, and medical history. Laboratory tests can be used to determine exposure status, such as blood lead levels. Physical measurements, such as height and weight to determine BMI. Special procedures such as bone density. Or existing records, in the context of an occupational cohort.

Slide 22: Cohort Studies – Selecting Comparison Group

Some studies of higher validity have a control or comparison group. Ideally we want to compare the outcome among an individual in the exposed group to what the outcome would have been if that same individual had not been exposed. This is counterfactual and impossible. However, we can compare the outcome among the exposed group to the outcome in a substitute population and the validity of the inference that we’re finding depends upon finding a valid substitute population, or a comparison population.
Slide 23: Cohort Studies – Internal Comparison Group

There are two types of comparison or substitute populations that we can use. First is an internal comparison group. These participants come from the same sample as the exposed group but they do not have the exposure. For example if participants came from the same town, the same occupation, HMO, school, military except they do not have the same exposure. This is typically used when the cohort is identified through a defined population. Strengths: this group is most comparable to the exposed group. Weaknesses: it may be difficult to identify and in some cases this population probably has similar exposures.

Slide 24: Cohort Studies – External Comparison Group

An external comparison group can also be used. Often the cohort is identified based on exposure status upfront. The comparison or the unexposed group should be susceptible to the same selective factors as the exposed. The strengths of this type of group is that the group is accessible and can provide stable data. Weaknesses: lack of comparability with the exposed group is sometimes seen. Results may suffer from the healthy worker effect, and that’s when the general population includes ill and well people and the occupational group generally includes well people only. People who are working are generally healthier than the people who are not working. Lastly, data on key variables may be missing in this external comparison group.
Slide 25: Cohort Studies – Measuring Disease

Now that the exposure status has been assessed among study participants, they are followed longitudinally and observed for the outcome of interest. The definition of the outcome must be defined before the study begins and the measurement methods must be comparable, use uniform criteria and the ascertainment as complete as possible for both the exposed and the unexposed participants. That is, disease ascertainment should be done by persons unaware of the exposure status of the participants. And that’s typically done by blinding the data collectors as to the exposure status of the participants. And there are various sources of information to measure disease. You could use death certificates if the outcome is death. Hospital records if hospitalization is required. You can use disease registries such as cancer or birth defects registries. Physician records, physical exam such as the Framingham Study used, among others. Laboratory tests or even questionnaires and have participants self-report their disease status.

Slide 26: Cohort Studies – Framingham Study

Here’s a quick aside on the Framingham Study, which was mentioned earlier in this presentation. Framingham Study is an example of a cohort study, which utilized a defined population based on location, that is, Framingham, Massachusetts. So in this study the participants were recruited and observed over time to determine exposure status and later the outcome of interest, namely cardiovascular disease. The exposures that they were interested in were smoking status, cholesterol levels, age, and alcohol used. And this cohort was identified nearly 50 years ago and there are multiple generations now.
Slide 27: Cohort Studies – Framingham Study Hypotheses

The Framingham Study had several hypotheses and generated much of what we know about risk factors for coronary heart disease (CHD) now. These hypotheses included: the incidence of CHD increases with age and occurs earlier and more frequently in men; persons with hypertension develop CHD at a greater rate than those who are normotensive; elevated blood cholesterol level is associated with an increased risk of CHD; and tobacco smoking and habitual use of alcohol are associated with an increased incidence of CHD.

Slide 28: Cohort Studies – Analytic Approach – Relative Risk

Now that the outcome measures have been defined and measured the data are complete it’s time to analyze the data. The principle measure of association used in cohort studies is the relative risk. Relative risk is simply the incidence of disease in exposed group divided by the incidence of disease in the unexposed group. For example, in a study examining the association between hormone replacement therapy and colorectal cancer among women, the relative risk would be the incidence of colorectal cancer in women who took hormone replacement therapy divided by the incidence of colorectal cancer in women who did not take hormone replacement therapy. Again, this is the measure of association used in cohort studies for deriving a causal inference.
So here’s the approach to calculating a relative risk. Set up your 2 by 2 table. You see here we have a row for the exposed group and a row for the unexposed group. And for each we indicate the number who develop disease, the number who did not develop the disease and the total number of exposed. For the unexposed group we indicate the number who develop the disease, the number who do not develop the disease and the total number of unexposed. So for each group, exposed and unexposed, we calculate the incidence of disease. For example in the exposed group the incidence of disease is calculated by dividing a by a+b. That is the number who are exposed and develop the disease divided by the total number of exposed. For the unexposed group you calculate the incidence by dividing c by c+d. That is, the number of the unexposed who developed the disease divided by the total number of unexposed. And again, the relative risk is the incidence of disease in exposed divided by the incidence of disease in the unexposed.

So how do we interpret this resulting relative risk? If the relative risk is 1 that means the exposure is not associated with the disease. That is, the incidence of disease in the exposed is equal to the incidence of the disease in the unexposed. If the relative risk is greater than 1, the exposure is associated with an increased risk of disease. The incidence of disease in the exposed was greater than the incidence of disease in the unexposed. If the relative risk is less than 1 the exposure is associated with a decreased risk of disease, that is, it is protective. The incidence of disease in the exposed was less than the incidence of disease in the unexposed.
Slide 31: Smoking and Coronary Heart Disease

So here’s an example involving smoking and CHD. So we set up our table, and we have 3,000 smokers in this cohort study and 5,000 non-smokers. Of the smokers, 84 developed the disease. Of the non-smokers, 87 developed the disease. We calculate our incidence of CHD among smokers by dividing 84 by 3,000 and we get 28 per 1,000. Then we calculated the incidence of CHD among the non-smokers. And we do that by dividing 87 by 5,000 and we get 17.4 per 1,000. We then divide the incidence of disease in the smokers by the incidence of disease in the non-smokers which is 28.0 divided by 17.4 and we get 1.61. That is the incidence of CHD among smokers was 1.61 times as great as in non-smokers. Or, smokers were 1.61 times as likely to develop CHD as non-smokers.

Slide 32: Cohort Studies – Other Analytic Approaches

Other analytic approaches include a life table, constructing a Kaplan-Meier plot, calculating incidence proportion or a hazard ratio, or creating a multiple logistic regression model.

Slide 33: Calculating Relative Risk

So let’s work through an example. Here’s a cohort study examining smoking and bladder cancer. There were a total of 1,000 people on the island. 400 were smokers and 600 were not. 50 of the smokers developed bladder cancer. 15 of the non-smokers developed bladder cancer. Calculate and interpret the relative risk.
Module 2: Observational Studies

Slide 34: Smoking and Bladder Cancer
The first step is to set up our blank 2 by 2 table. We have a row for smokers and a row for non-smokers and we’ll begin to fill it in.

<table>
<thead>
<tr>
<th>Smoking and Bladder Cancer</th>
<th>Bladder Cancer</th>
<th>No Bladder Cancer</th>
<th>Totals</th>
<th>Incidence of Bladder Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoke</td>
<td>50</td>
<td>350</td>
<td>400</td>
<td>0.125</td>
</tr>
<tr>
<td>No Smoke</td>
<td>15</td>
<td>585</td>
<td>600</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Slide 35: Smoking and Bladder Cancer
So as the text describes there were 400 smokers, 50 developed bladder cancer and 350 did not. Out of the 600 non-smokers, 15 developed bladder cancer and 585 did not.

<table>
<thead>
<tr>
<th>Smoking and Bladder Cancer</th>
<th>Bladder Cancer</th>
<th>No Bladder Cancer</th>
<th>Totals</th>
<th>Incidence of Bladder Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoke</td>
<td>50</td>
<td>350</td>
<td>400</td>
<td>0.125</td>
</tr>
<tr>
<td>No Smoke</td>
<td>15</td>
<td>585</td>
<td>600</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Slide 36: Smoking and Bladder Cancer
The next step is to calculate the incidence of bladder cancer among smokers and non-smokers. The incidence of bladder cancer among smokers is calculated by dividing 50 by 400. The incidence of bladder cancer among non-smokers is calculated by dividing 15 by 600.

<table>
<thead>
<tr>
<th>Smoking and Bladder Cancer</th>
<th>Bladder Cancer</th>
<th>No Bladder Cancer</th>
<th>Totals</th>
<th>Incidence of Bladder Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoke</td>
<td>50</td>
<td>350</td>
<td>400</td>
<td>0.125</td>
</tr>
<tr>
<td>No Smoke</td>
<td>15</td>
<td>585</td>
<td>600</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Slide 37: Smoking and Bladder Cancer – Relative Risk
Finally we calculate the relative risk we divide the incidence of bladder cancer among smokers by the incidence of bladder cancer among non-smokers which is .125 divided by .025 and we get 5.0. This means that the incidence of bladder cancer is 5 times as great in smokers as in non-smokers.

\[
\text{Relative Risk} = \frac{50}{400} \div \frac{15}{600} = 5.0
\]

Interpretation:
Incidence of bladder cancer is 5 times as great in smokers as in non-smokers.
Slide 38: Cohort Studies - Disadvantages

There are several disadvantages of cohort studies that should be mentioned. One is it is inefficient for the evaluation for rare diseases. You’ll find in a prospective cohort study that after you follow your participants you’ll have very few outcomes or diseases in which to compare the incidence between your two groups. If the outcome has a long latent period, the study can take a very long time, many many years. Next, cohort studies are generally quite expensive, given they have a long follow up period and require a large sample size. When conducting a retrospective cohort study you’re dependent on the availability of records to assess exposure status and disease status of participants. Lastly, the validity of your results can be seriously affected by losses to follow-up. Given you are following your cohort through time, including several years, participants may drop out. And a high degree of loss to follow-up can affect the validity of your results.

Slide 39: Cohort Studies - Advantages

However there are several advantages to cohort studies. They are useful when the exposure is rare. Which is particularly the case when you are recruiting participants based on exposure status. You can examine multiple effects of a single exposure status, that is, examine multiple outcomes of participants. If the study is prospective you can minimize the bias in the ascertainment of exposure. A prospective study allows you to examine the temporal relationship between exposure and disease. Given a cohort study begins with disease free people, you are ensuring that the exposure preceded disease. Similarly that allows you to directly calculate the incidence of disease in the exposed and unexposed groups which allows you to calculate the relative risk.

Slide 40: Case-Control Studies

So that concludes cohort studies. Next up is case-control studies.
**Module 2: Observational Studies**

**TRANSCRIPT**

**Slide 41: Epidemiologic Study Design**

You’ll see case-control studies sit further down on the hierarchy than cohort studies. They are less expensive than cohort studies and they may produce slightly less valid results. The decreased cost comes from the fact that there is no follow up of participants through time.

**Slide 42: Case-Control Study Design**

Here’s a figure depicting the general design of a case-control study. Participants are identified according to outcome or disease status. That is, the investigator selects cases with the disease and appropriate controls without the disease. The investigator then obtains data regarding past exposure to possible etiologic factors and makes comparisons between these groups. So in a case-control study, investigators cannot determine the incidence of disease as it was possible in a cohort study.

**Slide 43: “TROHOC” Studies**

Case-control studies have been referred to as TROHOC studies, cohort spelled backwards, because their logic seemed backwards and they seemed more prone to bias than other studies, particularly cohort studies. Participants were selected by disease status and then classified by past exposure status and that just seemed backwards. However, case-control studies are a logical extension of cohort studies and they can be seen as a more efficient way to learn about associations between exposures and diseases.
Next, when exposure data are difficult or expensive to obtain. And this is the case because you are collecting exposure data on fewer individuals, saving you time and money. Third, when a rare disease is being investigated it is easier to identify cases rather than waiting for them to occur. Similarly, if the disease has a long induction or latent period. Since there is no follow up in a case-control study, you are avoiding waiting around for years for cases to occur. Lastly, when you’re dealing with a dynamic underlying population. In a cohort study it can be difficult to identify and track this population so it is wiser to conduct a case-control study in which there is no follow-up.

Like the cohort studies, participants need to be identified, but in a case-control study participants are identified by disease or outcome status. To identify appropriate cases for the study, the definition for a case must first be established. In general the definition should lead to an accurate classification of diseased and non-diseased individuals. The definition should be broad enough so as to include all cases but narrow enough so that only true cases are included. Lastly once the definition is established it should be applied uniformly to all people in your target population.

In this case-control study examining the link between post-menopausal hormone replacement therapy (HRT) and endometrial cancer the following case definition was used. Black or White women (including Hispanic women self-identifying as Black or White) aged 50-79 years, who were residents in the contiguous nine-county Philadelphia, Pennsylvania, region at the time of diagnosis and newly diagnosed with endometrial cancer between July 1, 1999, and June 30, 2002.
So where do researchers find cases to participate in their studies? There are several sources that can be used. All cases of a disease in the population can be used depending upon the number of cases and the size of the population. This is often resource intensive. You could also take a representative sample of all cases in a population. You could use disease registries such as cancer or birth defects registries. This source depends on the availability of an appropriate registry for a population. All hospitals or a particular hospital or health system in a community or region could be used. However consideration must be given as to whether the particular hospital or health system is representative of all cases in a region. Lastly, physicians’ records could be used but this also highly resource intensive.

Now that cases have been defined and identified, investigators need to identify and recruit controls as a comparison group. The purpose of controls in a case-control study is to provide information on the exposure distribution in the source or reference population. Next controls must be identified independently of exposure status. Controls are a sample of the population that gave rise to the cases. They must satisfy what’s called the “would criteria”. A member of the control group who gets the disease “would” end up as a case in the study.
Case-Control Studies 
Sources of Controls

- General population
  - Used when cases are identified from well-defined population (e.g. residents of a geographic area)
  - Sources: RDD, voter reg lists, tax lists, neighborhood
  - Advantage: generally more representative of non-diseased with respect to exposure
  - Disadvantage: not as motivated, potentially lower data quality

Slide 49: Case-Control Studies – Sources of Controls

We can also talk about sources of controls for a case-control study. One source is the general population. And these are used when cases are identified from a well-defined population. For example the residents of a certain city or county. Specific sources include random digit dialing, voter registration lists, tax lists or neighborhood data. Advantage of this source is that they are generally more representative of non-disease people in the population with respect to exposure. However the disadvantage is that they are not as motivated and therefore they may provide potentially lower data quality. These controls are being contacted out of the blue as disease free people to participate in a study. Therefore they may not be motivated.

Case-Control Studies 
General Population Controls

Cases: active surveillance at 61 of 68 hospitals in 9 counties around Philadelphia
Controls: RDD controls were selected from the same geographic region as the cases

Slide 50: Case-Control Studies – General Population Controls

Back to the study examining hormone replacement therapy and endometrial cancer. The source of the cases was from active surveillance at 61 of 68 hospitals in 9 counties around Philadelphia. The controls were found from random digit dialing in the same area as the cases. The question is, does this group of controls satisfy the “would criteria”? Would a control, if they had come down with the disease, ended up as a case in this study? That is, a case identified through active surveillance at 61 of 68 hospitals in the 9 counties around Philadelphia.
Case-Control Studies

Sources of Controls

- **Hospital/Clinic**
  - Used when cases are identified from hospital/clinic rosters
  - Advantage: easily identified, readily available, more aware of prior exposure, same selection factors as hospitalized cases
  - Disadvantage: difficulty determining appropriate illness (unrelated to exposure and same referral pattern as cases)

Controls could be drawn from a surgical, orthopedic and medical unit of the same hospital. Specifically they could be patients with musculoskeletal and abdominal disease, trauma or other non-coronary conditions. You would not want to choose patients to serve as controls who have emphysema as it is related to smoking. One advantage of this type of control is that they are easily identified, they are readily available, since they are a diseased individual they are more aware of prior exposures, and the same selection factors were used as the cases to make them hospitalized. The disadvantage is that you have difficulty determining the appropriate illness to use as a hospital control. You need to make sure this illness or disease is unrelated to exposure and had the same referral pattern as the cases. Another advantage is that this type of control can certainly satisfy the “would criteria”. Would a control, if they had the disease, end up as a case in this study? Again the same selective factors move them toward this hospital where the cases were identified.

Slide 51: Case-Control Studies – Sources of Controls

Another source of controls is a hospital or clinic and they’re used when cases came from a hospital or clinic. For example, a study of cigarette smoking and myocardial infarction among women cases could be identified from admissions to a hospital coronary care unit.

Slide 52: Case-Control Studies – Sources of Controls

Another source is relatives, friends, classmates or coworkers. This is not as commonly used as the other types of control groups. Advantages include they’re motivated, they know the case, readily available in some circumstances, less expensive to identify, more similar to the cases in terms of neighborhood or social class, and they could be more representative of the healthy with regards to exposure. Disadvantage is that they may share exposures, for example alcohol use or the same occupation as the cases. Next, cases may be unable or may not want to nominate their friends or relatives to serve as disease free controls in this study.
Module 2: Observational Studies

TRANSCRIPT

Slide 53: Case-Control Studies – Matching

Since observational studies do not rely on randomization, cases and controls may differ in certain characteristics. So if a researcher finds an association between exposure and disease, is it due to differences between cases and controls other than the exposure of interest? Investigators can account for this by matching cases and controls on important variables such as age. That is, they can make cases and controls similar in terms of age. Matching can be done individually or at the group level. In individual matching, investigators select for each case one or more controls matched on a given variable. For example they match a control to a case within 5 years and/or they could match the control to a case based on gender. In group matching, controls are selected so that they have the same distribution of the matching variable as the cases. To do group matching, all cases must first be identified so investigators know the distribution of the given characteristic in the cases. For example, if 30% of the cases were female, researchers will pick a group of controls that is also 30% female.

Slide 54: Case-Control Studies – Matching

So in the HRT and endometrial cancer example, the controls were described as random digit dialing controls selected from the same geographic region as the cases, frequency matched to the cases on age (in 5-year age groups) and race (Black or White).

Slide 55: Case-Control Studies – Determining Exposure

Now it’s time to determine the exposure status of our cases and controls. Just like cohort studies researchers can determine exposure status among participants in a variety or a combination of ways. Again, questionnaires can be used, laboratory tests, physical measurements, special procedures, or existing records. Again one or a combination of these sources of exposure data can be used.
Slide 56: Case-Control Studies – Questionnaires

Back to our example again. In this study, telephone interviews, which averaged 60 minutes, were administered by trained lay interviewers with no knowledge of the study hypotheses. These lay interviewers assessed participants, that is, whether they used hormone therapy in the past.

Slide 57: Case-Control Studies – Analytic Approach – Odds Ratio

Now that exposure has been assessed for all participants, it is time to analyze the data. So like relative risk for cohort studies, odds ratios is the measure of association in case-control studies used for deriving a causal inference. So we have our data and it’s time to set up our 2 by 2 table. So in this table we have our cases and controls in the two columns and exposed and unexposed in the two rows. So in this 2 by 2 table the letter “a” denotes the number of exposed cases, “b” denotes the number of exposed controls, “c” denotes the number of non-exposed cases, and “d” denotes the number of non-exposed controls. So to get the odds ratio we first need to calculate the odds that a case was exposed and then the odds that a control was exposed and take the ratio of the two. The odds that a case was exposed is the probability that a case was exposed divided by the probability that a case was not exposed. The probability that a case was exposed is a divided by a+c, the number of exposed cases divided by the total number of cases. The probability that a case was not exposed is c divided by a+c, the number of non-exposed cases divided by the number of total number of cases. The a+c gets canceled out and we get a over c. Next the odds that a control was exposed and again the probability that a control was exposed divided by the probability that a control was not exposed. That is b over b+d, the probability that the control was exposed, divided by d over b+d. The b+d cancels and we get b over d. Now that we have the odds that a case was exposed as a/c, and the odds that a control was exposed as b/d, we take the ratio. That is, divide a/c by b/d, flip the denominator and we get a x d divided by b x c.
Slide 58: Smoking and Coronary Heart Disease

So now let’s apply that formula to some data. Here’s a study looking at smoking and CHD. We have our 2 by 2 table with our exposed and unexposed cases and controls. We can use this table to calculate an odds ratio of \( \frac{ad}{bc} \) and we get \( 84 \times 4913 \) divided by \( 2916 \times 87 \) and we get 1.63. This is interpreted as the odds that a person with CHD smoked is 1.63 times the odds that a person without CHD smoked.

Slide 59: When Does the Odds Ratio Provide a Good Estimate of Relative Risk?

We talked about relative risk in cohort studies, which uses incidence of disease to estimate risk of disease. However for a case-control study incidence of disease cannot be computed because we recruited participants based on disease status up front. Since we cannot do this, risk may be difficult to comment on. However, an odds ratio can estimate a relative risk when three conditions are satisfied. First the cases that are studied are representative, with regard to history of exposure, of all people with the disease in the population from which the cases were drawn. For example, if cases came from a certain hospital then cases identified in the study should be representative of all cases in the catchment area of the hospital with regard to exposure. Next, when the controls studied are representative, with regard to history of exposure, of all people without the disease in the population from which the cases were drawn. Lastly, when the disease studied does not occur frequently. This is what’s thought of as the rare disease assumption.

Slide 60: Case-Control Studies – Calculating Odds Ratio

So let’s work through an example of calculating an odds ratio. Suppose that a case-control study was conducted to evaluate the relationship between artificial sweeteners (AS) and bladder cancer. 3,000 cases and 3,000 controls were recruited. Among the cases, 1,293 had used AS in the past while 1,707 had never used AS. Of the controls, 855 had used AS and 2,145 had not. So let’s calculate and interpret the odds ratio.
Slide 61: Artificial Sweetener and Bladder Cancer

First thing we do is set up our blank 2 by 2 table. We have our cases and controls and exposed and unexposed.

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed to AS</td>
<td>1,293</td>
<td>855</td>
</tr>
<tr>
<td>Not Exposed</td>
<td>1,707</td>
<td>2,145</td>
</tr>
<tr>
<td>TOTAL</td>
<td>3,000</td>
<td>3,000</td>
</tr>
</tbody>
</table>

Odds Ratio = \( \frac{ad}{bc} \) = \( \frac{(1,293)(2,145)}{(855)(1,707)} \) = 1.90

Interpretation:
- OR = 1.90
- Odds that a person with bladder cancer used artificial sweeteners was 1.90 times the odds that a person without bladder cancer used artificial sweeteners.

Slide 62: Artificial Sweetener and Bladder Cancer

Next we use the description in the study to fill in our table. Of the 3,000 cases 1,293 have been exposed to AS. Of the 3,000 controls 855 have been exposed to AS. Calculate our odds ratio and we get 1.90.

Slide 63: Artificial Sweetener and Bladder Cancer

This odd ratio can be interpreted as the odds that a person with bladder cancer used AS was 1.90 time the odds that a person without bladder cancer used AS. That is, the odds that a case was exposed divided by the odds that a control was exposed.

Slide 64: Case-Control Studies – Disadvantages

Like all studies, case-control studies have advantages and disadvantages. The disadvantages include that we can only investigate one disease outcome at a time, that is inherent in the case-control study design. It is inefficient for rare-exposures. You cannot directly compute incidence rates of disease in exposed and unexposed because disease and non-disease are identified at baseline. The temporal relationship between exposure and disease may be difficult to establish given you are assessing the prior exposure of cases and controls and you don’t know when a person was diagnosed with a given disease, therefore you don’t know if the exposure preceded...
the disease. Lastly they are vulnerable to bias because they are retrospective. Specifically, recall bias may occur. The quality of data may differ between cases and controls since they are recalling their exposures from the past.

**Slide 65: Case-Control Studies – Advantages**

However there are advantages. One, they are efficient for rare diseases. Since you are recruiting based on disease status you can go out and find the relatively few cases of a rare disease. They are efficient for diseases with a long induction and latent period. Unlike cohort studies where you have to wait around for cases to occur, in case-control studies the cases have already occurred therefore there is no waiting around for diseases with a long induction period such as cancer. Next you can evaluate multiple exposures in relation to a disease. Although they all occur in the past, you can assess multiple exposures retrospectively. And since there is no follow up, like in a cohort study, case-control studies are relatively quick and inexpensive.

**Slide 66: Cross-Sectional Studies**

So that concludes case-control studies. The next study design type we will discuss is cross-sectional studies.

**Slide 67: Epidemiological Study Designs**

Cross-sectional studies are further down the hierarchy of epidemiologic study designs and we will discuss the reasons this is the case. However they do still serve a purpose.
Cross-sectional studies represent a snapshot in time. Exposure status and disease status of an individual are measured at one point in time. That is, disease prevalence in those with and without exposures or at different exposure levels are compared. Cross-sectional studies are very useful in health planning in that they do represent a snapshot in time, the current situation of a disease and an exposure and their association.

So here’s a figure representing the cross-sectional study design. We have our study population and simultaneously investigators gather data on exposure and disease status, producing four groups of participants: exposed with the disease, exposed without the disease, unexposed with the disease, and unexposed without the disease. Examples of cross-sectional studies include the various national and state-level surveys.

Like in cohort and case-control studies, investigators need to identify a study population in cross-sectional studies. Sometimes it is based on exposure of interest, if it’s readily identifiable. For example the prevalence of a disease in particular ethnic group or geographic area or occupational group could be identified. For relatively small numbers the entire population may be included, or a representative sample of this population can be included. For example an entire community or a random sample of households in the community could be included in the study.
Module 2: Observational Studies

Slide 71: Cross-Sectional Studies – Determining Exposure

Exposure status must be determined just like in cohort and case-control studies and similar sources are used. Questionnaires, records, lab tests, physical measures, and special procedures are used to assess exposure status of participants. One thing to keep in mind is that timing and duration of exposure is important to document, if possible, to relate to the onset of disease. Given this is a snapshot in time you can frame your exposure assessment in terms of when participants were exposed.

Slide 72: Cross-Sectional Studies – Measurement of Disease

Again, just like in other study design types, disease status must be measured among study participants. It can be determined by questionnaires if the disease status is simply symptoms, physical exam (joints for arthritis), special procedures (x-rays, lung function). For diseases with exacerbations and remissions such as asthma you need to ask participants if they are asymptomatic or if they’ve had symptoms in the past. Next, diagnostic criteria must be determined in advance and applied uniformly and systematically to all participants regardless of exposure status.

Slide 73: Cross-Sectional Studies – Analytic Approach

Once exposure and disease status have been determined it’s time to analyze the data. 2 x 2 tables can be developed and various measures can be calculated. One is the prevalence ratio and that is simply the prevalence of the disease in the exposed group divided by the prevalence of disease in the unexposed group. Next we can generate prevalence odds from the exposed and the unexposed and use those to calculate prevalence odds ratios. That is the ratio of the prevalence odds in exposed to prevalence odds in the unexposed.
Here’s an example of some data showing HIV infection and intravenous drug use (IVDU) among women in the New York State Prison System. Of the 136 IVDU, 61 were HIV+. Of the 338 women who were not IVDU, 27 were HIV+.

Using those data we can calculate a prevalence ratio. 61 of the IVDU were HIV+ divided by 27 of the non-IVDU who were HIV+ gives us a prevalence ratio of 5.61. This can be interpreted as IV drug users are 5.61 times as likely to be infected with HIV as non-IV drug users.

We can also calculate a prevalence odds ratio similar to how we calculated an odds ratio in a case-control study, that’s a x d divided by b x c. Using those data we get a prevalence odds ratio of 9.40. That can be interpreted as the odds that a HIV+ person uses IV drugs is 9.4 times the odds that a HIV+ person does not use IV drugs.
Slide 77: Cross-Sectional Studies – Disadvantages

Now let’s wrap up cross-sectional studies by talking about its disadvantages. One is a lack of temporal sequence of exposure preceding disease. Since it’s a snapshot in time you don’t necessarily know that exposure preceded disease. Cross-sectional studies tend to include cases with long duration, which may have different characteristics and risk factors than a series of incident cases. These cases with long duration may be less aggressive and have different characteristics than diseases with short duration. There could be potential misclassification of disease status if the disease of interest has exacerbations and remissions such as asthma, multiple sclerosis, or lupus, or if the disease is being treated, such as hypertension.

Slide 78: Cross-Sectional Studies - Advantages

However there are advantages. Cross-sectional studies often have reasonably good generalizability since they are often based on a sample of the general population and not just those seeking care. The generalizability of cross-sectional studies is dependent upon the type of sampling scheme used to identify participants.

Next the data are collected on individuals and not groups like ecological studies. Cross-sectional studies are often conducted in a relatively short period of time, since it is a snapshot in time in which disease and exposure status are simultaneously assessed. And, cross-sectional studies are less costly than both cohort and case-control studies.
In summary looking back over all three types of observational studies they can be thought of as natural experiments. That is, participants self-select their exposures and investigators are not allocating exposures to various groups of people. Cohort studies explicitly incorporate the passage of time. Cohort studies can be prospective or retrospective. All case-control studies are retrospective. Investigators identify diseased and non-diseased and then assess their past exposure status. Uniformity in data collection is key to increased validity. Exposure and disease status must be assessed uniformly and systematically without regard to the other across all study design types. We talked about the various measures of association: relative risk for cohort studies and odds ratio for case-control and cross-sectional studies. These are the key measures of association used to make a causal inference. That concludes this presentation on observational studies. Thanks.