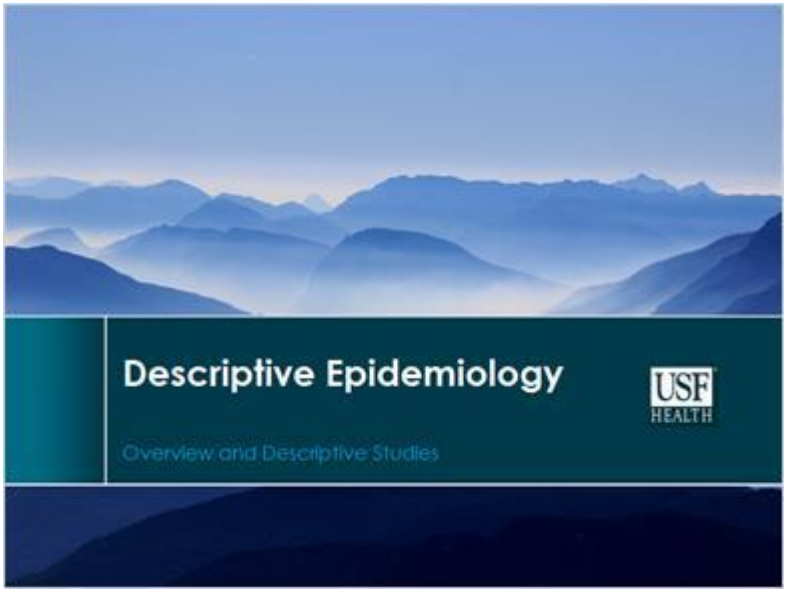
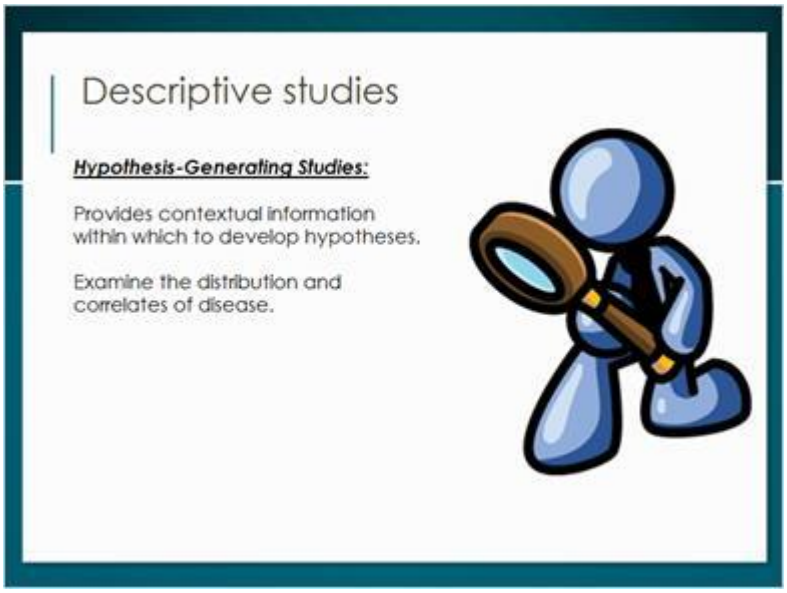
Lecture 1b Descriptive Epidemiology

1.1 Measures of comparison



In this lecture, we are going to focus on descriptive studies. We will start by identifying the different types of epidemiological studies and then focus in on specific descriptive designs, how they fit into epidemiology, and spend some time on the analyses issues with these studies.

1.2 Descriptive studies



Let's begin by reviewing an overview of descriptive studies. Descriptive studies are Hypothesis-Generating (Descriptive) studies:

Meaning, these studies provide contextual information within which to develop hypotheses.

Descriptive studies examine the distribution and correlates of disease.

1.3 Types of descriptive studies

There are a number of different types of descriptive studies.

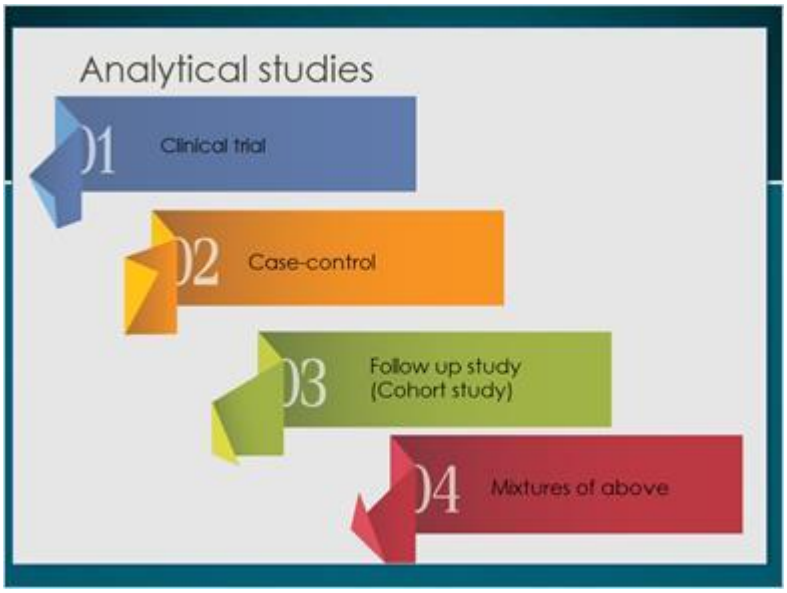
These include surveys, which may involve either primary or secondary data collection. Primary is when the researcher actively collects the data and secondary is when a researcher uses existing data. These surveys tend to represent a cross-section of the population. For example, we might ask a sample of USF college students about texting and driving, where we collect the information directly from students, meaning primary data collection. Another example would be using existing datasets, such as the NHANES dataset to look at heart disease and risk factors.

We can also look at correlation studies, including studies that are ecologic in nature that is those that compare rates of events between different communities. In these instances the unit of observation is not a person; it is a group or population.

Case reports describe an unusual situation in a patient. One often finds a thorough literature review as part of a case report because the author is explaining what is unusual or different about the person.

We also have case series which describe unusual findings in a group of people. Common features among the group may give us a clue about the disease and its cause.

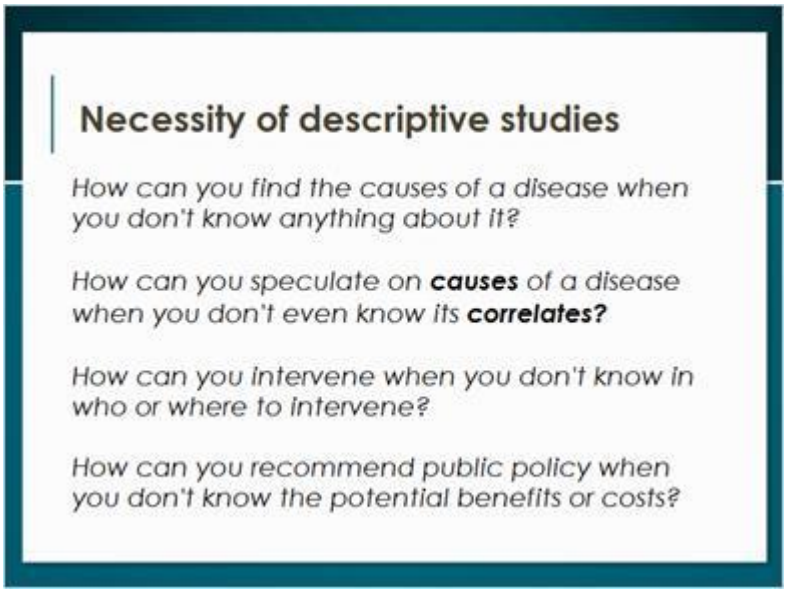
1.4 Analytical studies



In contrast to descriptive studies, we have analytical studies, which test hypotheses about different associations between exposures and diseases. The ultimate goal is to identify causation. These may be familiar to you from earlier epidemiology classes.

These include clinical trials, case-control studies, cohort studies, and mixtures of the above such as a nested case-control study.

1.5 Necessity of descriptive studies



We will discuss analytical studies in a future lecture, but let's start with focusing on descriptive studies

Why do we conduct descriptive studies? Think about these questions:

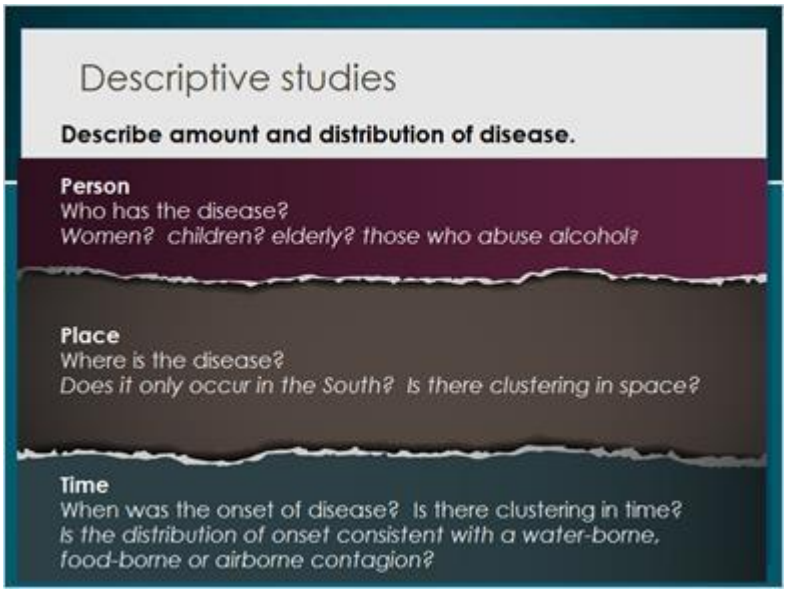
How can you find the causes of a disease when you don't know anything about it?

How can you speculate on causes of a disease when you don't even know its correlates?

How can you intervene when you don't know in who or where to intervene?

How can you recommend public policy when you don't know the potential benefits or costs?

1.6 Descriptive studies



So, how do descriptive studies work?

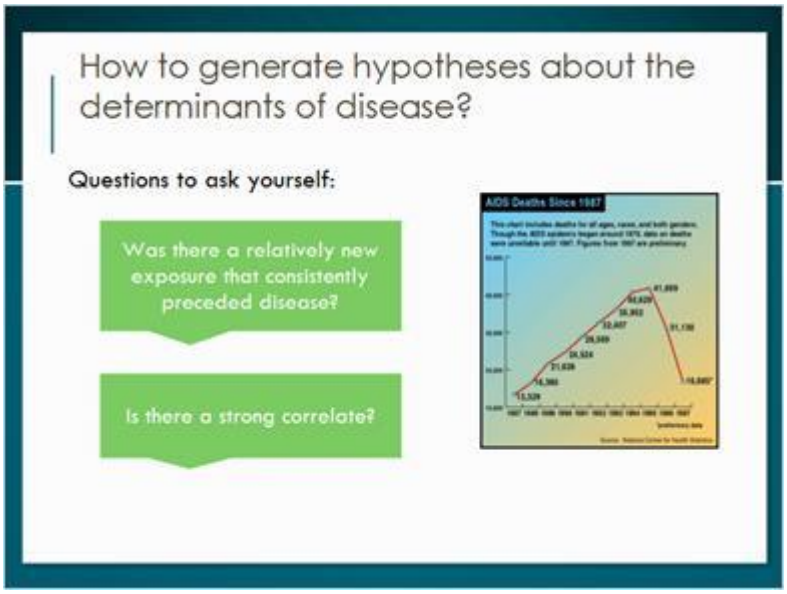
They focus on describing the amount and distribution of disease. We ask questions about three common determinants of disease.

person - Who has the disease? Women? children? elderly? those who abuse alcohol?

place - Where is the disease? Does it only occur in the South? Is there clustering in space?

time - When was the onset of disease? Is there clustering in time? Is the distribution of onset consistent with a water-borne, food-borne or airborne contagion?

1.7 How to generate hypotheses about the determinants of disease?

<table>
<tr>
<td> 
</td>
<td>So, how do we generate hypotheses about determinants of disease? This is a great challenge fro epidemiologists as well as epidemiology students. You will consider this when you are designing your thesis, special project, or dissertation.

It is helpful to start by asking yourself questions

Questions to ask yourself. For example in the situation where there is a new disease or a change in a disease you might ask yourself

• Was there a relatively new exposure that consistently preceded disease?

• Is there a strong correlate?

But we need the descriptive studies first so we can know that there is either a new disease or a change in a disease.

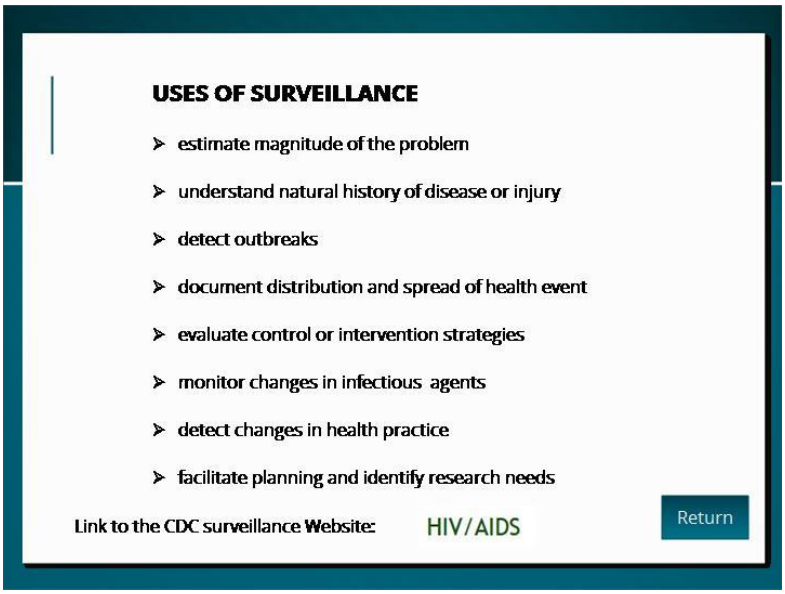
Click on the graph showing AIDS deaths to see a news report on early investigations into this at the time, a new disease</td>
</tr>
<tr>
<td>HIV video</td>
<td></td>
</tr>
</table>

1.8 Surveillance

<table>
<tr>
<td> 
</td>
<td>In addition to generating hypothesis, descriptive studies can be useful for surveillance. While surveillance can be accomplished by any type of study, it is an important function of descriptive epidemiology. The purpose of surveillance is to provide systematic and on-going assessment of health of a community, including timely data collection, analysis, interpretation, dissemination, and subsequent use of data.

Ongoing scrutiny, using methods distinguished by their practicality, uniformity, and frequently their rapidity, rather than by complete accuracy. (e.g.- get blood pressure data from routine office visits; high blood pressure deaths from the national death registry.)

Click on uses of surveillance to learn more about how surveillance is used</td>
</tr>
</table>



Surveillance is used for the following purposes:

- estimate magnitude of the problem
- understand natural history of disease or injury
- detect outbreaks
- document distribution and spread of health event
- evaluate control or intervention strategies
- monitor changes in infectious agents
- detect changes in health practice
- facilitate planning and identify research needs

Going back to one of our previous examples, HIV - in the US we have an HIV surveillance system set up by the CDC.

CDC funds state and territorial health departments to collect surveillance data on persons diagnosed with HIV infection; all personal identifiers are removed from these data before being transmitted to CDC via a secure data network. Data are analyzed by CDC and then displayed by age, race, sex, transmission category, and jurisdiction (where appropriate).

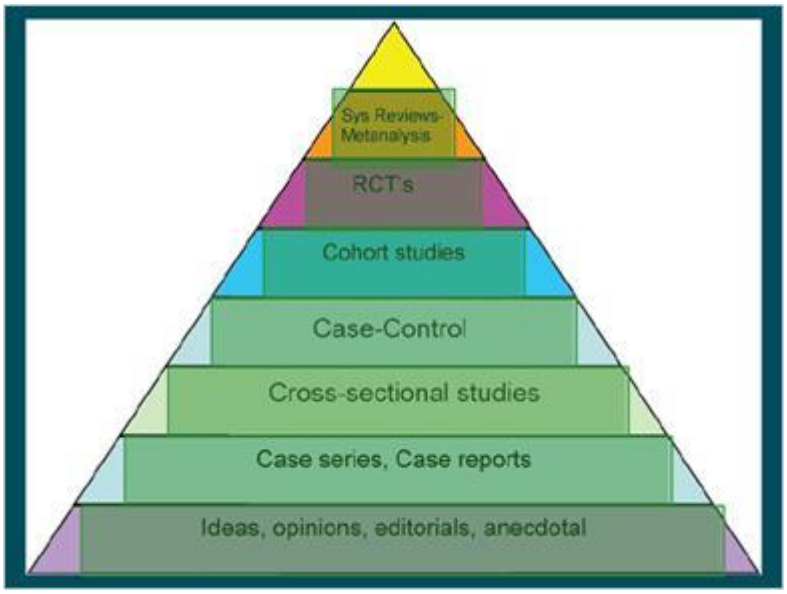
We can then examine the epidemiologic profile, which describes the burden of HIV on the population of an area in terms of sociodemographic, geographic, behavioral, and clinical characteristics of persons with HIV. The profile is a valuable tool that is used at the state and local levels by those who make recommendations for allocating HIV prevention and care resources, planning programs, and evaluating programs and policies.

The National Notifiable Diseases Surveillance System (NNDSS) from the CDC is a nationwide collaboration that enables all levels of public health-local, state, territorial, federal, and international-to share notifiable disease-related health information. Public health uses this information to monitor, control, and prevent the occurrence and spread of state-reportable and nationally notifiable infectious and noninfectious diseases and conditions.

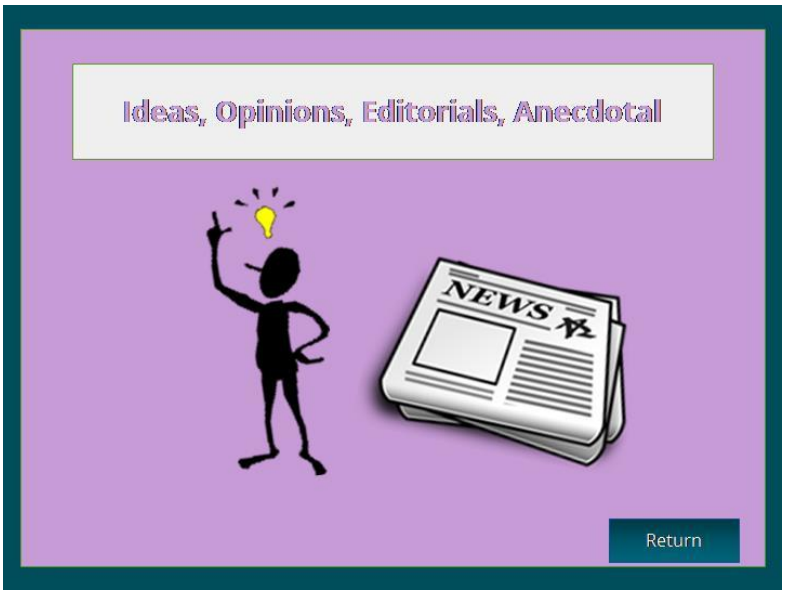
The Food and Drug Administration Med watch is used to report serious problems with human medical products.

The National Center for Health Statistics is requires public data collection and dissemination. Explore the website if you are interested in learning more about different types of datasets.

1.9 Research Studies

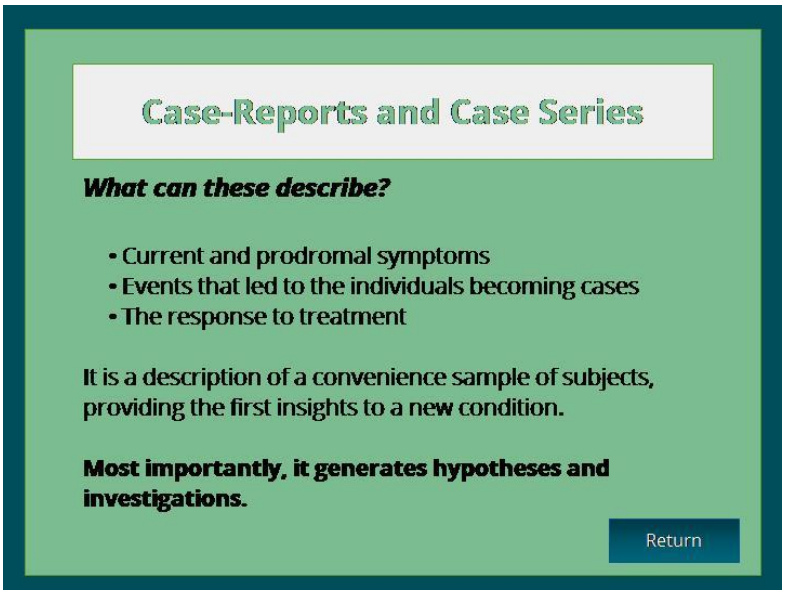
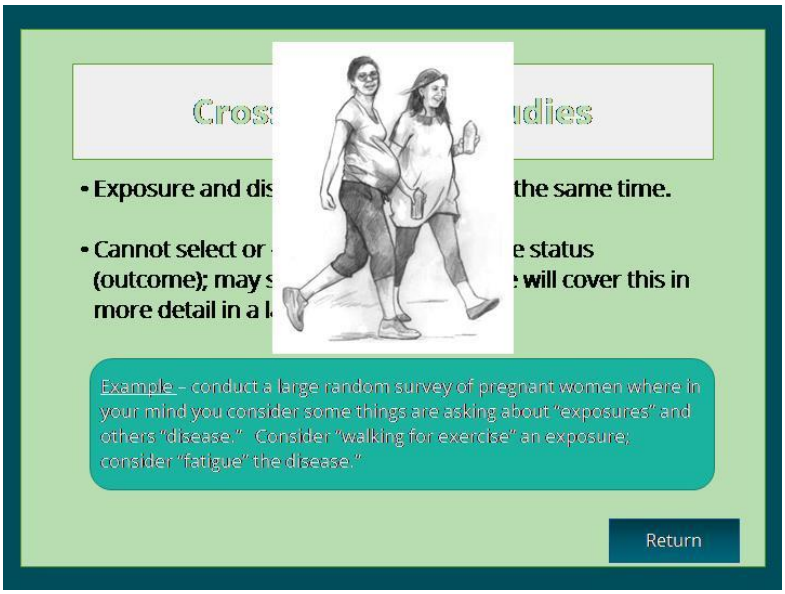


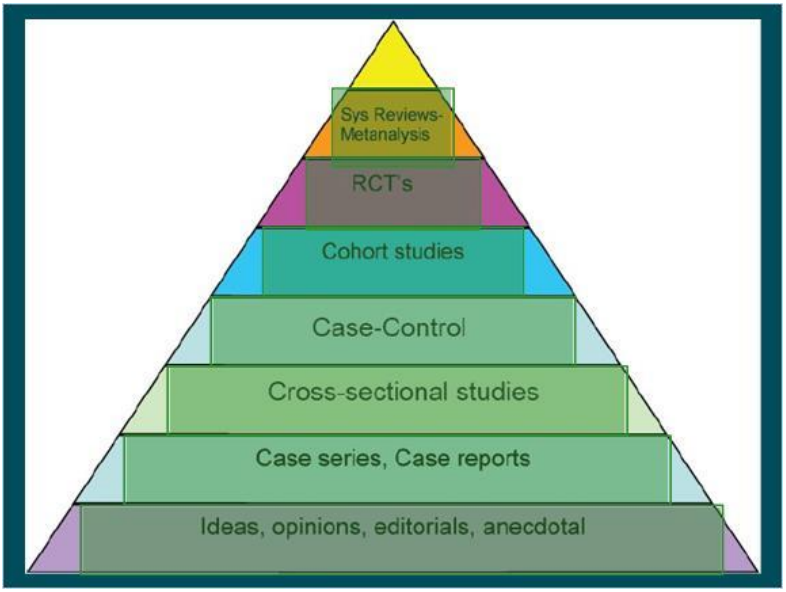
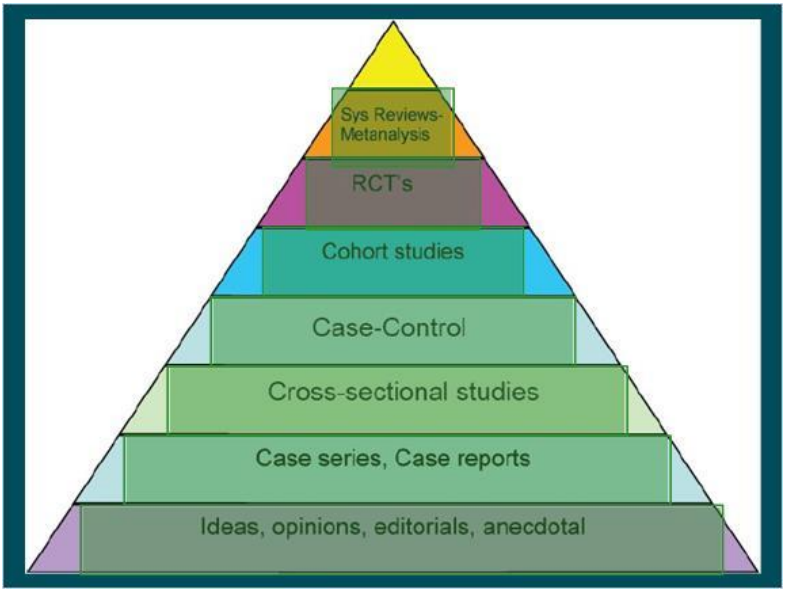
This model demonstrates the relative strength of different research studies in identifying associations between exposures and outcomes. As you go up the pyramid, you increase your ability to determine causation. In this lecture, we are going to focus on the three lower layers of the pyramid, which comprise the descriptive studies.

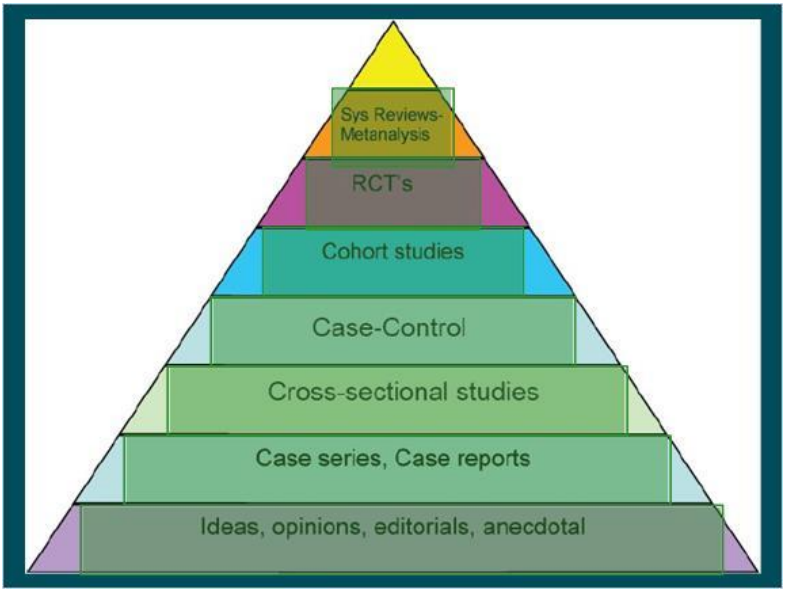
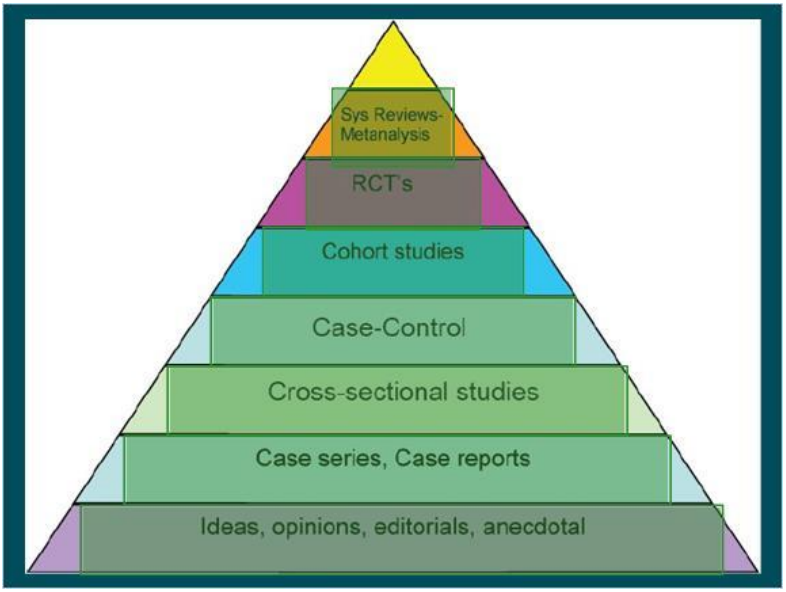


Ideas, opinions, editorials, anecdotal

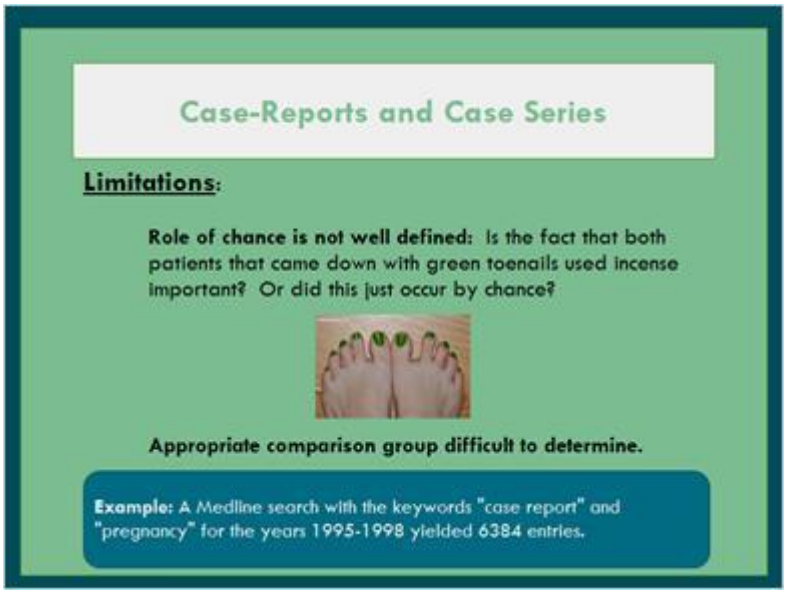
The first layer is identifying possible hypotheses through ideas, opinion, or anecdotal thoughts. This can help us in our hypothesis generating of asking is there a correlation or association? Of course, this alone does not provide scientific evidence for an association, but can start the thinking process.

<table>
<tr>
<td> 
</td>
<td>Case series, Case reports</td>
</tr>
<tr>
<td> 
</td>
<td>Cross-sectional studies</td>
</tr>
</table>

	Case-Control
	Cohort studies

 Sys Reviews-Metanalysis RCT's Cohort studies Case-Control Cross-sectional studies Case series, Case reports Ideas, opinions, editorials, anecdotal	RCTs
 Sys Reviews-Metanalysis RCT's Cohort studies Case-Control Cross-sectional studies Case series, Case reports Ideas, opinions, editorials, anecdotal	Sys

1.10 Case Reports and Case Series



Let's look at more on case reports:

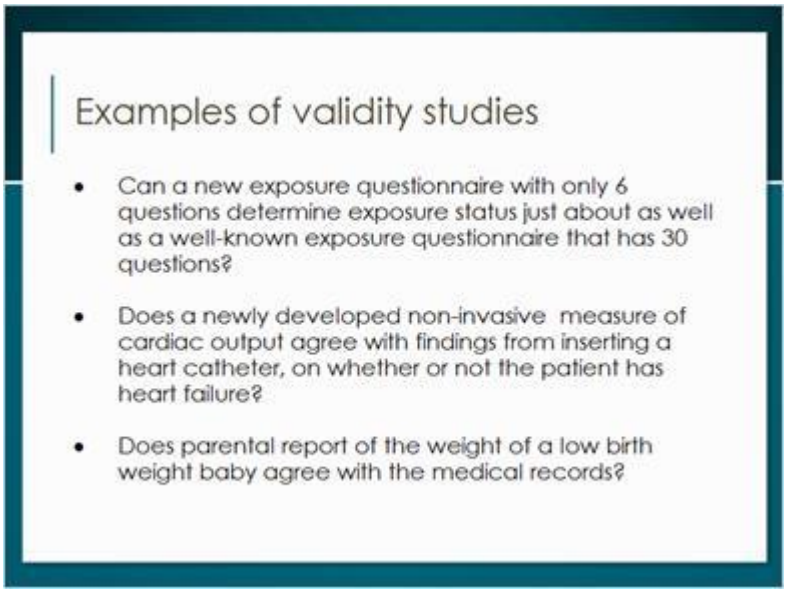
There are some serious limitations. First of all,

The role of chance is not well defined: Let's look at a case in which two patients presented to a doctor with green toenails. Upon the initial interval, they doctor discovered that both patients reported using incense, Is the fact that both patients that came down with green toenails used incense important? Or did this just occur by chance? Because a case series only has cases, there are no comparison groups. Appropriate comparison groups are difficult to determine since all we have are cases.

But this is not an uncommon design. For example,

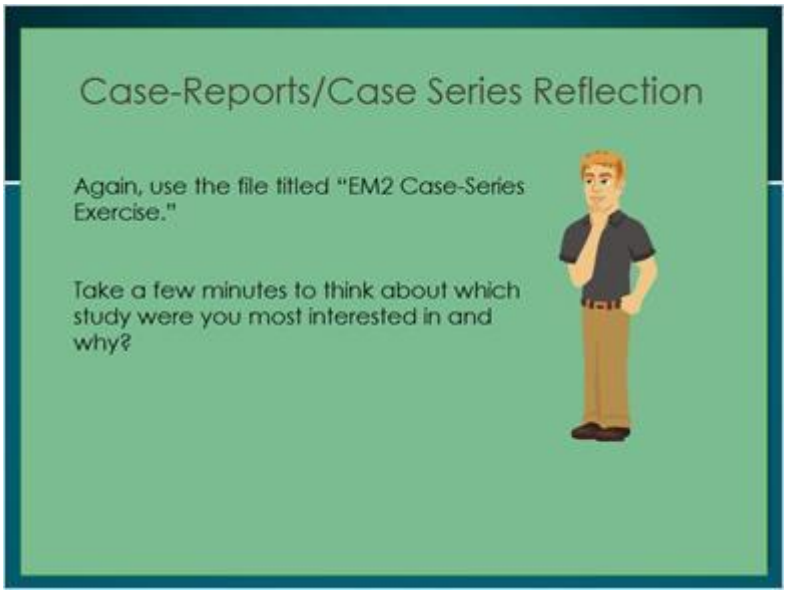
A Medline search with the keywords "case report" and "pregnancy" for the years 1995-1998 yielded 6384 entries.

1.13 Examples of validity studies



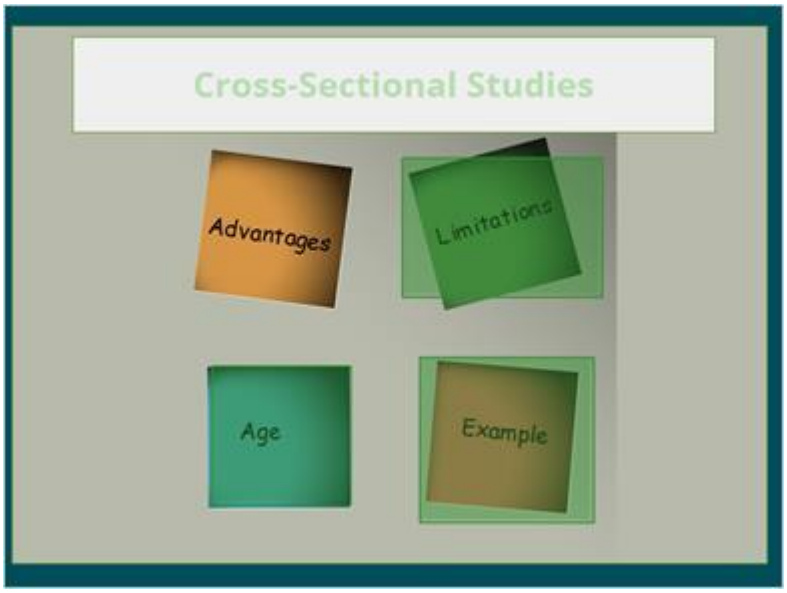
Let's digress for a minute to quickly give examples of research questions that can be answered by validity studies.

1.14 Case-Reports/Case Series Reflection



As a future epidemiologist, you need to think about what you are looking to do in the future. So take a few minutes to think of these abstracts and which studies appealed the most to you. If you get in the habit of doing this in your classes, you will start to understand the areas in epidemiology that you are most interested in.

1.15 Cross-Sectional Studies

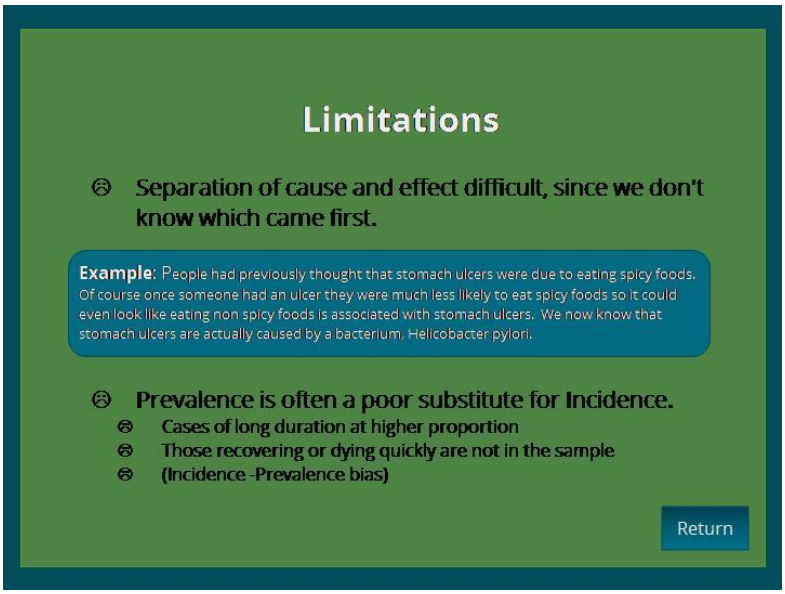


Here is some more information on cross sectional studies. Click on the words to learn more about these studies.

Advantages: The main advantages of cross-sectional studies are that they are usually easier to do and much less expensive than other study designs.

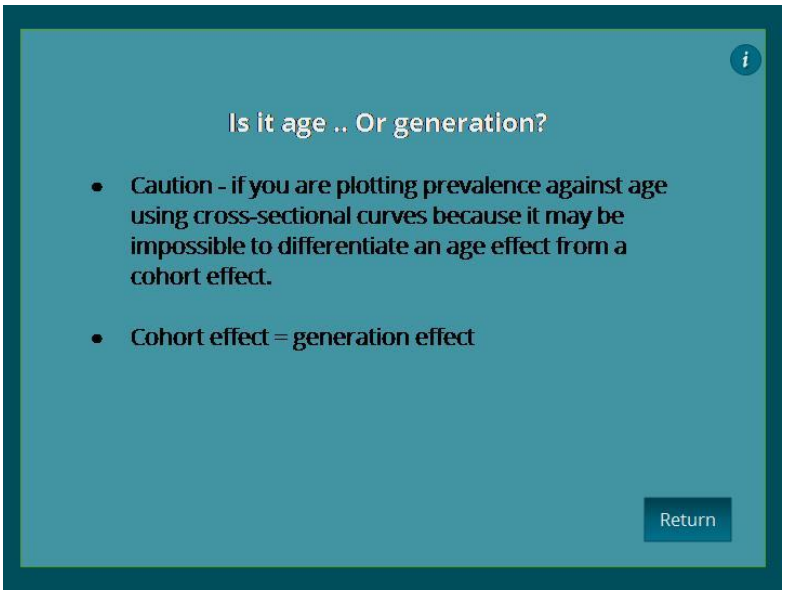
They also allow us to determine the prevalence of a disease or exposure as we are selecting subjects because they represent a population and not based on disease or exposure.

For example, if We work for a pharmaceutical company and we are making a drug to treat disease X, We would want to know (1) how many people have disease X and (2) whether certain subgroups are more likely to have disease X than others, so that we can better target marketing disease X. In this case prevalence of disease is very important as well as the prevalence of the disease among different subgroups.

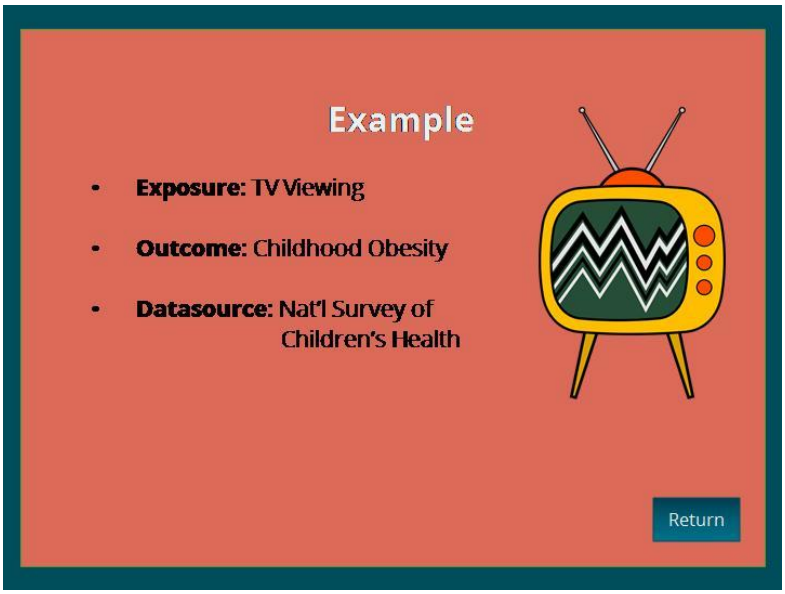


Limitations: Cross-sectional studies have a number of limitations as well. It may be difficult to identify the temporal sequence. For example, people had previously thought that stomach ulcers were due to eating spicy foods. Of course once someone had an ulcer they were much less likely to eat spicy foods so it could even look like eating non spicy foods is associated with stomach ulcers. We now know that stomach ulcers are actually caused by a bacterium, Helicobacter pylori. Thus, remember that association does not always imply causation.

The other limitation is that prevalence is a poor substitute for incidence. Cases with longer duration are more likely to be included in your cross sectional study. Those who recovered quickly or died quickly are much less likely to be in your study. We will discuss this later in the semester.



Age is another issue. There is something called a cohort effect (or generation effect). This occurs when there are changes over a generation by time. Click on the marker to see a definition of cohort effect.



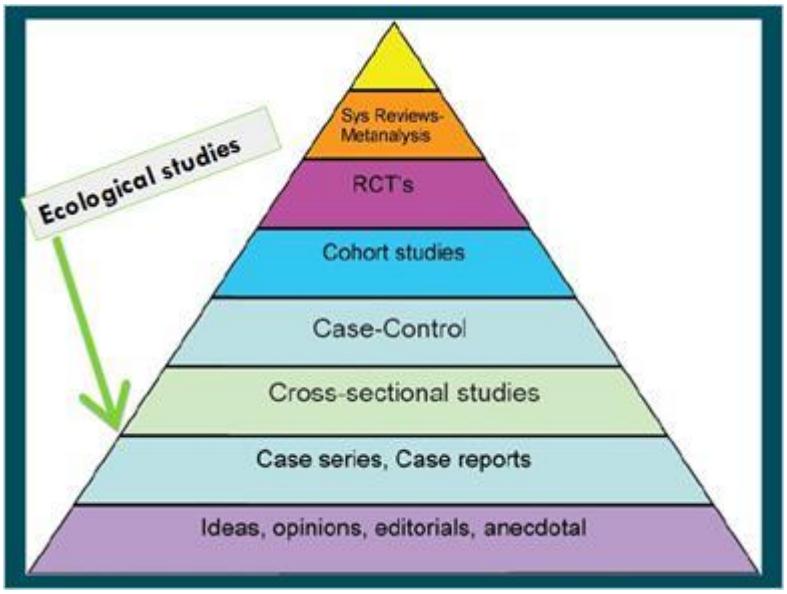
Example:

How would we measure this?

Benefits? easier to do and much less expensive than other study designs; prevalence of childhood obesity

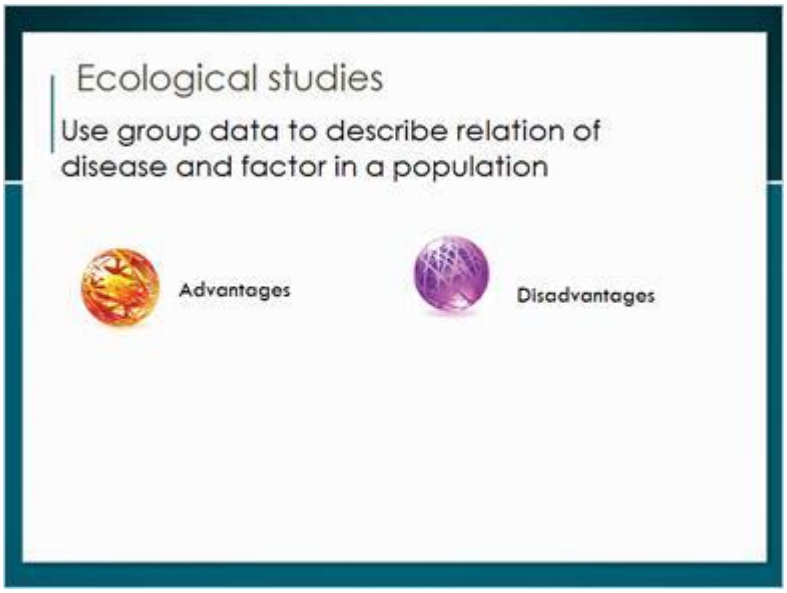
Limitations? difficult to identify the temporal sequence

1.16 Ecological studies



Let's return again to our model that demonstrates the relative strength of different research studies in identifying associations between exposures and outcomes. Remember, as you go up the pyramid, you increase your ability to determine causation. We have focused on the three lower layers of the pyramid, which comprise the descriptive studies. In these study designs, we have been talking about collecting data on individuals to help us with generating hypotheses. But sometimes, we don't always start with individual data, and therefore rely on large groups of data to help us with our hypothesis generating. We are now going to transition to talking about ecologic studies as a descriptive study.

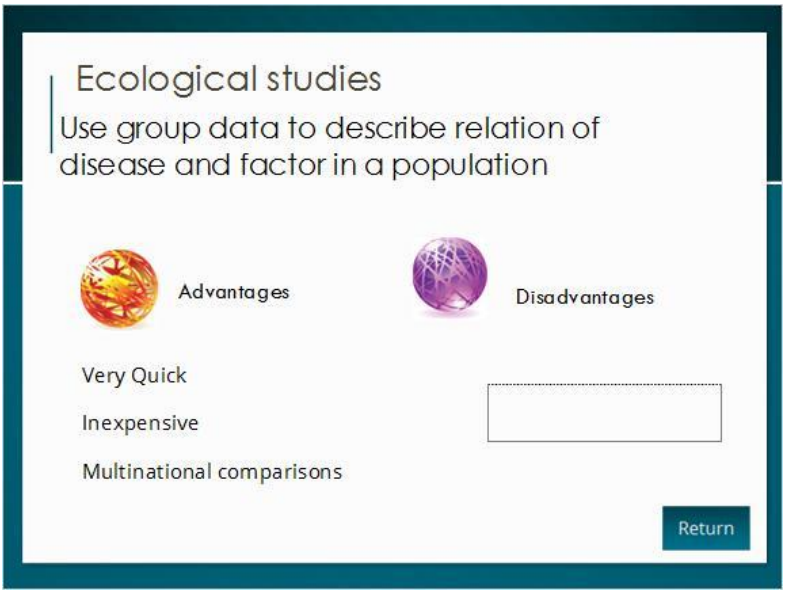
1.17 Ecological studies



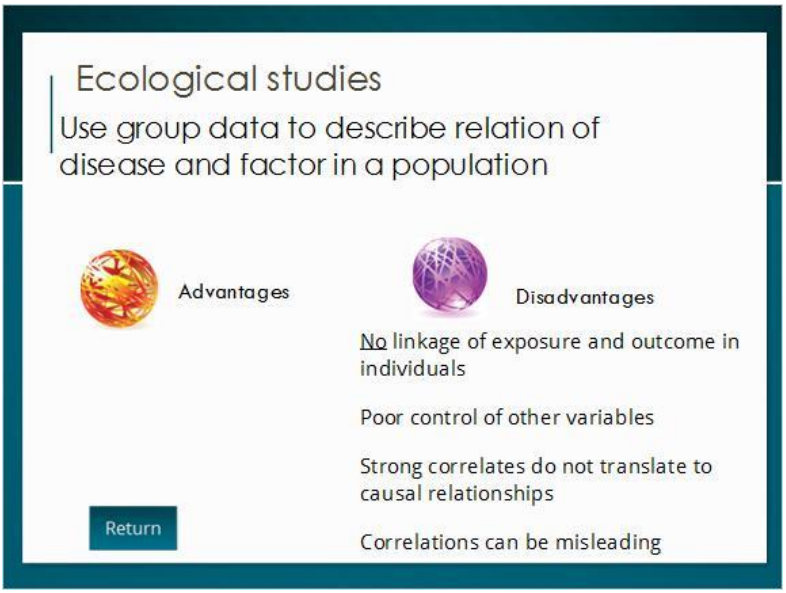
Now, let's shift our attention slightly to ecological studies.

An ecological study to be a type of correlation study, similar to a cross sectional study but the key difference is that it uses group rather than individual data. The correlation coefficient or regression slope measures the degree of association.

Click on the balls to learn more about advantages and disadvantages

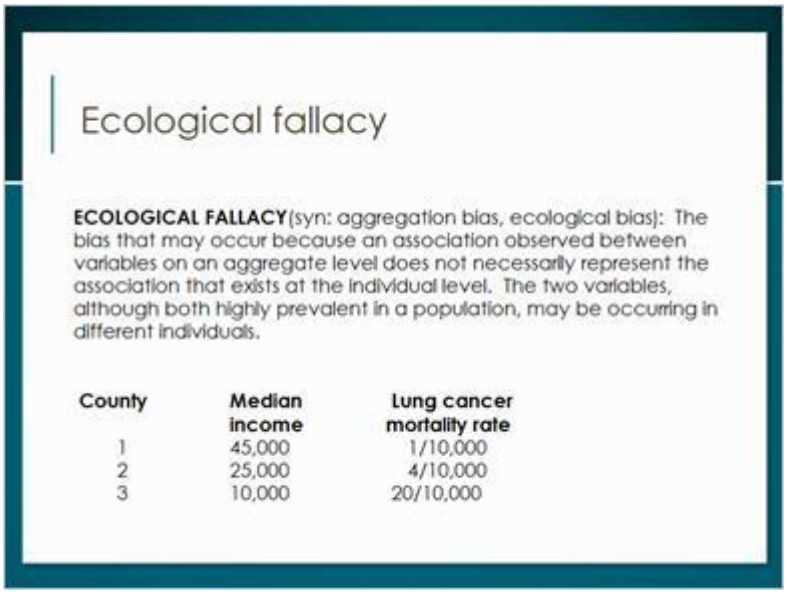


Advantages



Disadvantages

1.19 Ecological fallacy

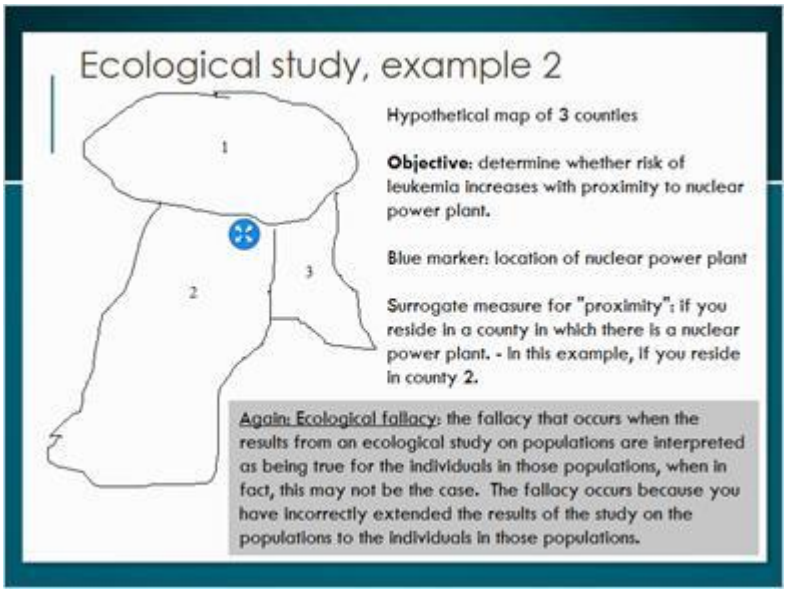


There is a key bias that can be present in ecological studies called the ecological fallacy. It states that while there is an association identified at the group level the individuals who experience the outcome may not actually experience the exposure.

So in the previous example, it may actually be the case that those who die of lung cancer are not the low income people. Suppose that in county 3, there is one extended *very wealthy family*. This family is, say, the Reynolds family, who owns all of the tobacco farmland in the county. Suppose that this family accounts for 18 of the 20 lung cancer cases ....

On the individual level, what is the association between being poor and lung cancer? Well, for county 3, it would be that the wealthy are more likely to die of lung cancer. This nullifies our conclusion based on the ecological data.

1.20 Ecological study, example 2



Let's look at a second example of ecological data. Here we have a hypothetical map of 3 communities. The objective is to determine whether risk of leukemia increases with proximity to nuclear power plant. On this map, the blue marker indicates the location of the nuclear power plant.

In a study, we might use a surrogate measure for proximity to be the county a person lives in and whether there is a power plant present. For example, county 2 would be marked as having close proximity to a power plant. So looking at this map, if we compared the rates of leukemia in the three counties, you would expect to have higher rates in county 2 compared to counties 1 and 3, right? Yet, the conclusion of this study found that there was no increased risk with proximity to the power plant. In fact, they found that there was a lower risk of leukemia in county 2 compared with counties 1 and 3.

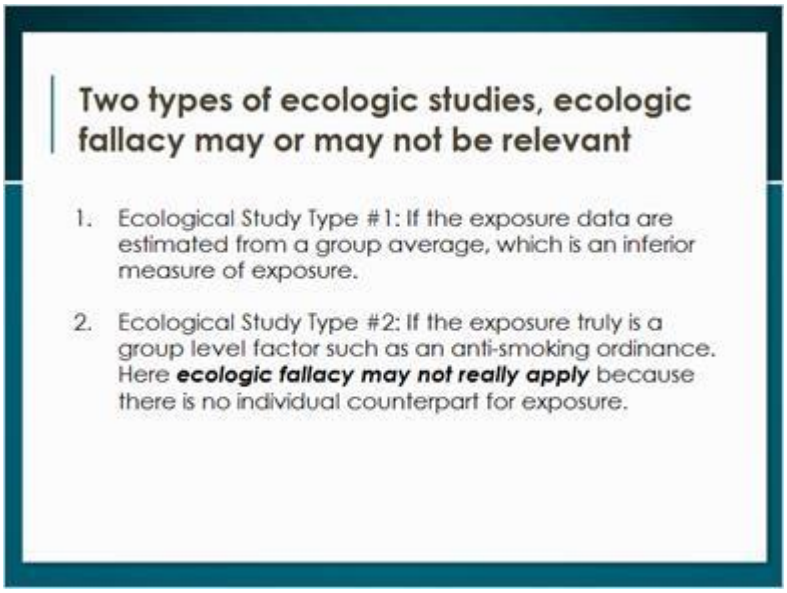
Why is this the case? Is this ecological fallacy, or just misclassification?

If in an ecological study, the bias occurs because of the inability to measure individuals, you can probably call it ecological fallacy.

Again: Ecological fallacy: the fallacy that occurs when the results from an ecological study on populations are interpreted as being true for the individuals in those populations, when in fact, this may not be the case. The

	fallacy occurs because you have incorrectly extended the results of the study on the populations to the individuals in those populations.

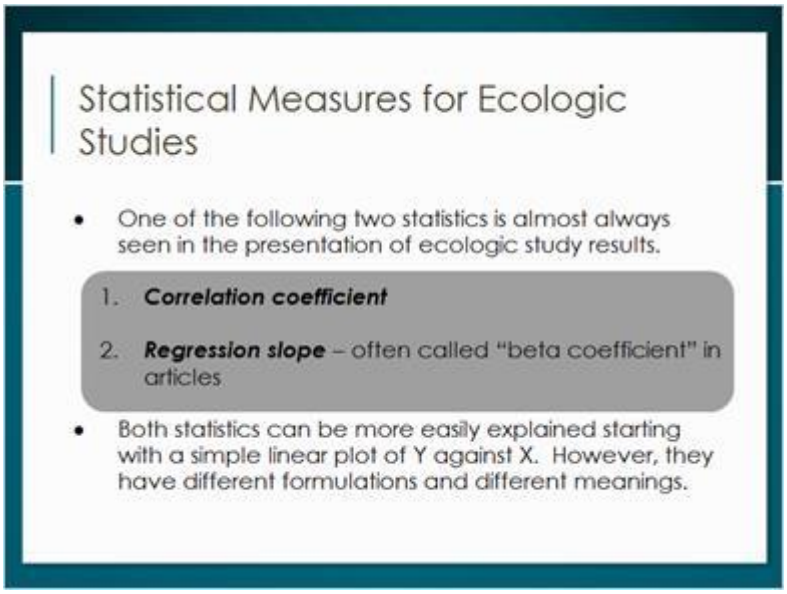
1.21 Two types of ecologic studies, ecologic fallacy may or may not be relevant



There are two instances in which the ecological fallacy may not be relevant

- The first way a study can be ecologic is if the exposure data are estimated from a group average, which is an inferior measure of exposure. Think of our last example - we could have had a more accurate measure of the exposure, distance to the power plant, for each individual.
- The second way a study can be ecologic is if the exposure truly is a group level factor such as an anti-smoking ordinance. Here ecologic fallacy may not really apply because there is no individual counterpart for exposure.

1.23 Statistical Measures for Ecologic Studies

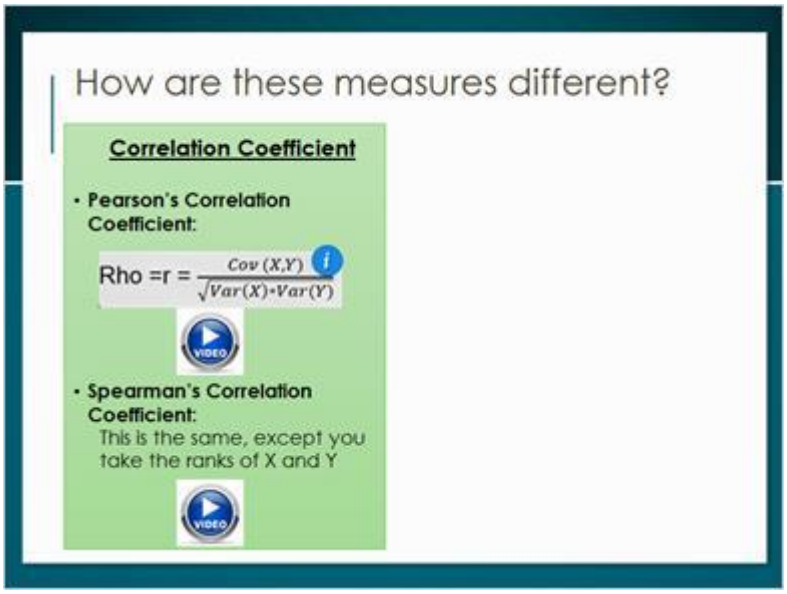


Now that we have a basic understanding of what an ecological study looks like, let's review how we see the data presented through statistics.

- One of the following two statistics is almost always seen in the presentation of ecologic study results.
- Correlation coefficient.
- Regression slope - often called "beta coefficient" in articles
- Both statistics can be more easily explained starting with a simple linear plot of Y against X. However, they have different formulations and different meanings.

Let's take a closer look at this.

1.24 How are these measures different?

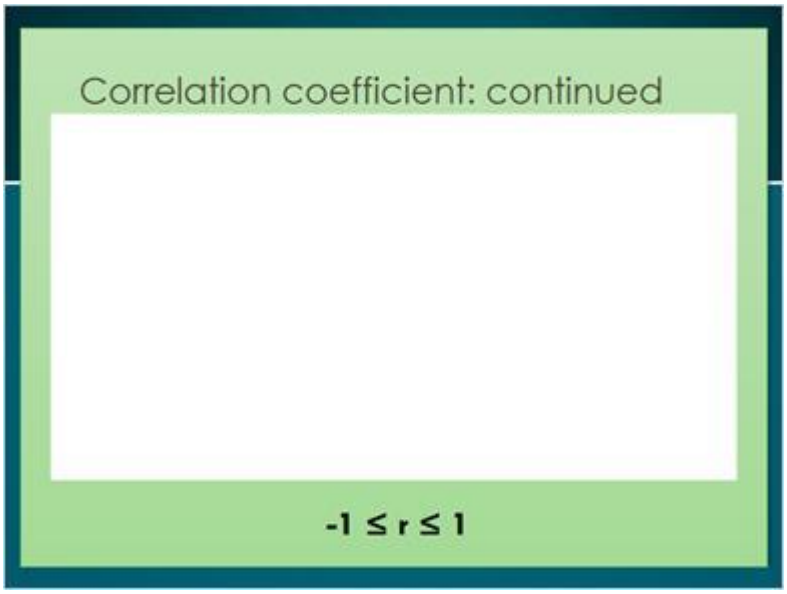
<table>
<tr>
<td> 
</td>
<td>We just learned, ecological studies can use two different types of measures.

The first is a correlation coefficient. And in this class, we will refer to two kinds. The take home message of these two correlation coefficients (pearson and spearman's) is that we are interested in the covariance of X and Y. First, we need to consider the concept of covariance. Covariance is the average amount by which X and Y covary, or change together. We denote the covariance between x and y using the notation cov(X, Y). The denominator is then the square root of the variance of x and y, respectively. This denominator results in a positive scaling factor os that r falls between -1 and 1. Click on the blue video buttons to obtain more information about each of these measures. You can also use the information markers to review refreshers on covariance and variance.

The second type of measure we can consider is a regression slope. This is the unadjusted slope of a regression line, also known as a beta coefficient. This measure has the covariance of x and y in the numerator and the variance of x in the denominator. How is this different than the last equation? Well, take a look at the denominator. The slope does not have that "positive scaling factor", which means the beta coefficient is not restrained to -1 to 1, instead it is limitless.

Finally, we need to be careful with how we talk about these two different measures. Both the correlation coefficient and the slope are loosely referred to as a "coefficient"</td>
</tr>
<tr>
<td>Peason's r Correlation video</td>
<td></td>
</tr>
<tr>
<td>Spearman Correlation video</td>
<td></td>
</tr>
</table>

1.25 Correlation coefficient: continued



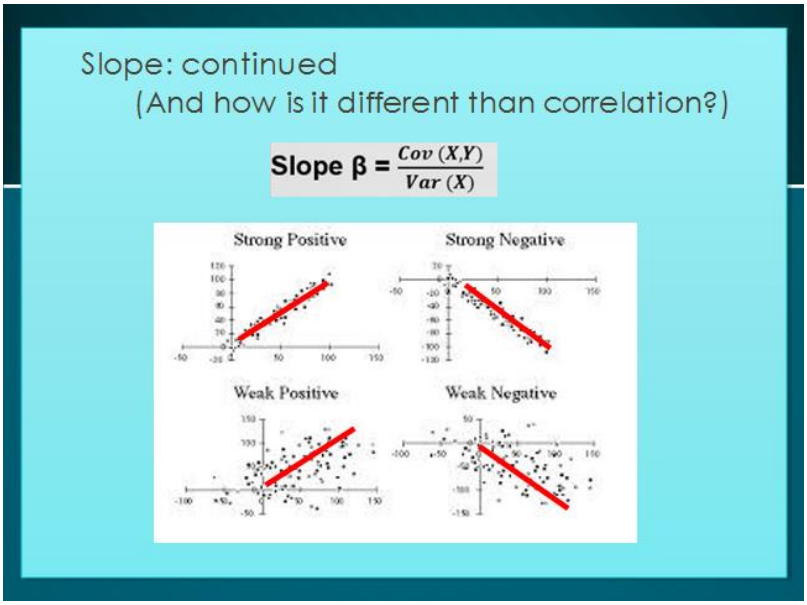
Correlation coefficient: continued

- Recall, that the correlation coefficient can only range from -1 to 1
- The correlation coefficient of data falling on any straight line with a positive slope is equal to 1

The correlation coefficient of data falling on any straight line with a negative slope is equal to -1

The correlation coefficient of data all falling on a horizontal line or vertical line is undefined because the variance of y=0 or the variance of x=0, and therefore the denominator =0.

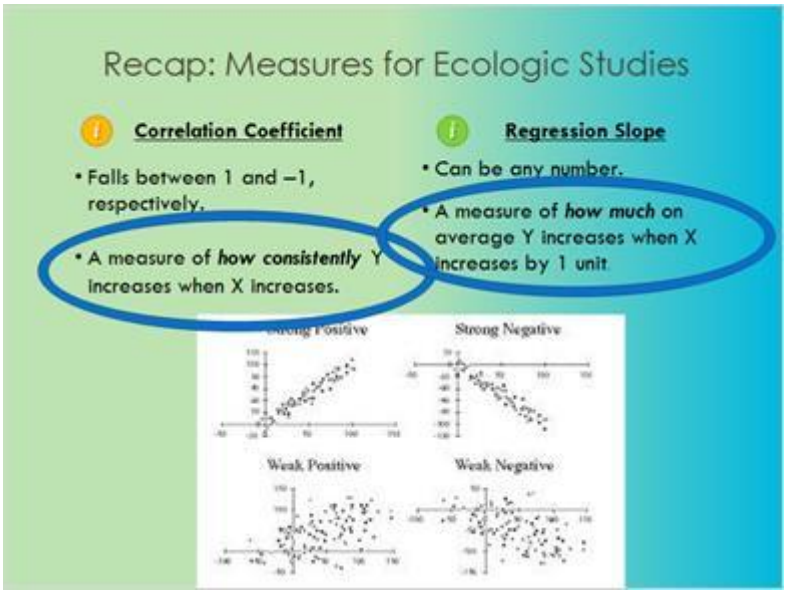
1.26 Slope: continued



Now to describe the slope. The slope is a measure of the change in Y with a one unit change in X. That's why we have the variance of x as our denominator.

While the correlation coefficient gives us an idea of strength of the points, such as weak or positive as shown on this slide, the slope is does not provide that type of information. In fact, these slopes of the lines are the same, regardless if they are weak or positive. Let me elaborate: The two figures on the left have the same positive slope -but the one on top shows a strong correlation while the one on the bottom shows a weak correlation. The two figures on the right the same negative slope -but the one on top shows a strong negative correlation while the one on the bottom shows a weak negative correlation

1.27 Recap: Measures for Ecologic Studies



So, let's recap on these two measures for ecologic studies.

- A correlation coefficient falls between 1 and -1 respectively.
- A slope can be any number.

If you think about it,

- The correlation coefficient is a measure of how consistently Y increases when X increases.
- The slope is sort of a measure of how much on average Y increases when X increases by 1 unit.
- Intuitively the two complement each other and it would be help to have both statistics reported in a study.
- Funny thing: I've never seen both reported in a study; its either one or the other. But now that I think about it, I would report both.

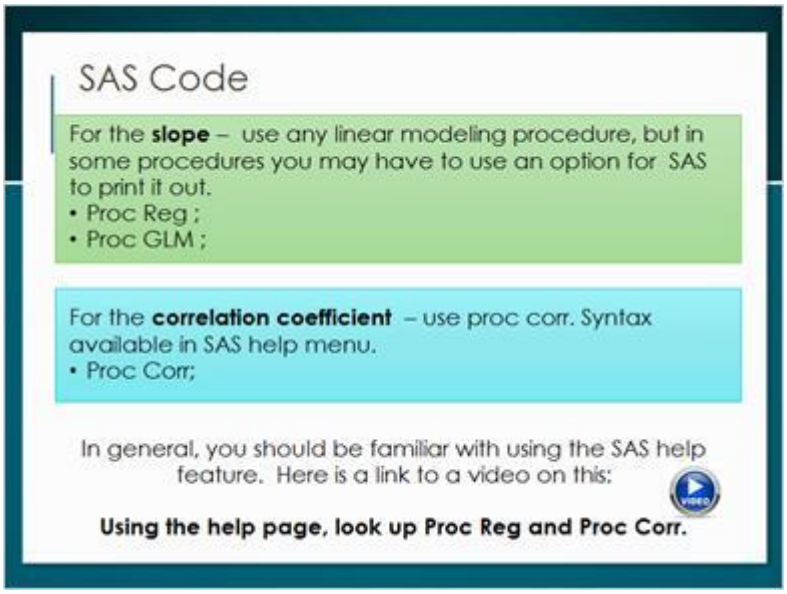
Click on the markers to read these explanations again.

Now to provide an tangible example of comparing these two measures. Let's say we plot an outcome Y against exposure X. Both the slope and the correlation coefficient are measures of association between X and Y.

The correlation coefficient is a measure of how close the points are to a straight line. A reasonable eyeball guess is that the top left and right hand graphs have correlations of 0.95 and -0.95 respectively, while the bottom two has correlation coefficients of 0.3 and -0.3 respectively.

The slope is a measure the change in Y with a one unit change in X. The slope is a really a more direct measure of strength of association, but without taking into account variability or significance. If our eyeball estimation of a 45 degree line is correct, then the left-hand graphs both have slopes of 1.0 while the right hand graphs have slopes of -1 because on average Y increases 1 unit when X increases 1 unit.

1.28 SAS Code



Now that we have the basic ideas on these two measures, how can we calculate them?

To obtain the slope in SAS, you can use any slope, you can use any linear modeling procedure, but in some procedures you may have to use an option for SAS to print it out.

Some of you are very comfortable with SAS and some are not.

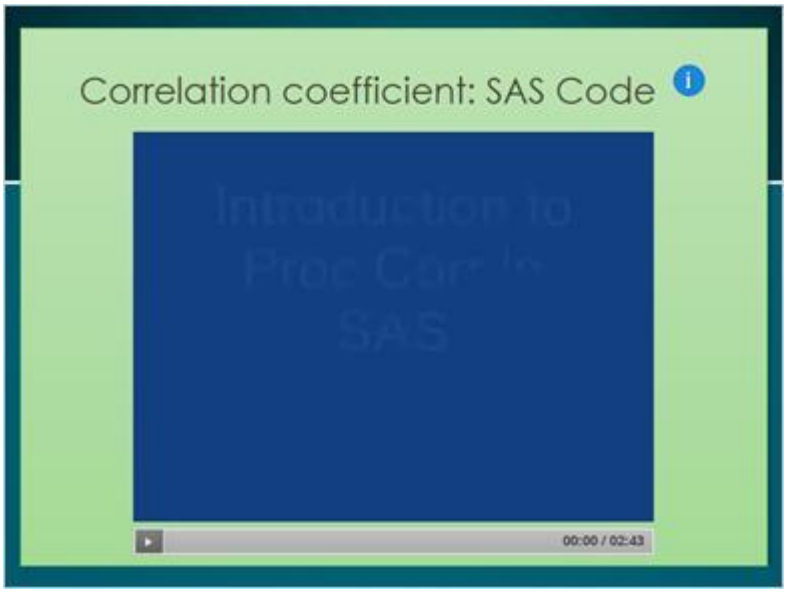
Proc Reg ;

Proc GLM ;

One great tool in SAS is the help feature. I have attached a link to a Youtube video showing how to use the help feature. If you are not extremely familiar with this feature, you need to watch the video. Once you have done so, open SAS and using the help page, look up Proc Reg and Proc Corr. Be sure you can see the code listed for these procedures, and that you would be able to copy that code and apply it to your own file

SAS help video

1.29 Correlation coefficient: SAS Code



- Here is a brief example for using proc corr in sas.

1.31 Slope: SAS code

Here is a brief video on prog reg and sas

1.33 Credits