PHW250B Week 9 Reader

Topic 1: Information Bias, Selection Bias, and Generalizability

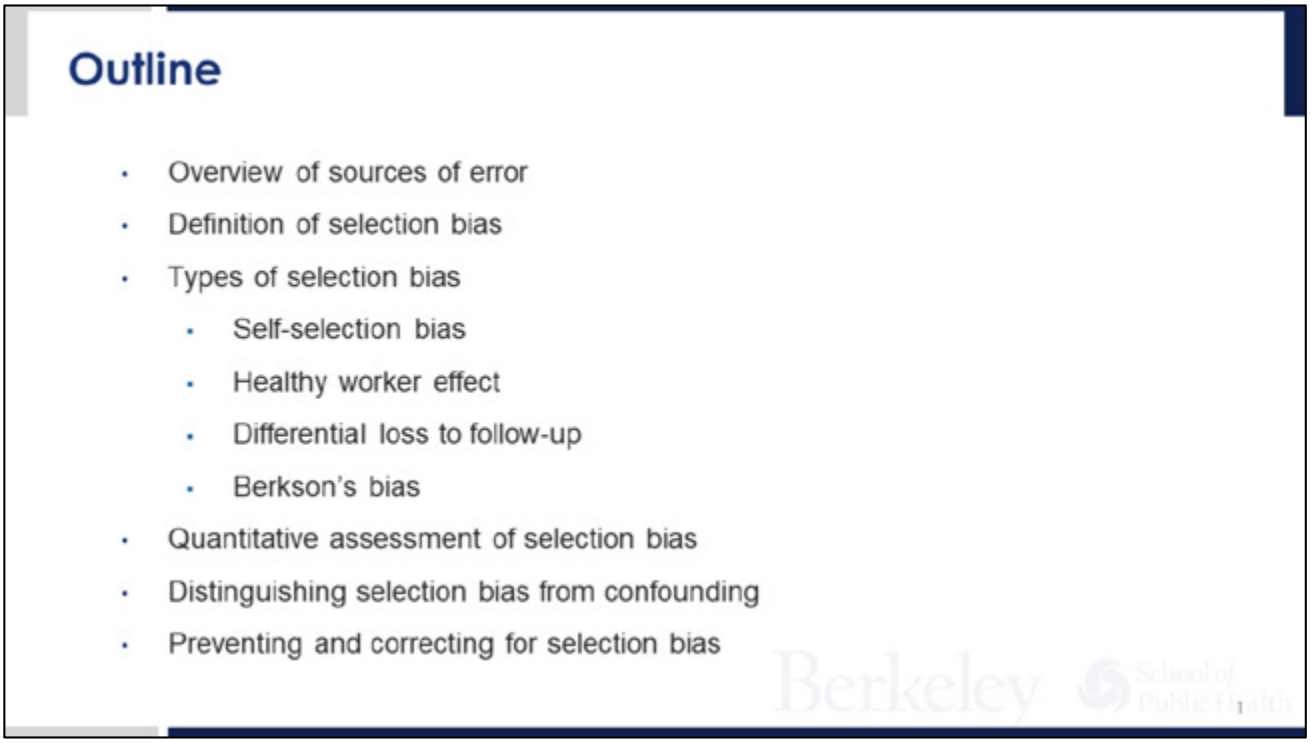
Topic 2: Diagnosing Bias Using DAGs

Podcast

Journal Club

Lecture: Selection bias

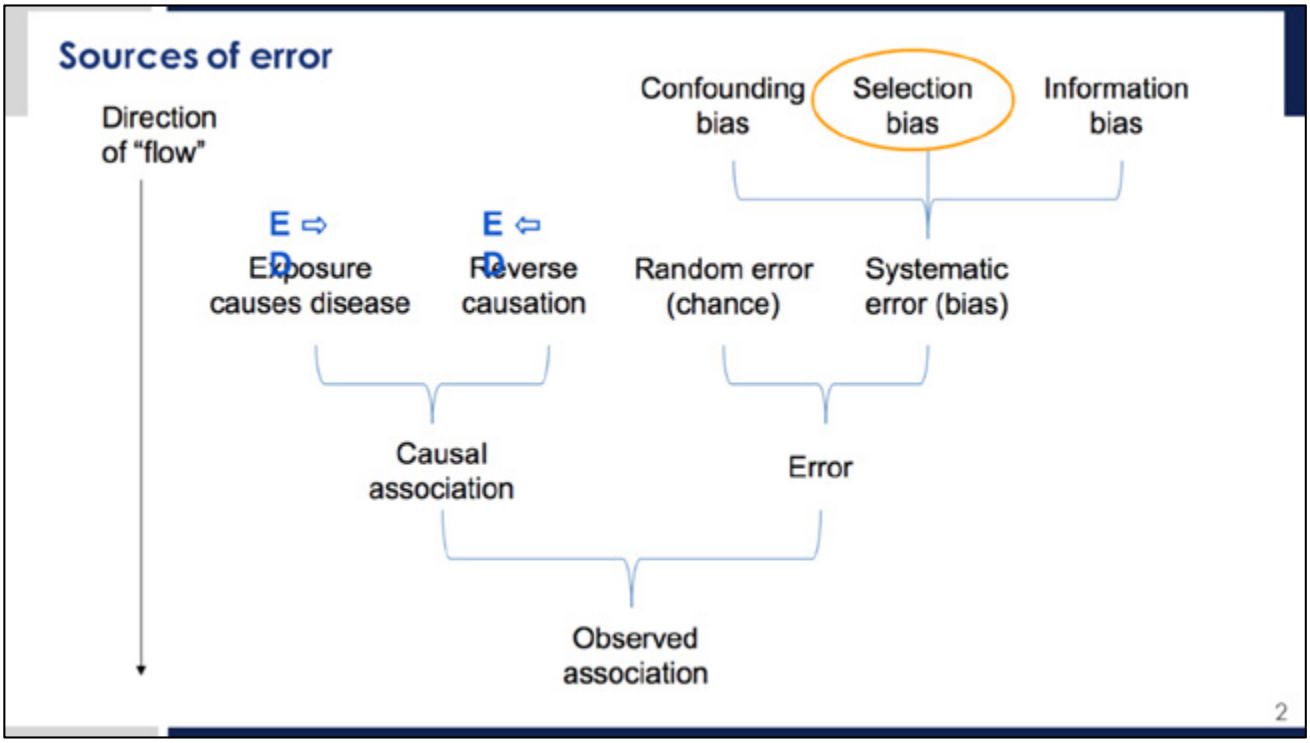
JACK COLFORD: This module is going to discuss the issue of selection bias. I like to think of epidemiology analysis as plagued by things that can go wrong. In the next set of videos, we're going to be talking about the various big picture items that can go wrong in an epidemiology study and what we can do to prevent them.



So to structure our discussions about this topic, we'll talk about the sources of error and the defined selection bias. Because selection bias is just one of several different types of biases.

Then we'll talk about specific examples of selection bias that can occur when self-selection is a problem in a study or when we're following healthy workers or when we differentially lose participants during the study. That is, from different groups we might lose different types of people. And that can create bias in our interpretation when we compare the experience of one group to the other.

We'll talk about a very specific type of bias named after someone, Berkson, that has to do with using a hospital as a source of controls for a study. Then we'll look at some of the ways to quantitatively assess whether selection bias is present, learn how to distinguish selection bias from confounding. Confounding is itself a different type of bias. And then we'll look for ways to prevent and correct for selection bias.



Let's just take a big picture view of this idea that so many things can go wrong in an epidemiology study. You know, when we think about the direction of flow of our thinking and observation in a study, we flow from the top of this figure down to the bottom. Lots of different relationships might be true in what we're studying.

Hopefully we're going to find a relationship where we can properly measure how exposure causes disease. That's what we're about as epidemiologists. But we might also be observing a relationship where we see associations because there's reverse causation, where the disease causes the exposure.

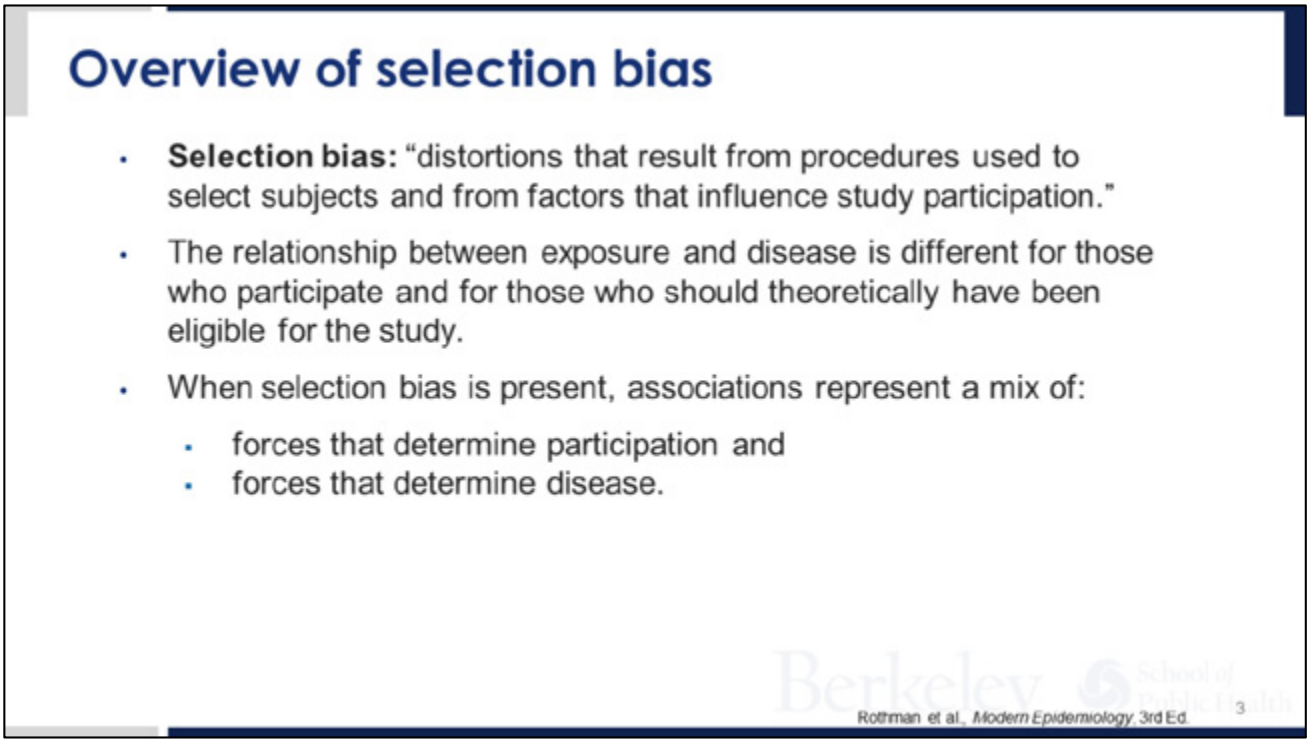
An extreme example of that, for instance, might be, say, someone develops lung cancer and then becomes depressed and starts smoking because they're depressed, having never smoked before. Well, that would be an example where the disease has caused the exposure. So that would be reverse causation.

There are error prone sources that can distort what we see in a study. And these are generally broken down into either random error or chance, which we try to reduce or eliminate through statistical means, or systematic error or bias, which we try to reduce by thinking ahead about our study and then analyzing our study properly.

In the category of systematic error or bias, we tend to group these into three different types of biases-- we call confounding, selection bias, and information bias. We'll go through each of these three, but today's little talk is focused on selection bias.

If you group these sources of error all together, all of them are weighing on our interpretation of the causal relationships between exposure and disease. They interact to produce what we observe as our association.

If we've measured cigarettes and smoking, we've calculated some risk to the smokers compared to the risk to the nonsmokers and then related those risks to each other through a ratio or a difference and called that our association. We're trying to understand how error might have affected our observed association and not lead us to properly conclude whether exposure was causing disease and by how much.



Selection bias can be defined as a distortion that arises from procedures that are used to select subjects and from factors that influence study participation. So either of those two things, the selection of subjects or factors that are influencing the choice of subjects to participate can lead to selection bias. And we'll go through specific examples of each type.

Now the relationship between exposure and disease is different for those who participate and for those who theoretically should have been eligible for the study. So in other words, generally, or often, the people in our study may be different from the whole universe of people we wish we had a sample of in our study.

If we study smokers, they may be a reasonable representation of all smokers, but they may be quite different than the pool of all smokers because by choosing to participate in a study or the way they were recruited, they may somehow be different and not as representative as we hope.

So when selection bias is present, the associations that we see represent a mix of forces that are determining their participation of the subjects in our study, as well as the forces that determine disease. We're really trying to study as epidemiologists, the forces that determine disease. But we have to take care of or adjust or understand these forces that determine participation and that may have made our subjects different than the full universe of our subjects we hoped we had studied.

Self-selection bias

- This bias occurs when people are allowed to self-select into a study and their desire to participate in the study is associated with the exposure and outcome.
- **Example**: In a study of leukemia among troops who worked at the Smoky Atomic Test in Nevada, 18% of troops with known leukemia status asked investigators about participation in the study (instead of being contact by investigators).
- Among troops with known leukemia status,
  - 22% of self-selecting troops had leukemia
  - 5% of non self-selecting troops had leukemia
- This suggests that self-selection bias was present.

Rothman et al., *Modern Epidemiology*, 3rd Ed. 4

There are specific types of selection bias. One of them is called self-selection bias. This occurs when people are allowed to choose for themself, or self-select into a study. And their desire to participate in the study is associated with the exposure and outcome. Here's an example.

When studying leukemia among military troops who worked at the Smoky Atomic Test Site in Nevada, 18% of the troops with known leukemia status asked investigators about participation in the study instead of being contacted by the investigators. Because they knew they had leukemia, they wanted to get into the study.

Now among the troops with known leukemia status, 22% of the self-selecting troops had leukemia and 5% of the non-self-selecting troops had leukemia. So you can already see how different these two groups are based on this self-selection. This differential representation here suggests that self-selection bias was present.

Healthy worker effect

- The healthy worker effect occurs when a study focused on the health of workers compares health outcomes among workers to the general population and the health of workers is better than that of the general population.
- The general population includes people who are not able to work because of illness disability, retirement, etc.
- Traditionally this effect is considered to be a type of self-selection bias, but in the modern view it is classified as confounding (more on this later).

Berkeley School of Public Health

Rothman et al., *Modern Epidemiology*, 3rd Ed. 5

Another type of selection bias we worry about is called the healthy worker effect. This occurs when a study is focused on the health of workers and is trying to compare health outcomes among workers in our study to the general population, and that the health of the workers is better than that in the general population.

And why might that be? Well, because the general population includes people who weren't able to work because of illness, disability, retirement, et cetera. Whereas our healthy workers in our factory or who are in our study are likely on average to be different from the general population.

So traditionally, this effect is considered to be a type of self selection bias. But there are other views, more recent views, that would treat this as confounding, and we'll talk more about this later. But just to know that you'll sometimes see healthy worker effect referred to as a type of selection bias.

Differential loss to follow-up

- Occurs when exposed vs. unexposed follow-up rates differ or disease vs. non diseased follow-up rates differ.
- Most common in cohort studies.
- Can think of as "backwards" selection bias since it occurs during and at the end of the study rather than during enrollment.

Rothman et al., *Modern Epidemiology*, 3rd Ed. 6

Another type of selection bias that can occur and alter the observed associations we see in a study away from the truth is that differential loss to follow-up can occur. Imagine that a cohort is moving along, and we have an exposed population and unexposed population. And the follow up rates might differ, or the disease rates might differ between the exposed and the non-exposed.

So this differential between the exposed and non-unexposed in these follow up rates can occur and lead to a differential estimate of what's happening in the two groups that's different from the truth. This is most common in cohort studies. And you can think of this as backward selection bias since it occurs during and at the end of the study rather than during enrollment.

Berkson's bias (or "Berksonian bias")

- First described by Berkson in 1946.
- Occurs when both exposure and disease are associated with the risk of hospitalization. If so, a false association is induced between exposure and disease.
- Bias occurs because people are more likely to be hospitalized for two conditions than one.
- **Example**: a hospital-based study of the effect of hypertension on skin cancer.
  - People with hypertension are more likely to be hospitalized (regardless of skin cancer status)
  - People with skin cancer are more likely to be hospitalized (regardless of hypertension status)
  - The proportion of people in the study with neither exposure nor disease will be smaller than in the general population

Rothman et al., *Modern Epidemiology*, 3rd Ed.

And finally, another type of selection bias to be aware of is called Berkson's bias, or Berksonian bias, first described in 1946. And this occurs when both exposure and disease are associated with the risk of hospitalization.

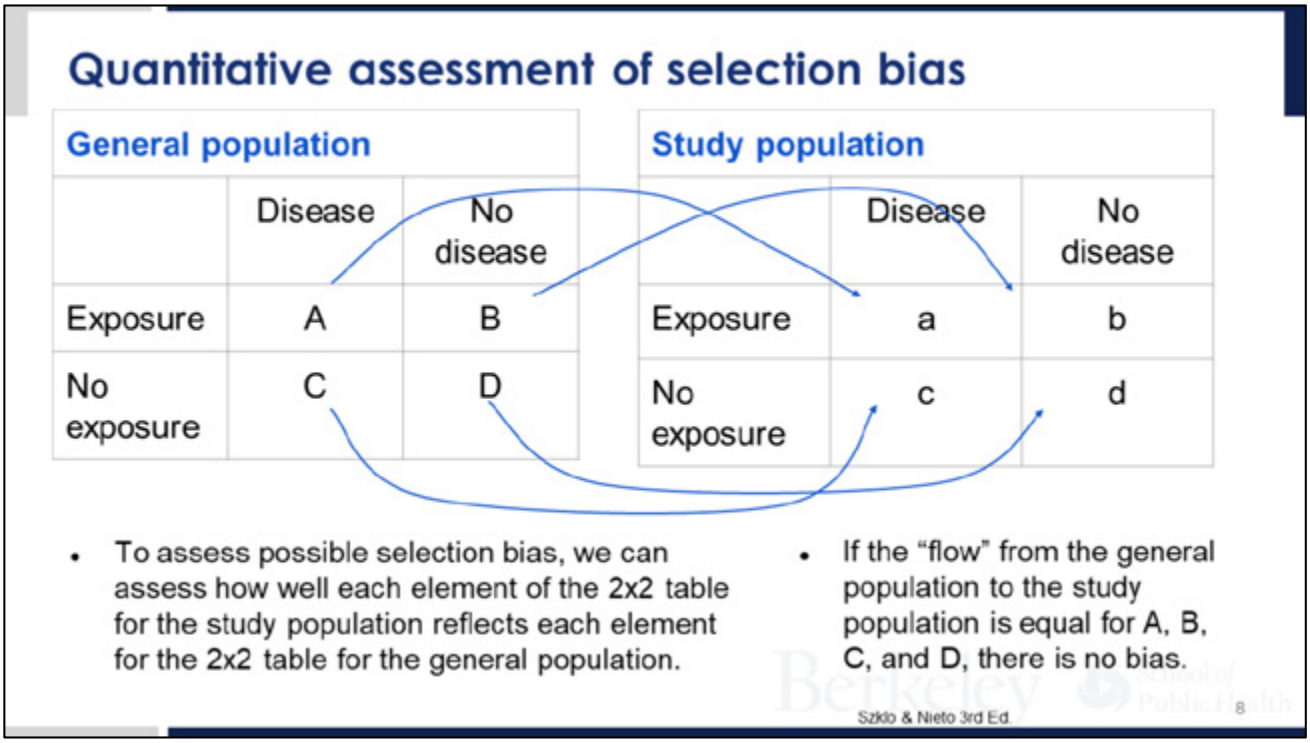
So for example, in this case, a false association is induced between exposure and disease because of this use of the hospital as a way to capture patients. This bias occurs because people are more likely to be hospitalized for two conditions than one.

So here's an example. This is a hospital-based study of the effect of hypertension on skin cancer that's detailed much more extensively in your reading from the Rothman third edition of Modern Epidemiology.

In this study, people with hypertension are more likely to be hospitalized regardless of their skin cancer status. Again, this is a study looking at the effect of hypertension on skin cancer. Hypertension is the exposure. Skin cancer is the outcome.

People with skin cancer are more likely to be hospitalized regardless of hypertension status. So you see how both the exposure and the outcome are changing the probability of being a patient in the hospital.

So the proportion of people in the study with neither exposure nor disease will be smaller in the general population because these two affects this over-representation of exposure and outcome in the hospital population, making it different from the general population. Where we really wish we could be doing the study is Berkson's bias, is this selection bias.



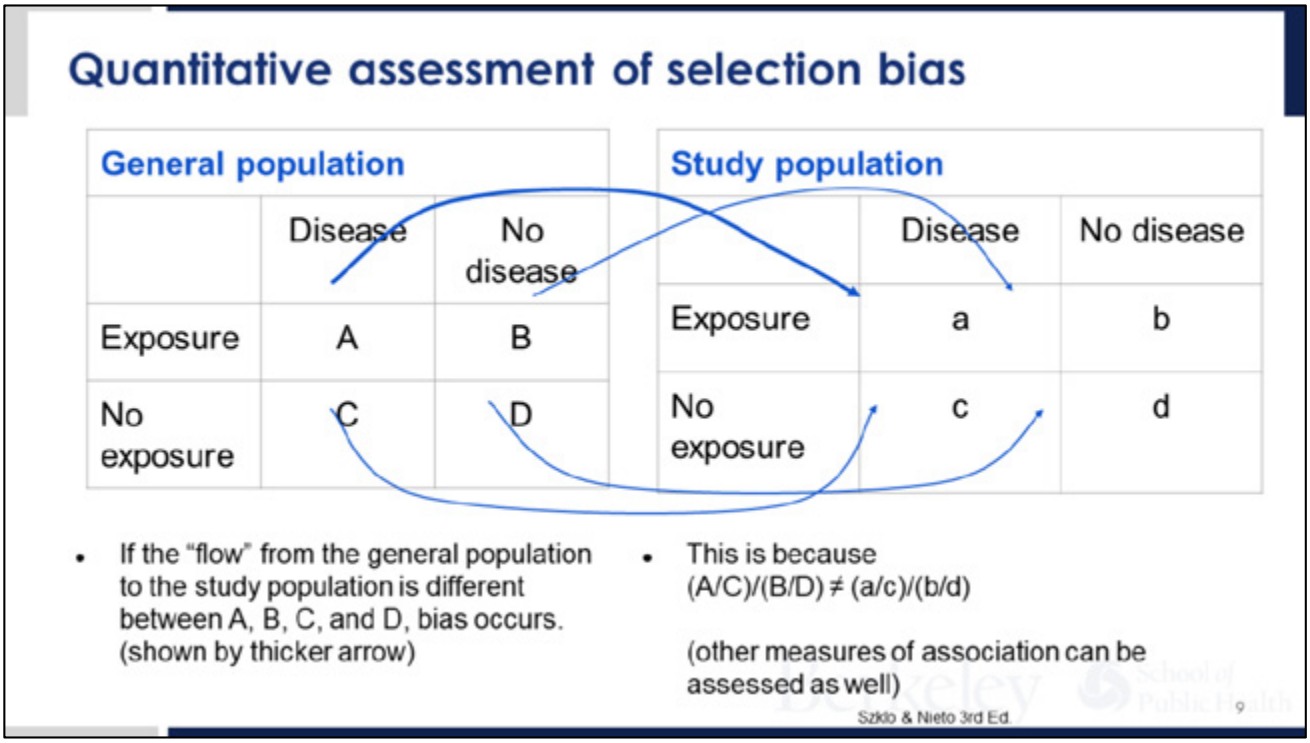
Let's look in a quantitative way at how to assess selection bias. Generally, when we do an epidemiology study, we wish that we were studying the general population. But we can't afford to do that. We can't afford to study everybody in a population. So what we do is we take a sample and that's called our study population.

Now these two two-by-two tables are meant to show us what's occurring when we move from working with the general population theoretically to working with the study population in actual, or practical terms.

In order to assess possible selection bias, we want to assess how well each element, each cell of the two-by-two table for the study population reflects where it sort of derived from the cell in the general population. So if the flow from the general population of the study population is completely equal for all cells, ABCD, there is no bias.

So in other words, if the population is a million people, the general population, and if we were able to study everybody, there were 300,000 people or 30% in cell A, if our study population is only 1,000 people, if study A is properly representing the general population in our sample, in our study population, then cell A would have 30% of 1,000 people, or 300 people.

We want these proportions of ABCD in the study population, the much smaller study population to be representative in the proper proportions to the general population.

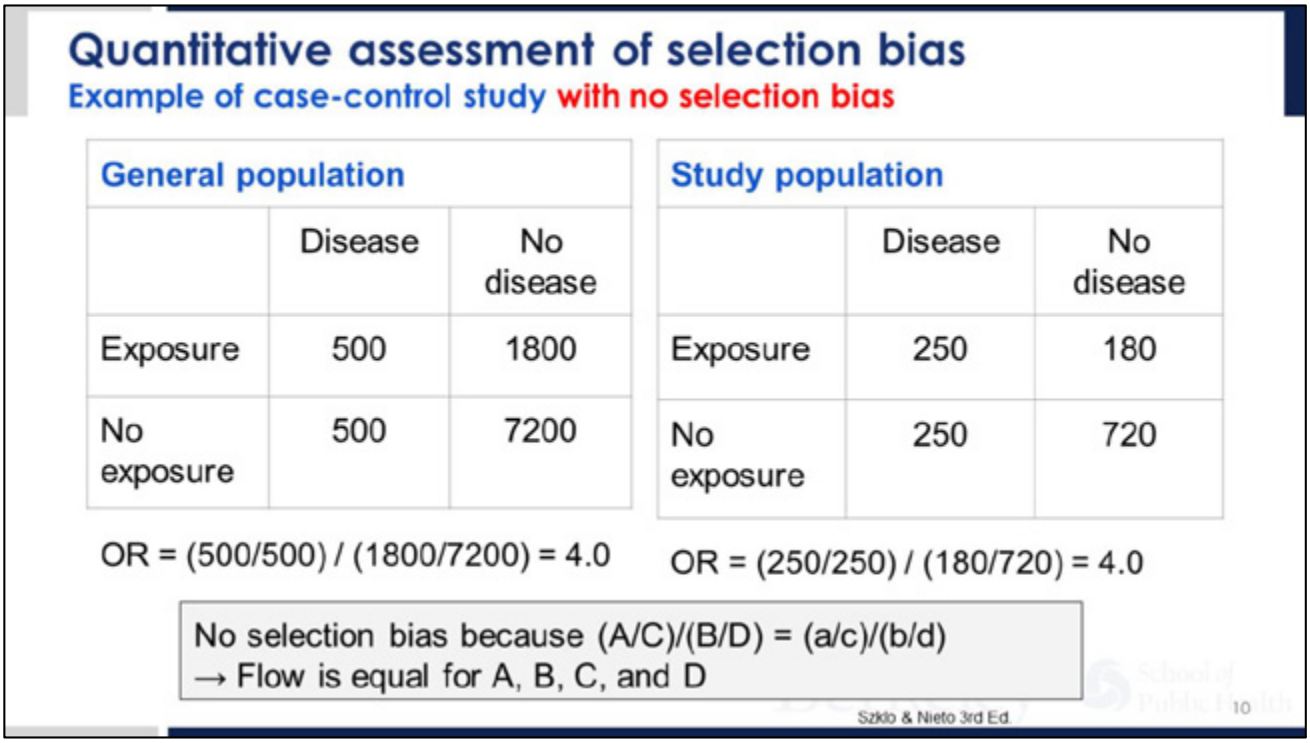


What happens if our selection of patients into our study is influenced in some way by any of the four specific types of selection bias we talked about earlier in this video.

Well, imagine that there's differential representation. So let's just say that cell A is overrepresented. There are too many people in cell A. And the other cells perhaps are correctly represented, but cell A has been distorted.

Now the relationship in cells AB and CD, for instance, if we were doing an odds ratio of the general population where the true odds ratio is A divided by C over B over D. You can see that in the study of the sample population, the study population, A over C divided by B over D is no longer going to be the same quantitative result because the cell A has been distorted by selection bias.

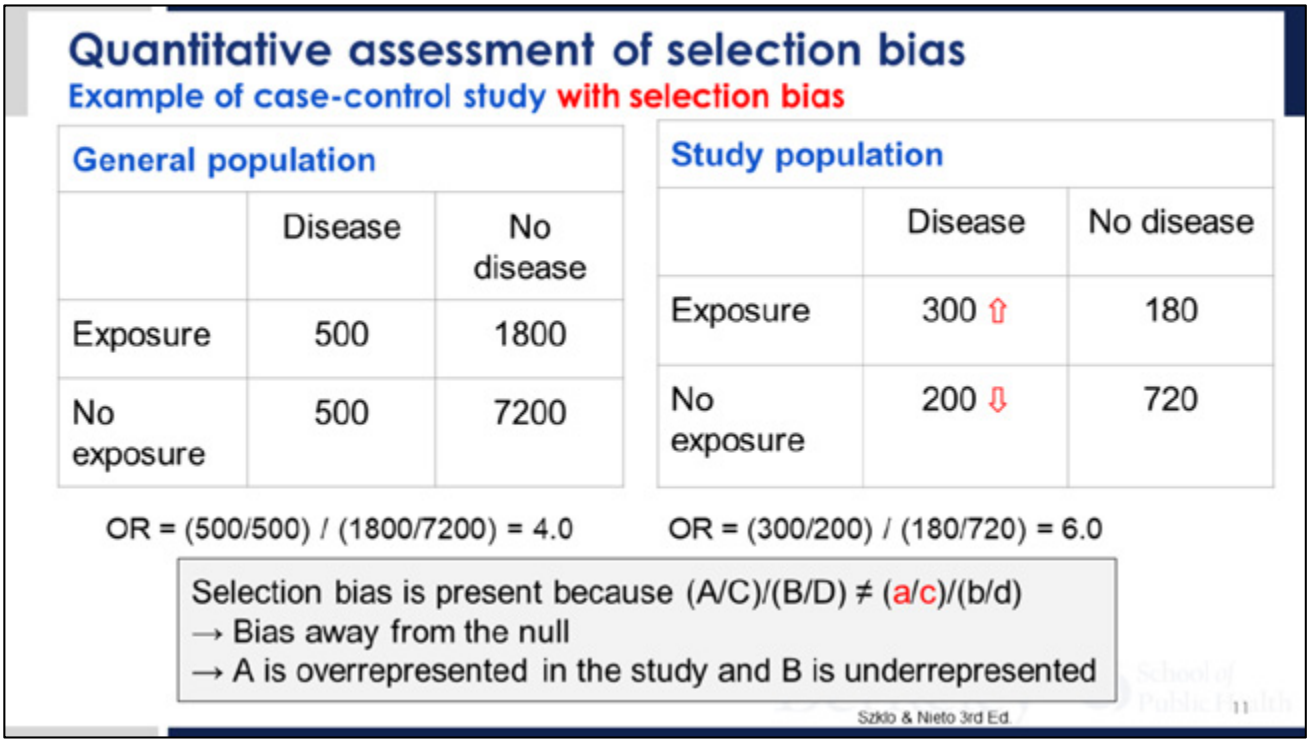
And this doesn't apply just to the odds ratio, you could do this for the cumulative incidence ratio, the incidence density ratio, the prevalence ratio, and so forth.



Here's an example of a case control study with no selection bias. On the left, we see the true or general population, the population we wish we could capture. It is everybody in the study. And if we do an odds ratio for that study, you come up with an odds ratio of 4. So let's call that in this theoretical example, let's call that an odds ratio that's the truth.

In our study population where we can't afford to sample everybody, but we take a sample of, among the diseased people maybe we take 50% of them. So instead of 500 exposed people with disease, we have 250. And instead of 500 diseased people with no exposure, we have 250. And for the non-diseased people, the farthest right column, we take a 10% sample. So now we have 180 and 720.

Now the odds ratio is calculated. And once again, because there was no differential selection into the different groups, we still see an odds ratio of 4. You can think of this as the flow was equal, both for the disease population and no disease population. If you drew arrows from cell A to cell little a and cell B to cell little b, the proportion represented was the same in both the exposed and non-unexposed.

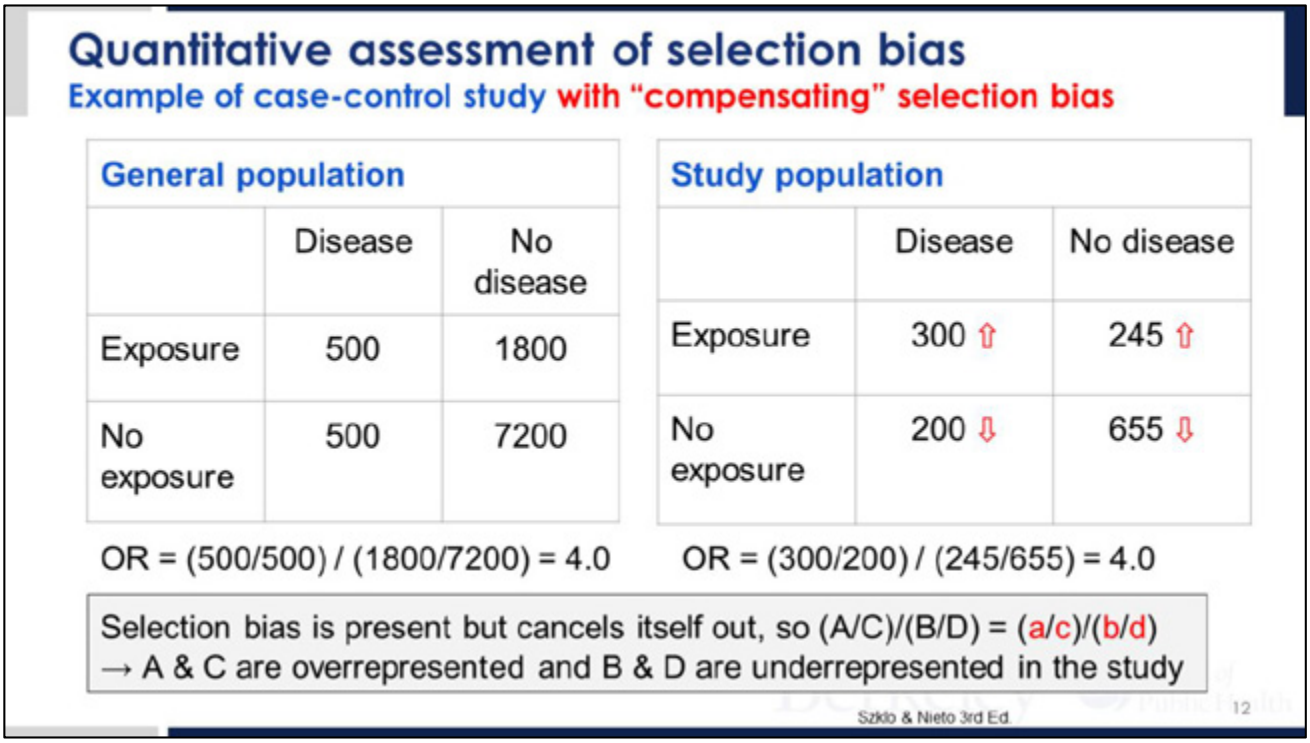


Now let's do an example where there is selection bias. So let's look at what happened here. We have the same truth, the odds ratio is 4 in the general population. But now when we've selected our 50% sample of the disease patients, there are 500 disease patients altogether in our study population.

But now instead of being equally balanced, 250 and 250 in cells A and C, they are now as you see, 300 and 200. So cell A has been overrepresented. Cell C has been under-represented.

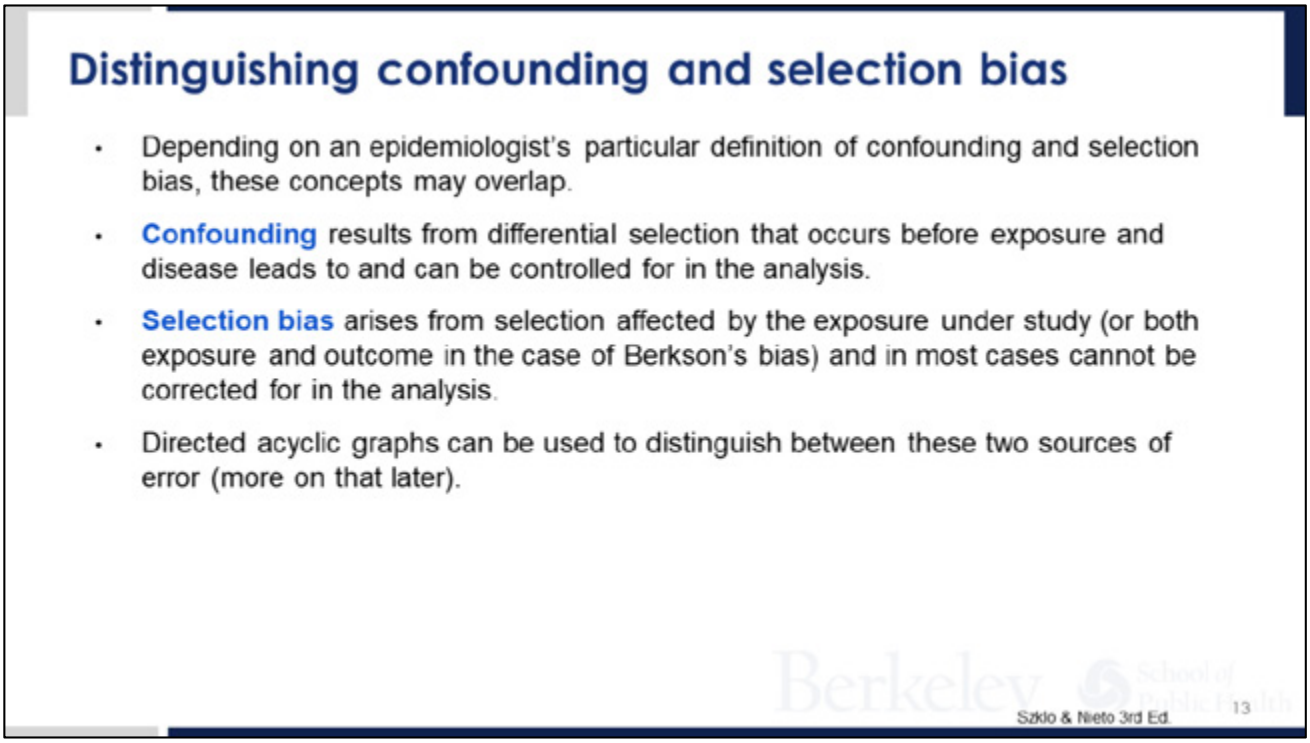
And when we put these new numbers into our odds ratio calculation, we come up with a conclusion that the odds ratio is 6. So that's our estimate from our sample, or our study population. But we know that's not true because the true value is 4.

This is because the selection bias that led to over-selection of exposed people with disease and under selection of non-exposed people with disease. And this is somewhat similar to the leukemia example earlier.



Now you can have a situation, and it's hard to actually know if this has happened. But you could have a situation in which there is both under and over selection, leading to a balancing or compensating selection bias that in a sense, accidentally comes up with the right answer.

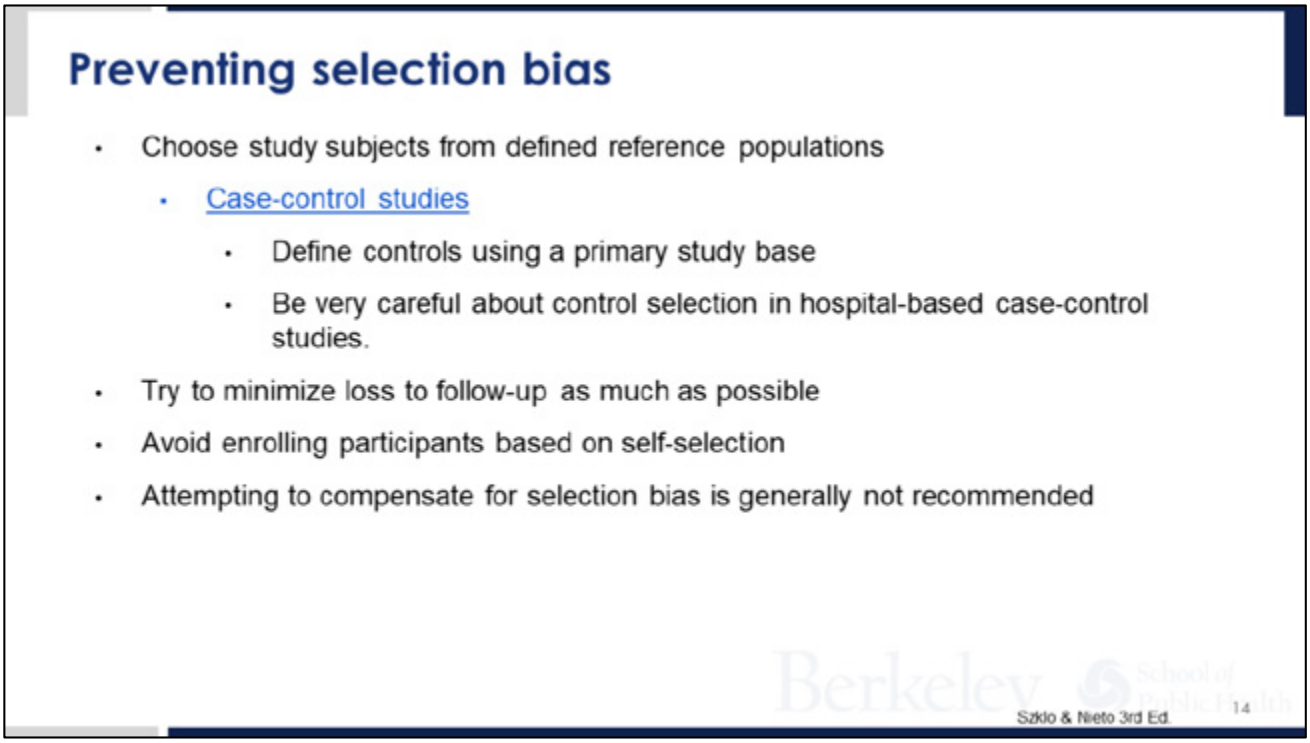
So the implication in this figure, these two, two-by-two tables is same general population. But now in our study population, we have under and over representation of the disease and no diseased populations as you see here with the up and down arrows. But the effect ends up leading us to a correct conclusion about the general population. Now I just want to say, it can be very hard to know that this is happening when you're doing a study.



We're going to talk much more about confounding in the course. But it'll be important to understand confounding and selection bias and how people think about these in different ways. So it sort of depends on your particular definition of confounding and selection bias. But that's what causes the overlap of these concepts.

Confounding results from differential selection that occurs before exposure and disease and leads to and can be controlled for in the analysis. Selection bias technically arises from selection affected by the exposure under study, or both exposure and outcome in the case of Berkson's bias and in most cases can't be corrected for in the analysis.

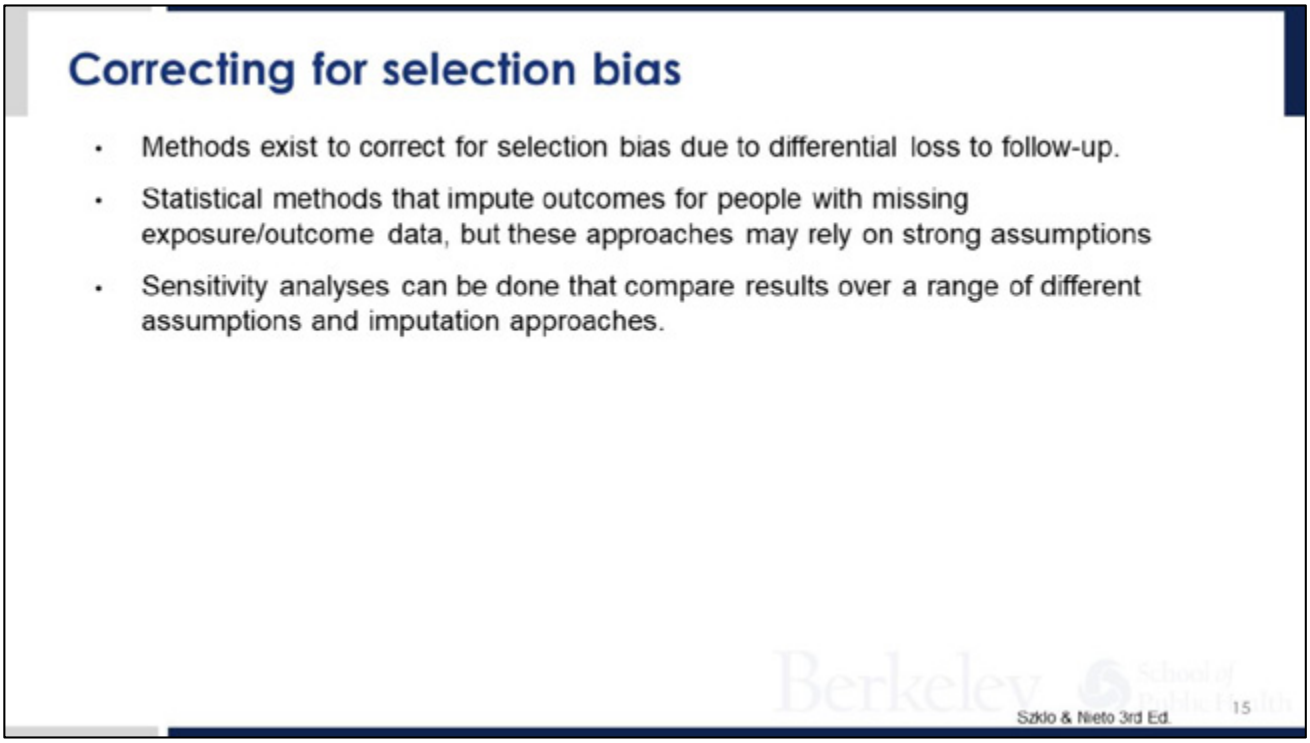
Directed acyclic graphs can be used to distinguish between these two sources of error. So once again, we're going to see how we bring in these causal inference tools to understand these epidemiology principles.



What can we do to prevent selection bias? Well, several things, we can choose study subjects from well-defined reference populations. If we're working in a case control study, we could define the controls using a primary study base. Go back and look at your earlier notes from 250w about what a primary study base is.

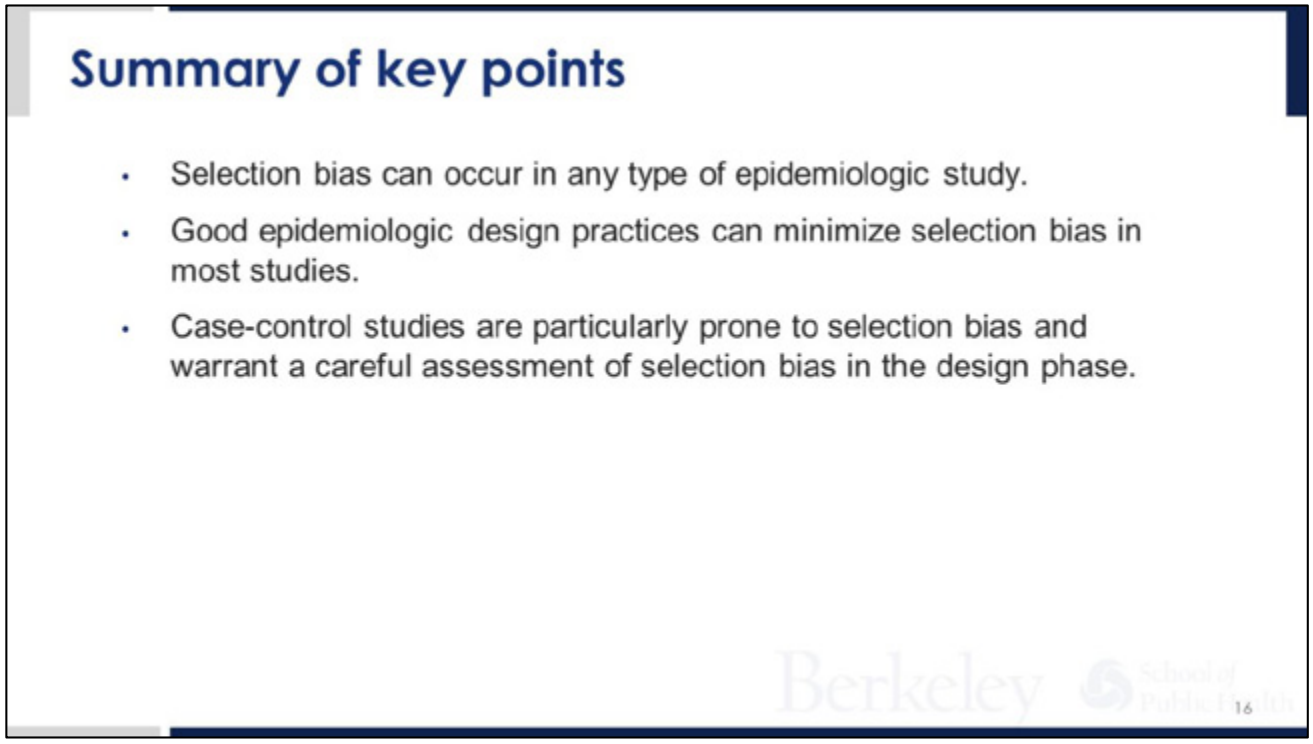
Be very careful about control selection and hospital case control studies, that's the Berkson's bias problem we've talked about. Next, we can try to minimize loss to follow up as much as possible. And we want to avoid enrolling participants based on self-selection for the reasons that we described earlier.

Finally, note that attempting to compensate for selection bias is generally not recommended. This is something you really want to get right at the design beginning of the study.



So to correct for selection bias, there are methods for the particular example where there's a differential loss to follow up. They are statistical methods that will impute outcomes for people with missing exposure or outcome data. But these approaches rely on some strong assumptions that can't always be met.

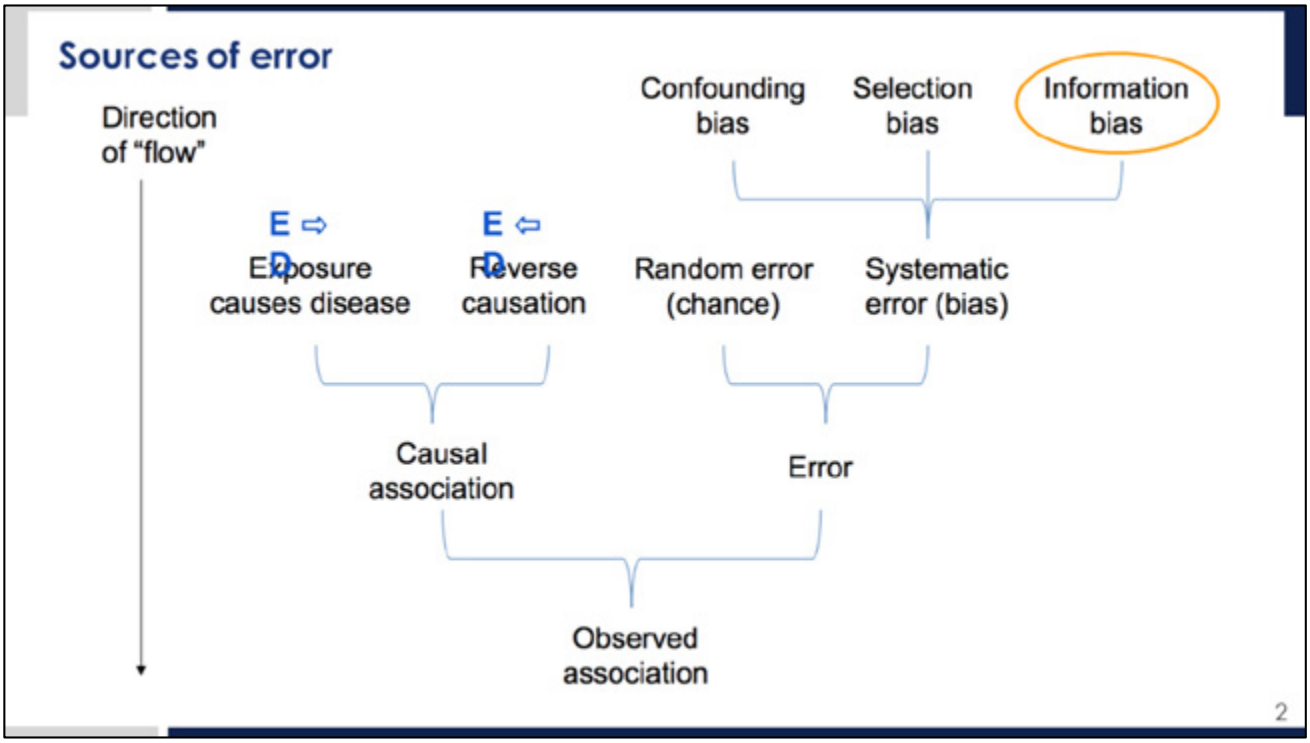
And we can do sensitivity analyzes that can be done that compare results over a range of different assumptions and imputation approaches. But again, wouldn't it be much better to avoid needing to use these techniques by avoiding selection bias in the design of the study?



So to summarize the key points in this talk on selection bias, it can occur in any type of epidemiology study. Good design practice can minimize selection bias in most studies. And case control studies are particularly prone to selection bias and warrant a careful assessment of selection bias in the design phase.

JACK COLFORD: In this video, I'll discuss information bias in epidemiology studies. Previously we discussed selection bias. And in the future, we'll be discussing confounding bias. This topic is information bias.

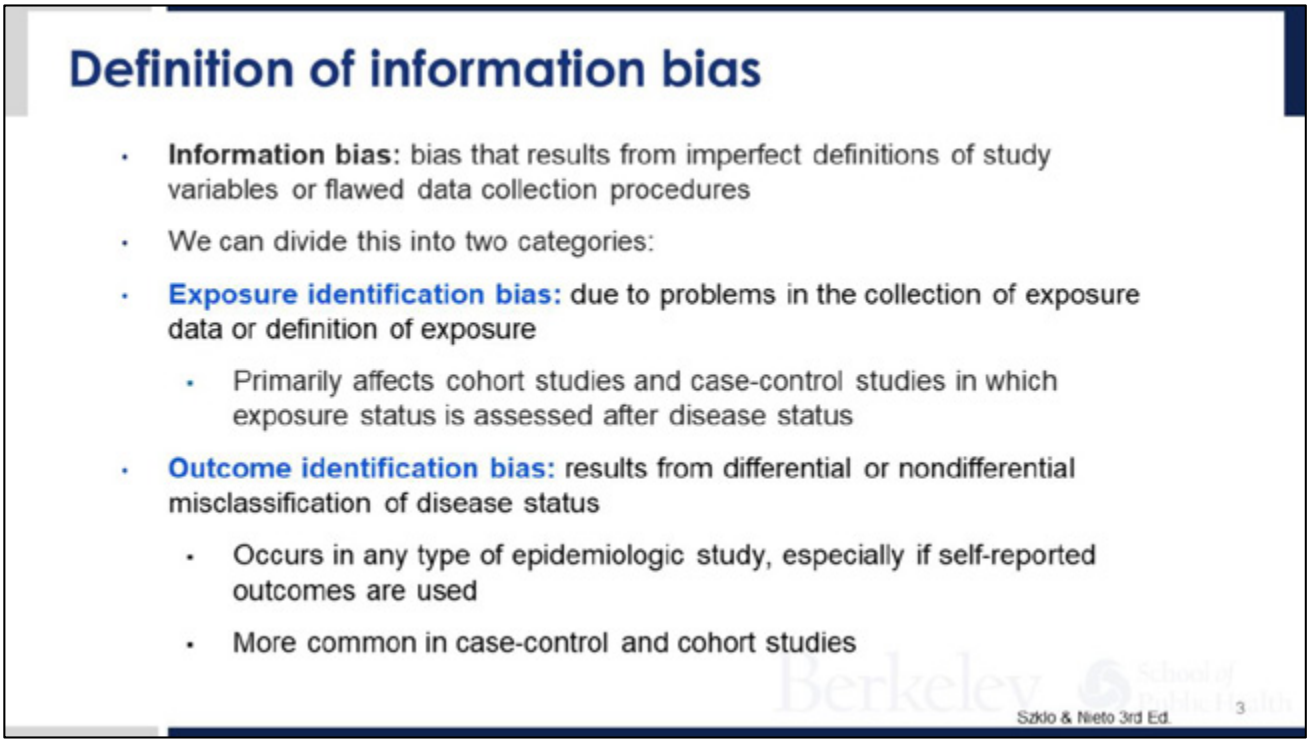
We'll focus on a couple of different types of information bias that can influence epidemiology studies-- in particular, two types that we call recall or respondent bias and a second type interviewer or observer bias. The result of a study suffering from information bias is the potential for misclassification of data about the subjects in the study. And this misclassification can itself either be non-differential or differential misclassification. We'll also talk about some ways to quantitatively assess how much misclassification is present in a study.



Let's start back with our summary slide of the sources of error, kind of the master plan of all the things that can go wrong in a study. We talked about how what we hope will happen is that exposure causing disease is what we are measuring properly. However, sometimes when we see an association we might have experienced reverse causation. That's an error in our study.

Other sources of error include random error, or chance, which, as I've mentioned before we try to eliminate through statistical means, and finally systematic error or bias of which there are three types that we're focused on in this course, confounding bias, selection bias, and information bias. We'll talk about information bias now.

These pathways down through either a causal association or through error lead us to whatever we observe. And if there is no error, then the observed association hopefully represents a causal association if we haven't experienced reverse causation.



Information bias results from imperfect definitions of study variables or flawed data collection procedures, or both things can happen in the same study. We can divide information bias into two categories based on our exposure or outcome. So exposure identification bias refers to problems in the collection of exposure data or in the definition of exposure.

This primarily is a problem for us in cohort studies or case control studies, in types of studies in which exposure status is assessed after disease status. So in other words, if we were going back and measuring exposure status in a participant who's already developed disease, we're at a high risk for exposure identification bias.

Outcome identification bias results from differential or non-differential misclassification of disease status itself. This can occur in any epidemiologic study, and especially when self-reported outcomes are used. So for example, if I'm measuring something that's totally dependent on the participant or are relative to a report, such as diarrhea, that's particularly subject to outcome identification bias. This is more common in case control and cohort studies.

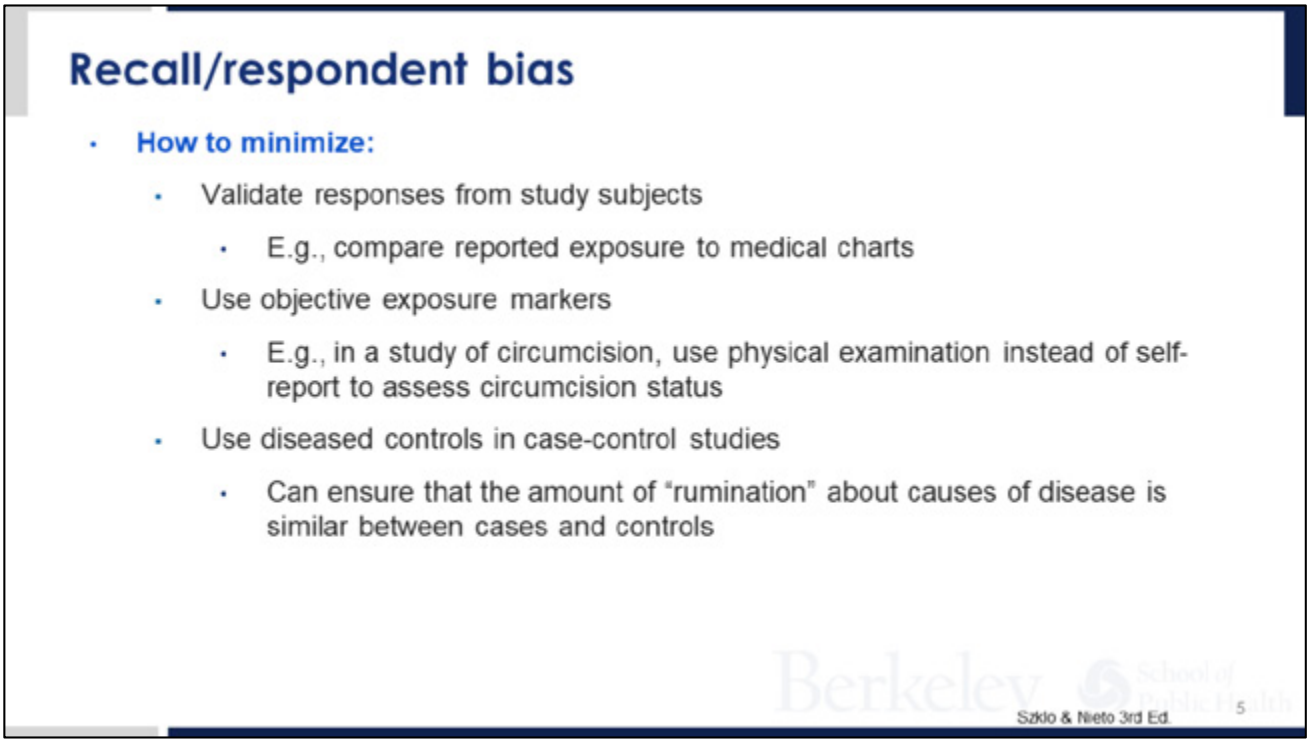
Recall/respondent bias

- Results from inaccurate recall of past exposure or disease.
- Particularly a concern in case-control studies when cases and controls know their disease status since cases and controls may differentially recall exposure status based on their disease experience.
- Can occur in any study design (e.g., a trial with self-reported outcomes)
- More likely to occur with long recall periods

Let's talk about recall or respondent bias. This is when a participant provides inaccurate recall of past exposure or disease that he or she has experienced. This is particularly a concern for us in case control studies because here, cases and controls know their disease status usually.

The cases and controls may therefore differentially recall exposure status based on their disease experience. So for example, if a mother is being asked about exposure to pesticides of her child, if her child has already been diagnosed with leukemia, her memory of what the child was exposed to may become much more acute, much different than a mother whose child does not have leukemia.

This can occur in any study design. And for example, it can occur even in a trial with self-reported outcomes. So even though a trial is a very powerful design, if self-reported outcomes are present, it's still subject to this recall, or respondent bias. And this is more likely to occur and gets worse with long recall periods.



There are some techniques for trying to minimize recall or respondent bias. One is to validate the responses that are gotten from study subjects. And for example, in a subsample you might compare the recall of the participant to the medical chart to see if their recall of what was recorded at a prior time is accurate.

One could use objective exposure markers to reduce recall or respondent bias. For example, in a study of circumcision, physical examination rather than self-report could be used to assess circumcision status in children.

And finally, use disease controls in case control studies as a way to minimize recall or respondent bias. So here you can ensure that the amount of thinking or rumination about the cause of disease is similar between cases and controls.

Interviewer/observer bias

- Occurs when interviewers are not masked:
  - with regard to disease status of study participants when measuring exposure status (e.g., in a case-control study)

    or
  - with regard to exposure status of study participants when measuring disease status (e.g., in a cohort study or trial)
- Can occur if interviews ask or clarify questions in a different way depending on exposure / disease status
- E.g., in a trial if the interviewer knows the intervention arm a study participant is in, they may consciously or unconsciously ask questions in a way that bias results towards a beneficial effect

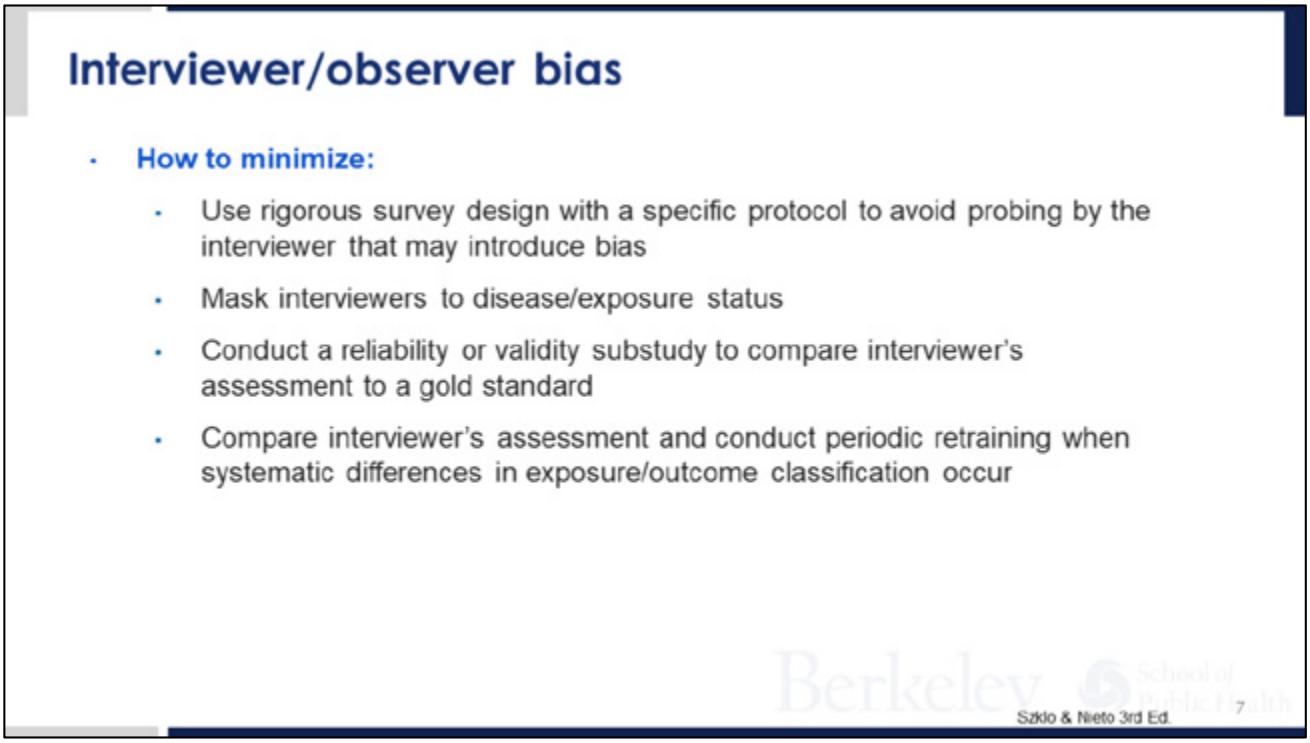
Berkeley School of Public Health

Szklo & Nieto 3rd Ed. 6

Another type of information bias is interviewer or observer bias. This often occurs when the interviewers are not masked. And by masked, I mean they know either the exposure or the outcome status of the participant, and it might influence the interviewer.

With regard to disease status of a study participant when measuring exposure status, this can occur in a case control study that an interviewer may introduce a bias into the reporting of the participant's condition, or with regard to exposure status of the study participants when the interviewer is measuring disease status. This might occur in a longitudinal study like a cohort study or a trial.

Interviewer or observer bias can occur if interviewers ask or clarify questions in a different way, depending on exposure or disease status. So if an interviewer knows, for example that a participant has disease, an interviewer then asks more pointed or more detailed or longer questions or different questions, that could introduce differential classification of the patients. And therefore is a type of interviewer or observer bias.

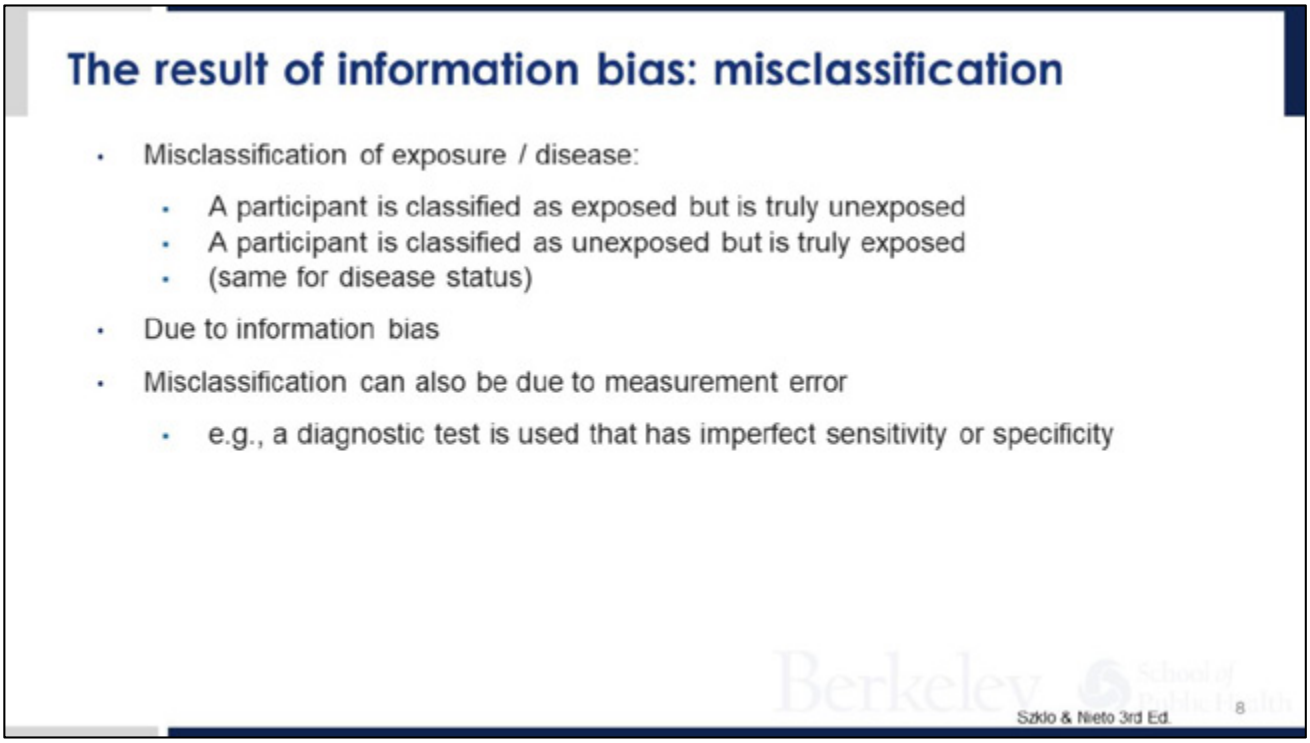


There are some techniques for trying to minimize interviewer or observer bias. Epidemiologists try to use rigorous survey designs with specific protocols to try to avoid probing by the interviewer in order to avoid that interviewer bias that can be introduced.

We can mask interviewers to the disease or exposure status of the participant when this is possible. This is a good technique for minimizing interviewer or observer bias. Epidemiologists can conduct a reliability or validity sub-study to try to compare the interviewers assessment to a gold standard in a very small or affordable subsample of the participants.

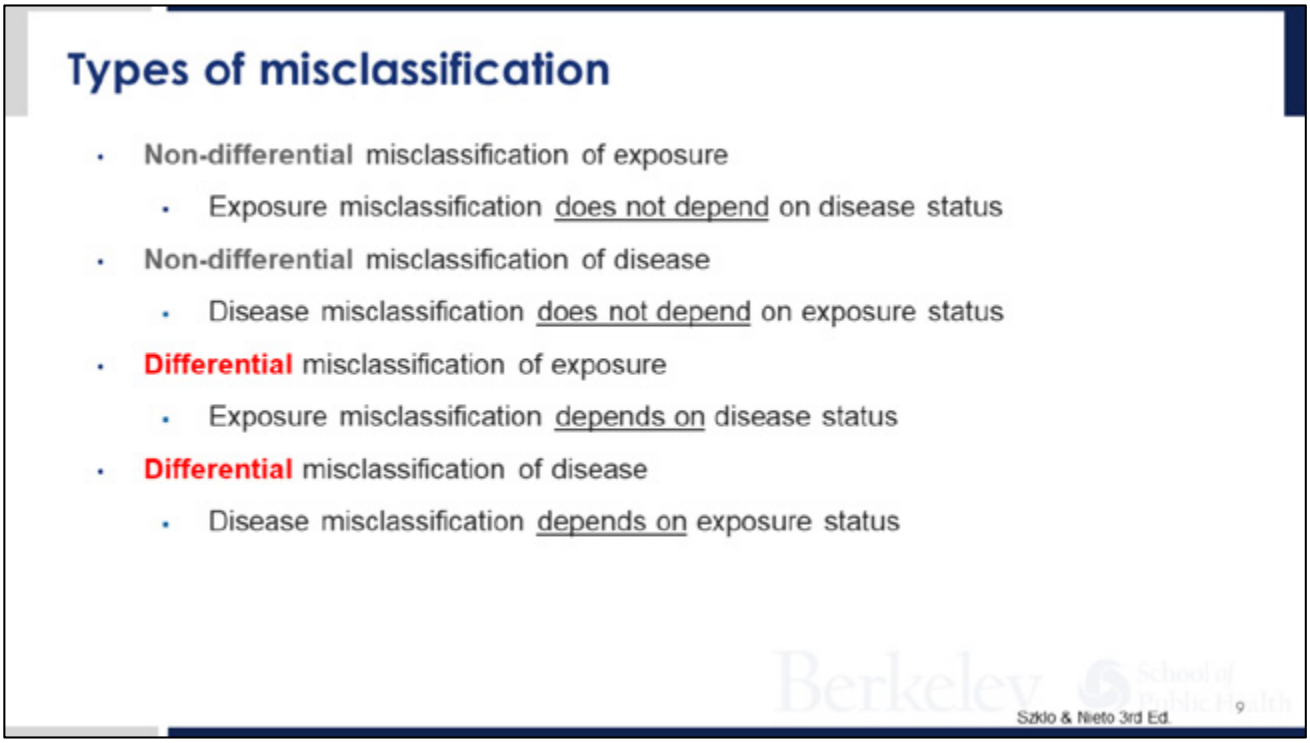
The epidemiology team could compare their responses obtained by interview with their truth if measured through some gold standard, a test or a medical chart or some other way of knowing for sure what really happened.

And then finally, in order to reduce interviewer or observer bias, we can compare the interviewer's assessment and conduct periodic retraining and systematic differences in exposure or outcome classification occur.



One result, an important result of information bias is misclassification. Misclassification can occur for either exposure or disease. For example, if a participant is classified as exposed but is truly unexposed, that's a misclassification. If a participant is classified as unexposed but is truly exposed, that's a misclassification. These same problems apply if this is with respect to disease status. So this is a type of information bias.

Misclassification can also be due to measurement error. For example, a diagnostic test that one is using might have imperfect sensitivity or specificity. So the test is then giving a wrong answer because it is an imperfect test due to its poor sensitivity or specificity.



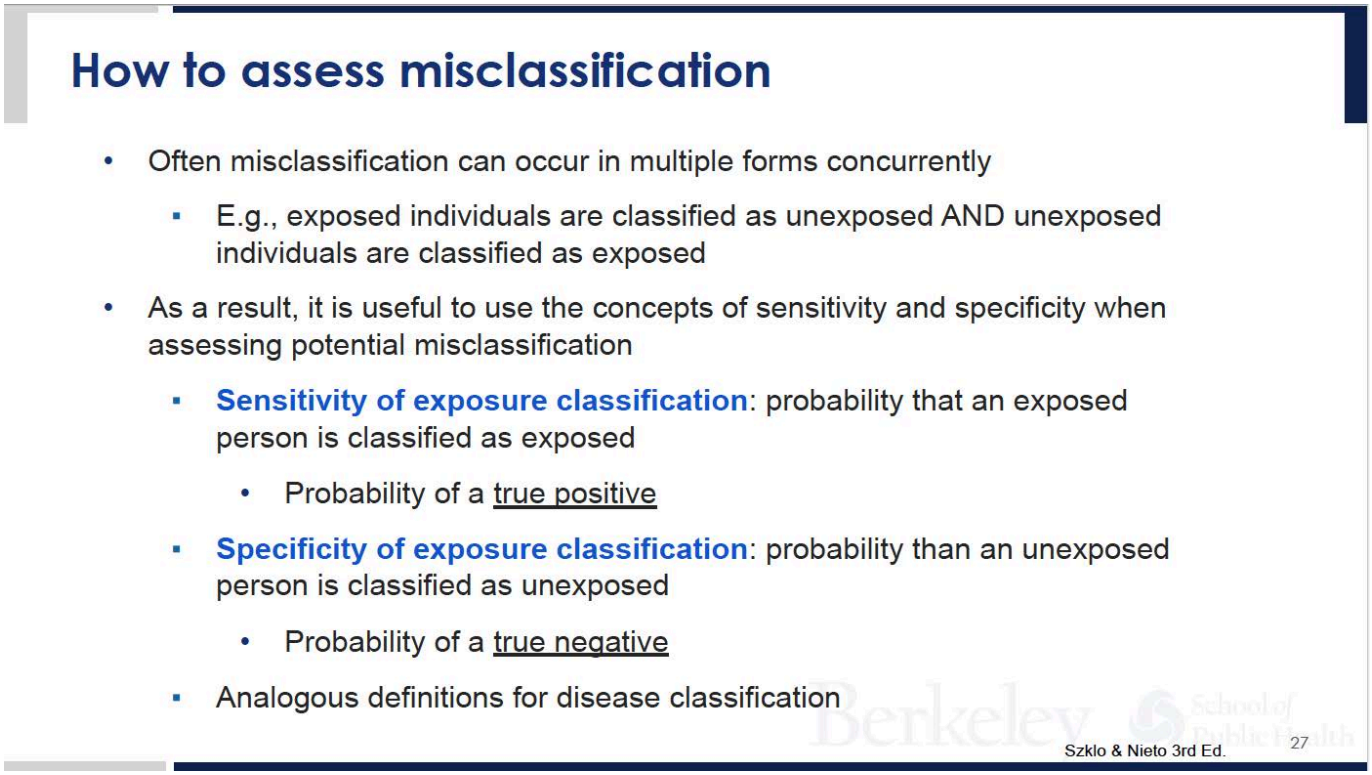
Misclassification can be categorized into differential or non-differential and this can apply to either exposure or disease. Non-differential misclassification of exposure occurs when exposure misclassification does not depend on disease status. So there is some misclassification occurring but it's not related to disease status.

Non-differential misclassification of disease is a situation in which the disease misclassification is occurring but isn't dependent upon the exposure status. Let's contrast that out here in the red with differential misclassification of exposure and differential misclassification of disease.

In the situation with differential misclassification of exposure, the exposure misclassification here is dependent somehow on disease status. So that, for example here, this would be if participants in a lung cancer and smoking study, if the participants with lung cancer had their smoking exposure measured with a tool or lab test that was different than the way smoking was measured in the non-lung cancer patients.

And finally, differential misclassification of disease can occur, and when it occurs, disease misclassification is dependent upon the exposure status of the participants.

Here the situation might be, if you think of smoking and lung cancer again, if smokers have much more rigorous measurement of their lung cancer status. For example, let's say that smokers in a study have MRI examinations done on their lungs but nonsmokers only have chest x-rays, that's a differential misclassification that can be introduced dependent upon the exposure status.

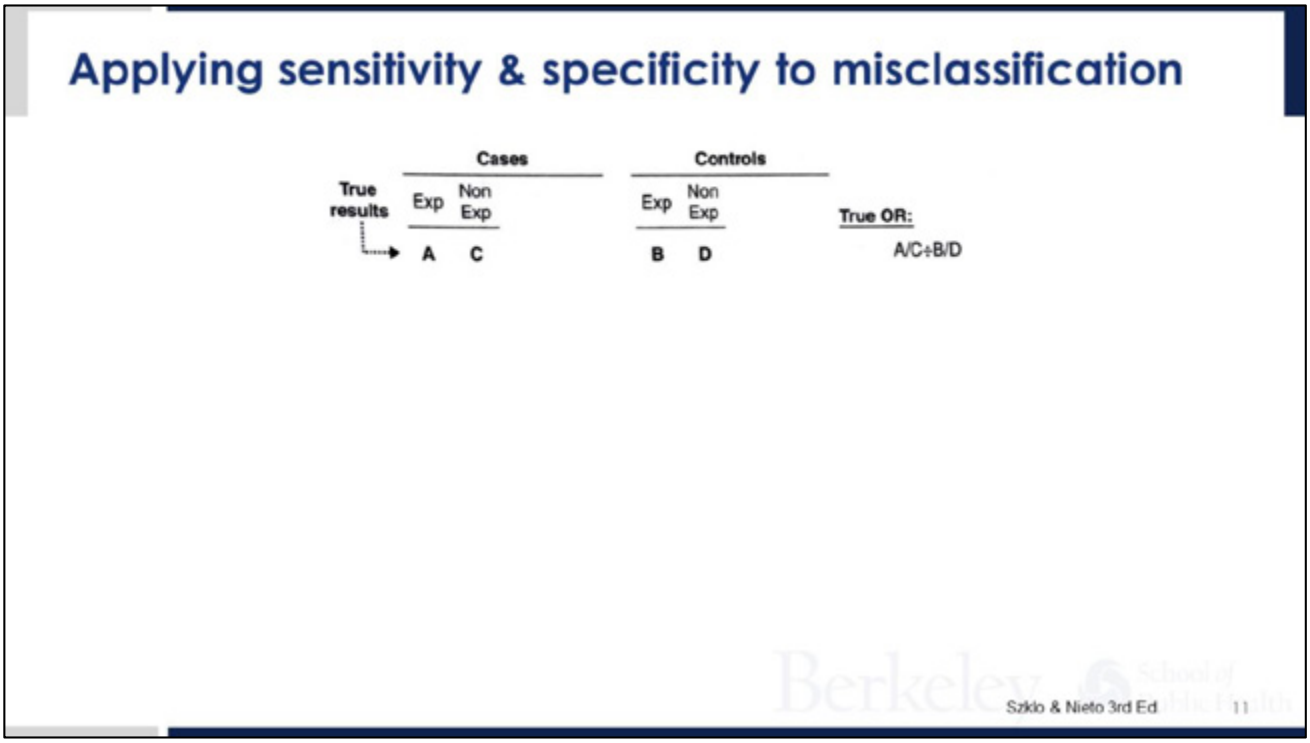


In order to assess misclassification, we have to think about the fact that sometimes these can occur concurrently, so several types of misclassification. Exposed individuals might be classified as unexposed, and unexposed individuals are classified as exposed. So there's multiple problems going on here.

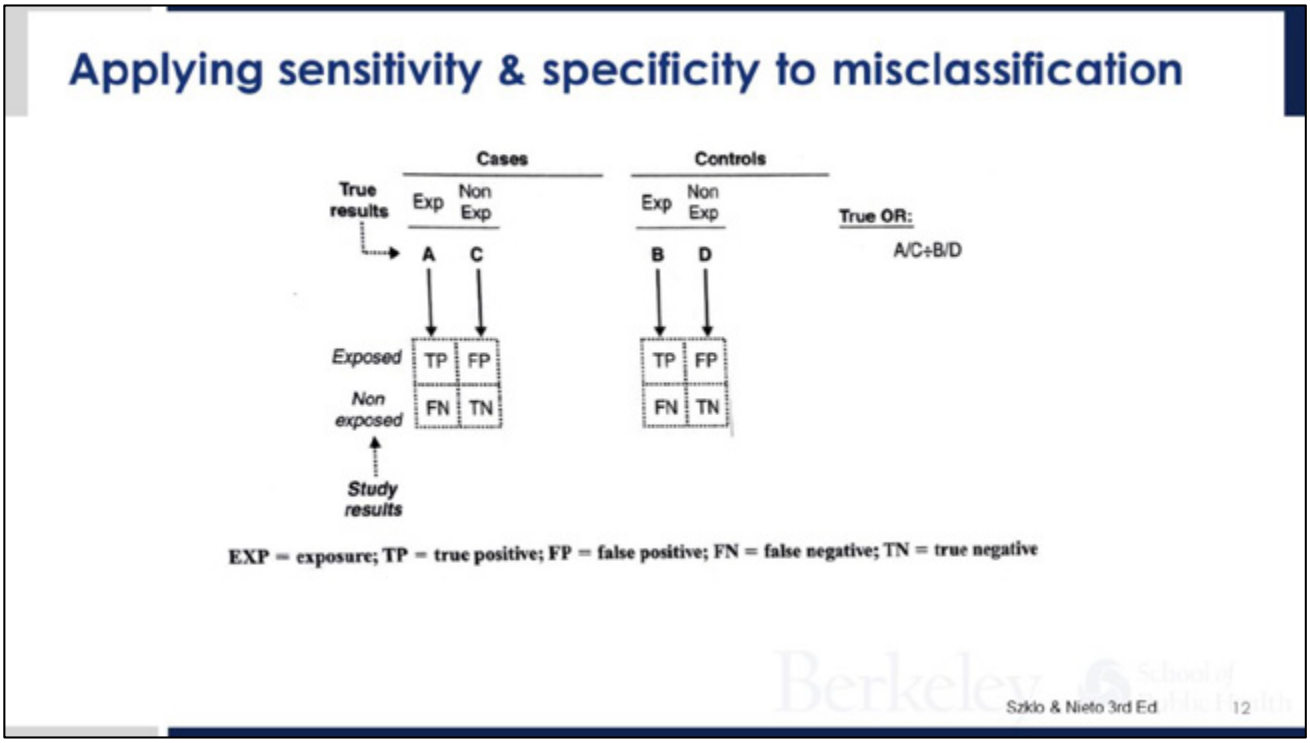
As a result, it's useful to use the concepts of sensitivity and specificity when assessing potential misclassification. The sensitivity of exposure classification is the probability that a person classified as exposed is truly exposed. We might call this the probability of a true positive. This is sensitivity.

And the specificity of exposure classification is the probability that a person classified as unexposed is truly unexposed, or you might think of this as the probability of a true negative classification.

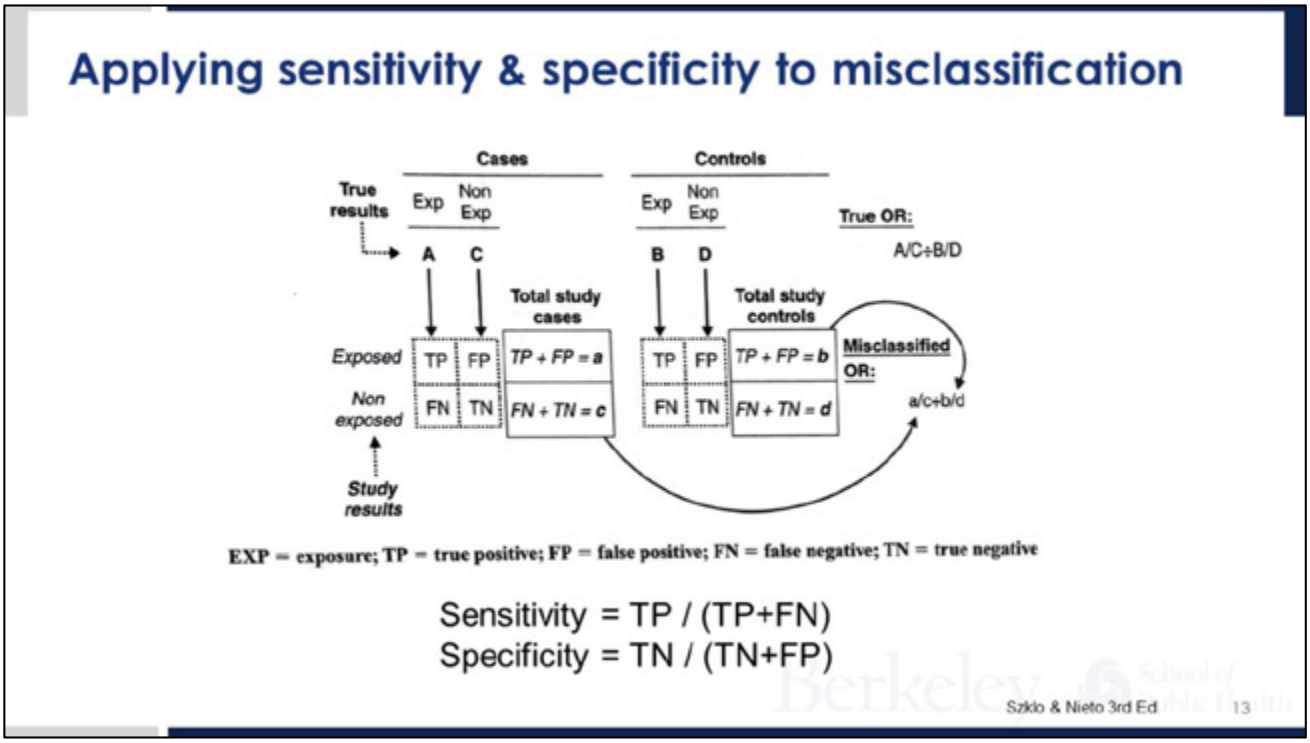
You can make analogous definitions for disease classification with the probability of a true positive classification and the probability of a true negative classification for disease status.



Let's look at how this use of sensitivity and specificity can be applied to misclassification status. We're seeing here tables that are going to develop for cases and controls, the true results for the exposed and the non-exposed. We see here this is essentially a two-by-two table but written out as a row. AC BD, A over C divided by B over D is our true odds ratio. And we know this to be the truth.



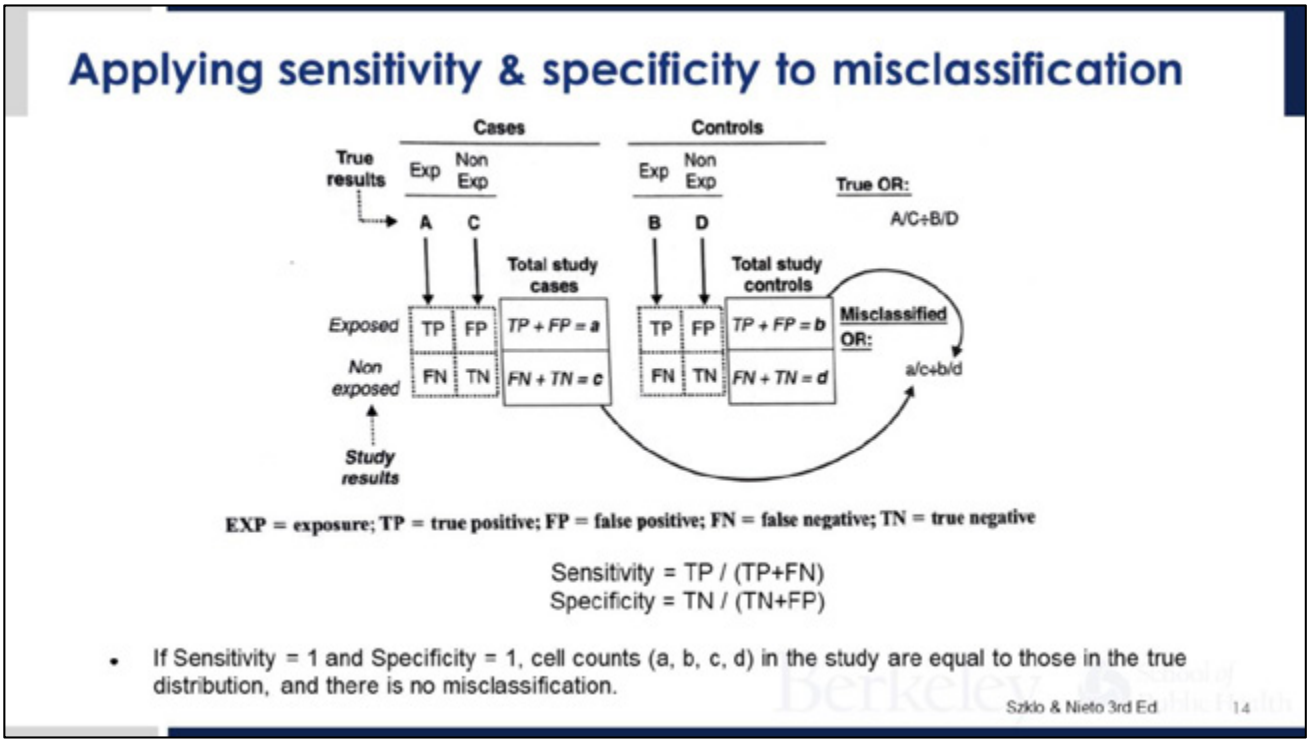
But if we do a study in which we have misclassified participants, both the cases and controls, what is going on here is we're showing what's happening to the classification of each cell, A, C, B, and D. Based on the true positivity and false positivity and true negativity and false negativity of each cell that has been measured. So in other words, we're looking at all four cells here and applying the sensitivity and specificity to all four of the cells.



Now when we have used the sensitivity and specificity of all four of the cells, we have new values for A, B, C, and D. So the total study cases who are exposed would be the sum of the true positives and the false positives. Let's call that little a. And the non-exposed would be the false negatives plus the true negatives. Let's call that little c. And similarly, we develop those values for the control.

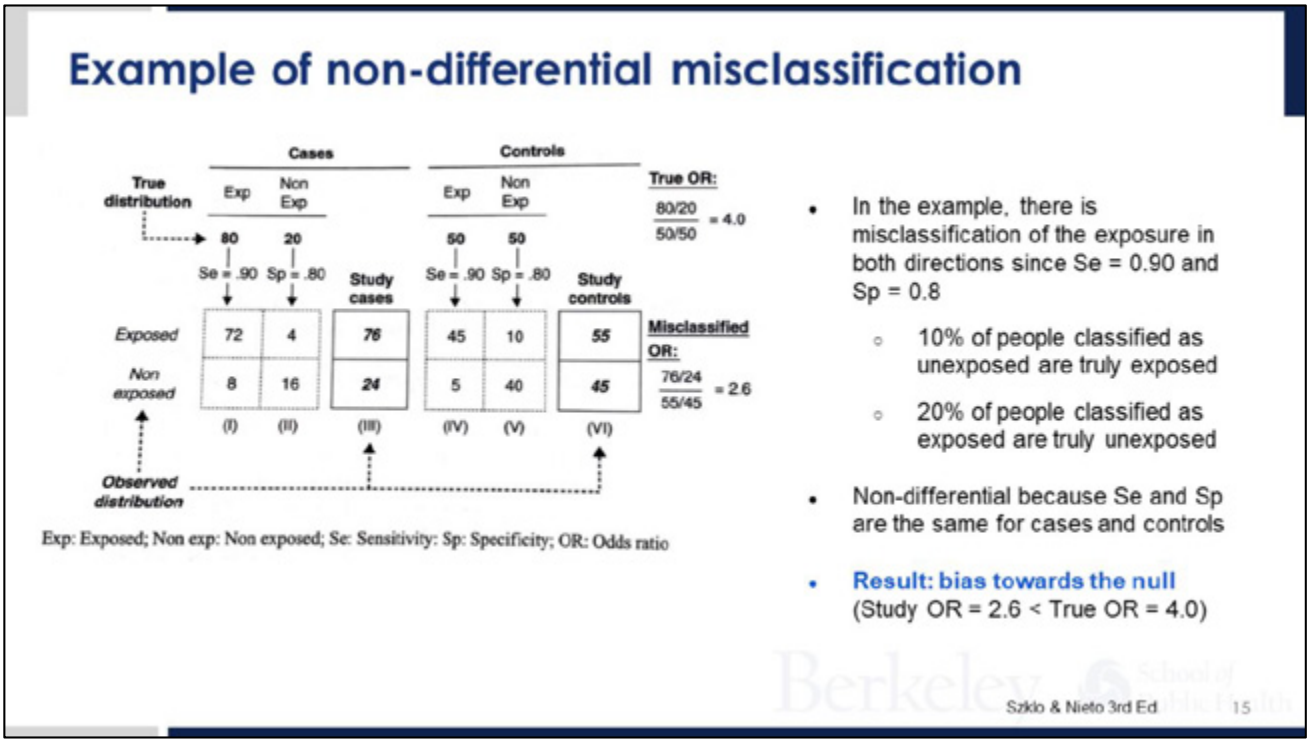
So for cells B and D, we now have little b and a little d. When we put all of these together, the little letters, a over c divided by b over d, gives us a misclassified odds ratio. So the true odds ratio is represented by the capital letters a over c divided by b over d. The misclassified odds ratio is represented by the numbers for little a over little c divided by little b over little d.

And just to recall at the bottom here, the sensitivity and specificity are defined as we learned in an earlier course, the true positive over true positive plus false negatives is the sensitivity. And the specificity is the true negative divided by the quantity true negative plus false positive.



One way to understand this misclassified odds ratio is, think what the situation is if sensitivity and specificity are perfect. If our test is perfect, sensitivity is equal to 1, and specificity is equal to 1. Then the cell counts, little a, little b, little c, little d, turn out to be capital A, capital C, capital B, and capital D. So in the situation with a perfect test and no misclassification, the odds ratio represented by the little letters is the same as the odds ratio represented by the big letters, as you'd expect.

But since sensitivity and specificity are seldom equal to 1, whatever the values of sensitivity and specificity are going to change the values of little a, little b, little c and little d. So the misclassified odds ratio will generally be different than the true odds ratio.

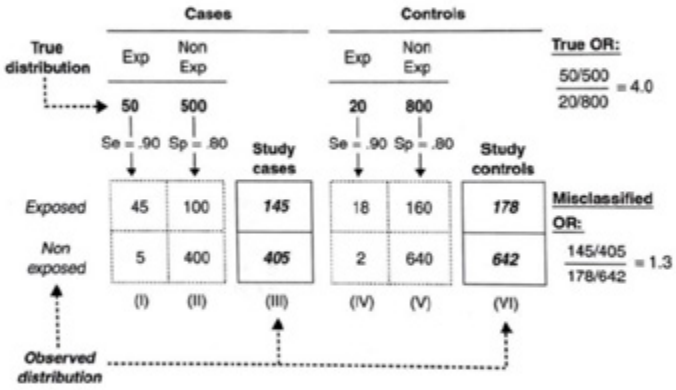


Here's an example of non-differential misclassification. In the first row again, we see the truth. So we for some reason know exactly what the right answer is.

Here in this example, this thought experiment, there's a sensitivity of 0.9 of the test and a specificity of 0.8. So let's apply those sensitivities and specificities to our cells A, B, C, and D, which were 80, 20, 50, and 50. And look at what happens.

So 10% of the people classified as unexposed are truly exposed. That's because of the sensitivity of 0.9. 20% of the people who were classified as exposed are truly unexposed. That's because of the specificity. This is a non-differential misclassification because sensitivity and specificity, the 90 and the 80 are the same for both the cases and the controls.

Example of non-differential misclassification with low prevalence of exposure



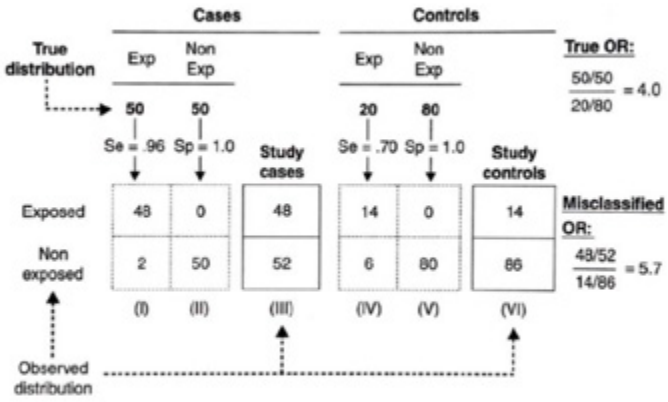
Exp: Exposed; Non exp: Non exposed; Se: Sensitivity: Sp: Specificity; OR: Odds ratio

- Same as previous example but the exposure is less common (26% of cases and 22% of controls are exposed)
- Non-differential because Se and Sp are the same for cases and controls
- **Result: stronger bias towards the null**
  (Study OR = 1.3 < True OR = 4.0)

Let's look at an example of non-differential misclassification when there's a low prevalence of exposure. This situation here is the same as the previous example, but here the exposure is less common. Only 6% of the cases and 22% of the controls are exposed.

This is non-differential. Again because sensitivity and specificity are still the same for cases and controls. So what we see here is that the bias is stronger towards the null. And this is because the exposure is less common.

Example of differential misclassification with imperfect sensitivity and perfect specificity



Exp: Exposed; Non exp: Non exposed; Se: Sensitivity; Sp: Specificity; OR: Odds ratio

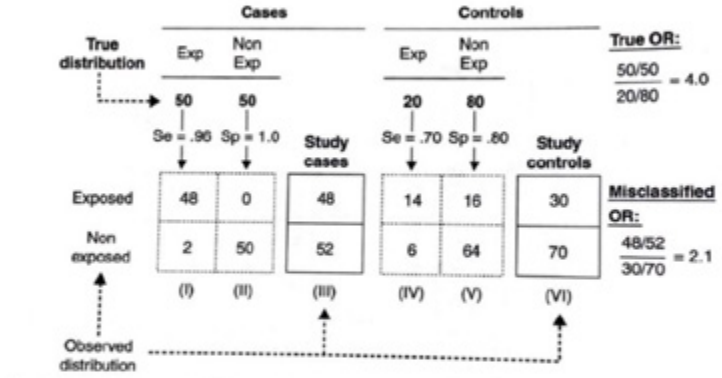
- In the example, there is misclassification of the exposure in one direction since Se<1 and Sp=1
  - 4% of cases and 30% of controls classified as unexposed are truly exposed
- Differential because Se differ between cases and controls
- **Result: bias away from the null** (Study OR = 5.7 > True OR = 4.0)

Let's shift now to an example with differential misclassification where we have imperfect sensitivity but perfect specificity. So the specificity is 1.0 but the sensitivity is going to be less than 1.0. Here, 4% of the cases and 30% of the controls are classified as unexposed even though they are truly exposed.

This is a differential misclassification because this misclassification is different for the cases than it is for the controls. The result here is a bias away from the null. Here the study odds ratio is 5.7, and that's greater than the true odds ratio of 4.0.

But let's remember, or let me make the point that differential misclassification is unpredictable in its direction. So it won't always be away from the null, unlike the non-differential misclassification we discussed earlier that was always toward the null.

Example of differential misclassification with imperfect sensitivity and specificity



Exp: Exposed; Non exp: Non exposed; Se: Sensitivity; Sp: Specificity; OR: Odds ratio

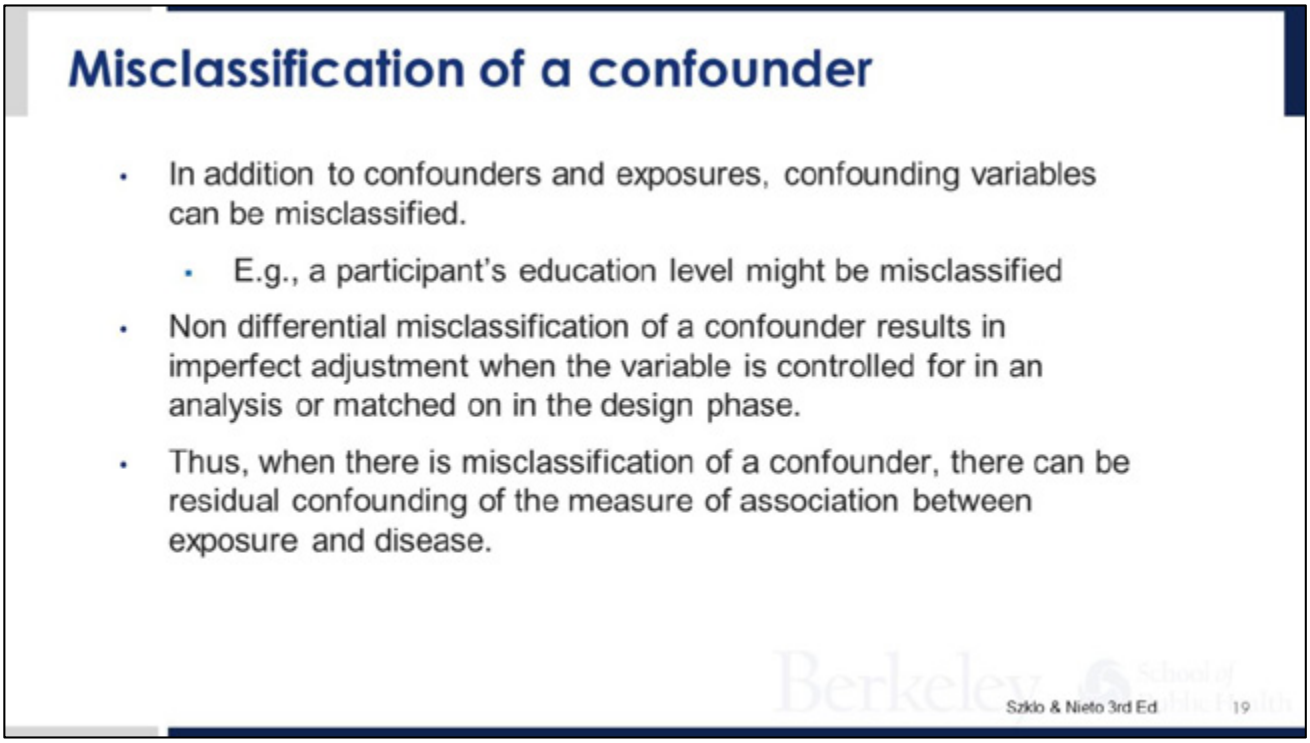
- In the example, there is misclassification of the exposure in both directions since neither Se or Sp equal 1
- Differential because Se and Sp differ between cases and controls
- **Result: bias towards the null** (Study OR = 2.1 < True OR = 4.0)

Berkeley School of Public Health

Szklo & Nieto 3rd Ed 18

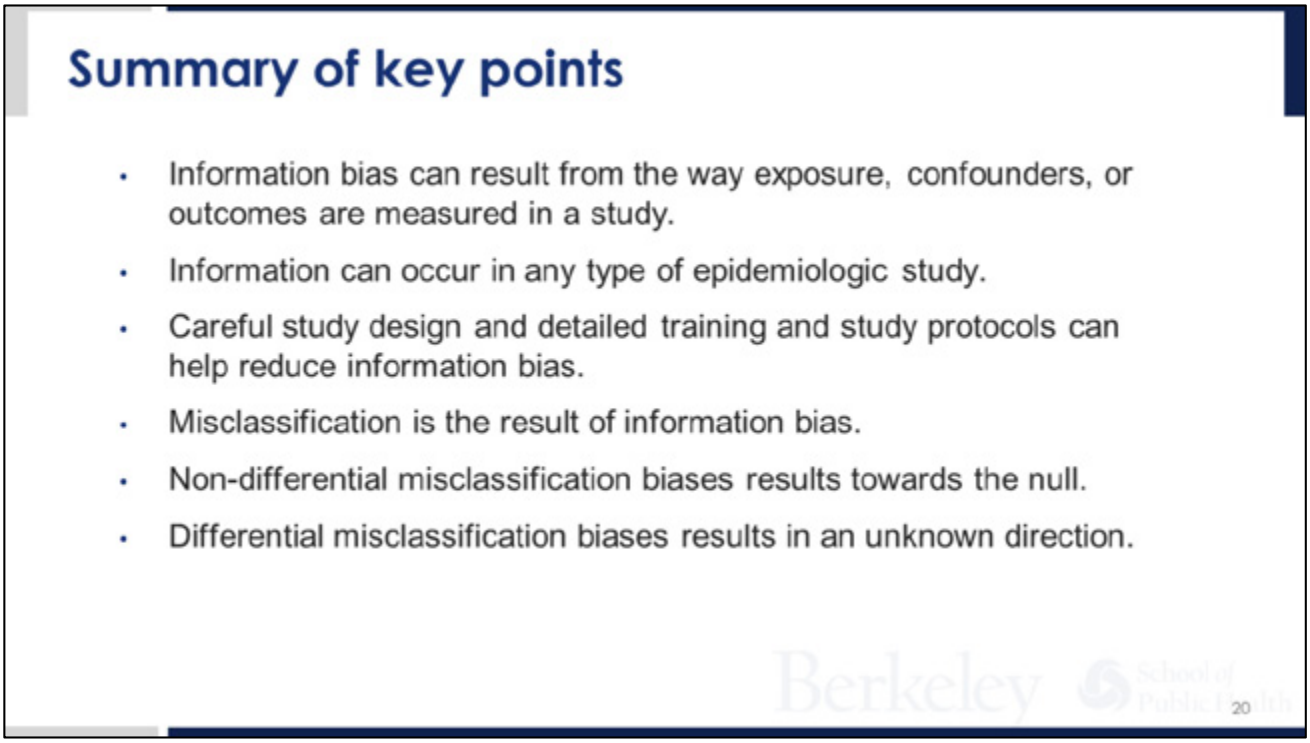
Let's now do an example of differential misclassification with both imperfect sensitivity and imperfect specificity. Things are moving in both directions here because sensitivity and specificity are both not perfect.

Here, working through the example, the result is a bias toward the null. And the study odds ratio is 2.1, where once again, the true odds ratio was 4.0. So once again, to make the point that with differential misclassification, it's unpredictable what's going to happen to the direction of the misclassification.



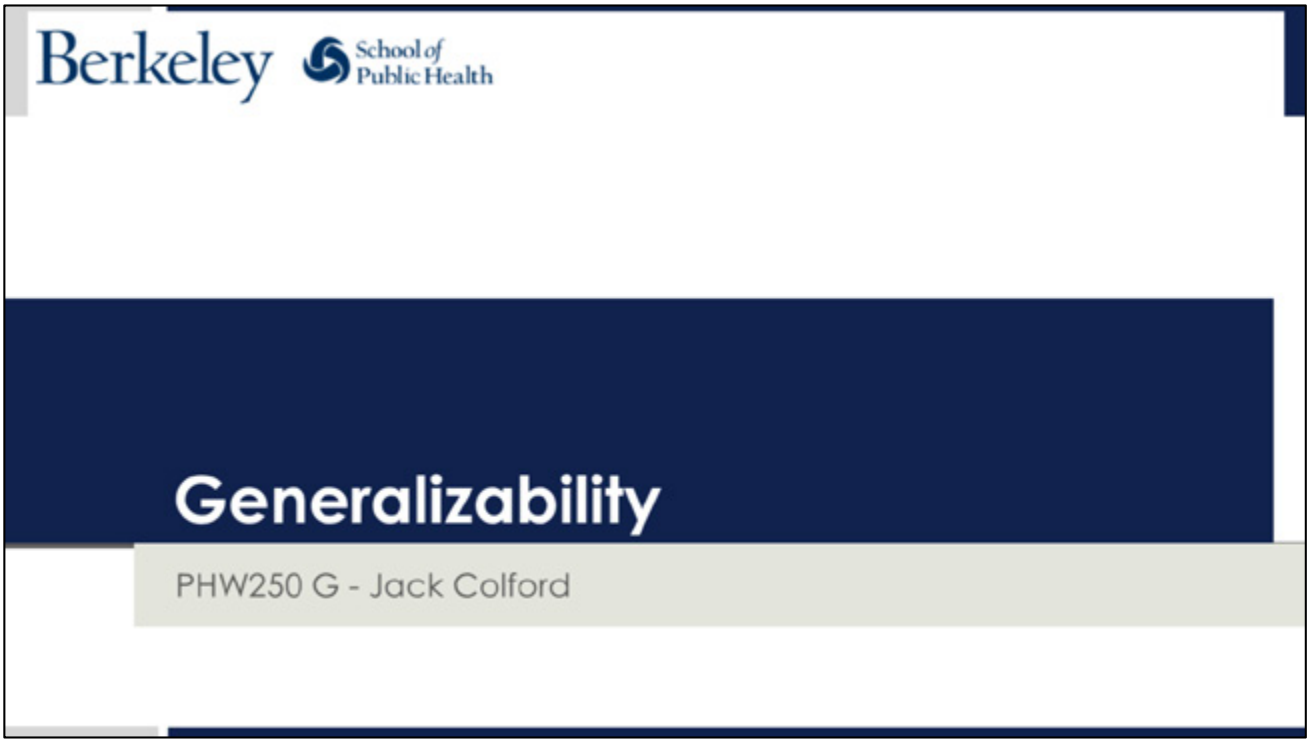
In addition to Misclassification of our exposure or our outcome, we could also have misclassification of a confounder. For example, education might be a confounding variable in a study of smokers and lung cancer, or in many studies. And education might be misclassified.

So non-differential misclassification of a confounder results in imperfect adjustment when the variable is controlled for in an analysis, or is matched on in the design phase. When there's misclassification of a confounder, there can still be residual confounding of the measure of association between exposure and disease because of the misclassification of the confounder.

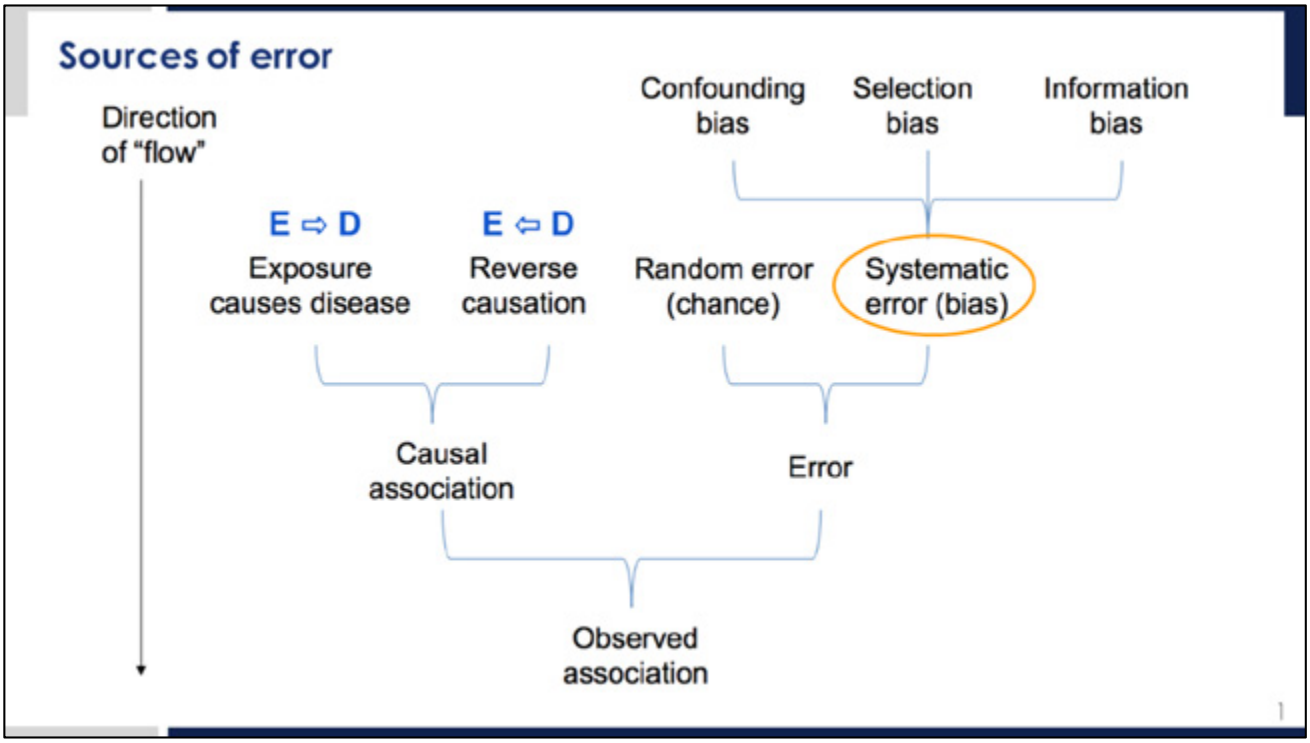


So to summarize, information bias can result from the way exposure, confounders, or outcomes are measured in a study. Information bias can occur in any type of epidemiologic study. Careful study design and detailed training and study protocols can help reduce information bias.

Misclassification occurs when information bias is present. Non-differential misclassification biases result in misclassification towards the null. But differential misclassification biases result in an unknown direction of misclassification.



JACK COLFORD: In this session, we're going to talk about a very important concept in epidemiology, which is generalizability.



When we do a study as epidemiologists, obviously we go out and we find some sample of the population to work with. We hope that whatever we find in that sample is representative of the broader truth or the broader world in which the sample has been derived from.

Now, there are many different ways for things to go wrong in the measurement of our exposure and disease and the relationship between exposure and disease in our sample. And this flowchart here, Sources of Error, is a reminder of kind of all the different things that can go wrong. We've discussed this several times in earlier videos, but just to review quickly, we hope that were correctly measuring exposure of causing disease, and we want to see what the relationship is-- whether exposure causes disease, and how strongly exposure causes disease, or whether exposure might prevent disease, like a vaccine, for example.

We hope that what we've detected in our measurement is a causal association, but we know that all we get to report is an observed association because we can never provide a perfect counter-factual. But what are all the different things that can go wrong in our measurement of exposure and disease? Well, one we've talked about before is the possibility of reverse causation-- that the disease itself might be increasing or changing exposure. So that's a reversal of what we think we are measuring when we see this association.

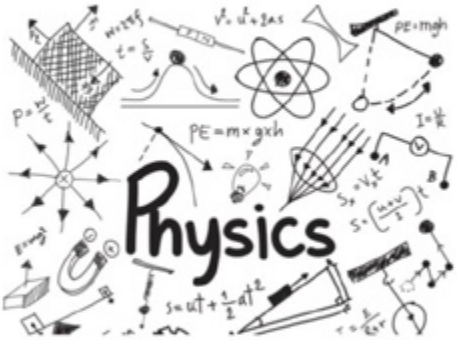
And then there are multiple types of bias that we've been talking about that can cause

error in our estimation of exposure and disease. And these include confounding bias, selection bias, and information bias, all of which together we treat as systematic error, or bias. So that's a source of error that can distort our observed association.

And then finally, there is random error or chance that, just because of randomness in nature, can lead to different estimates if we repeat our measurement several times. We try to account for that with confidence intervals and by doing multiple observations and by trying to use the best equipment and best measurement tools that we can find.

Generalizability in scientific disciplines

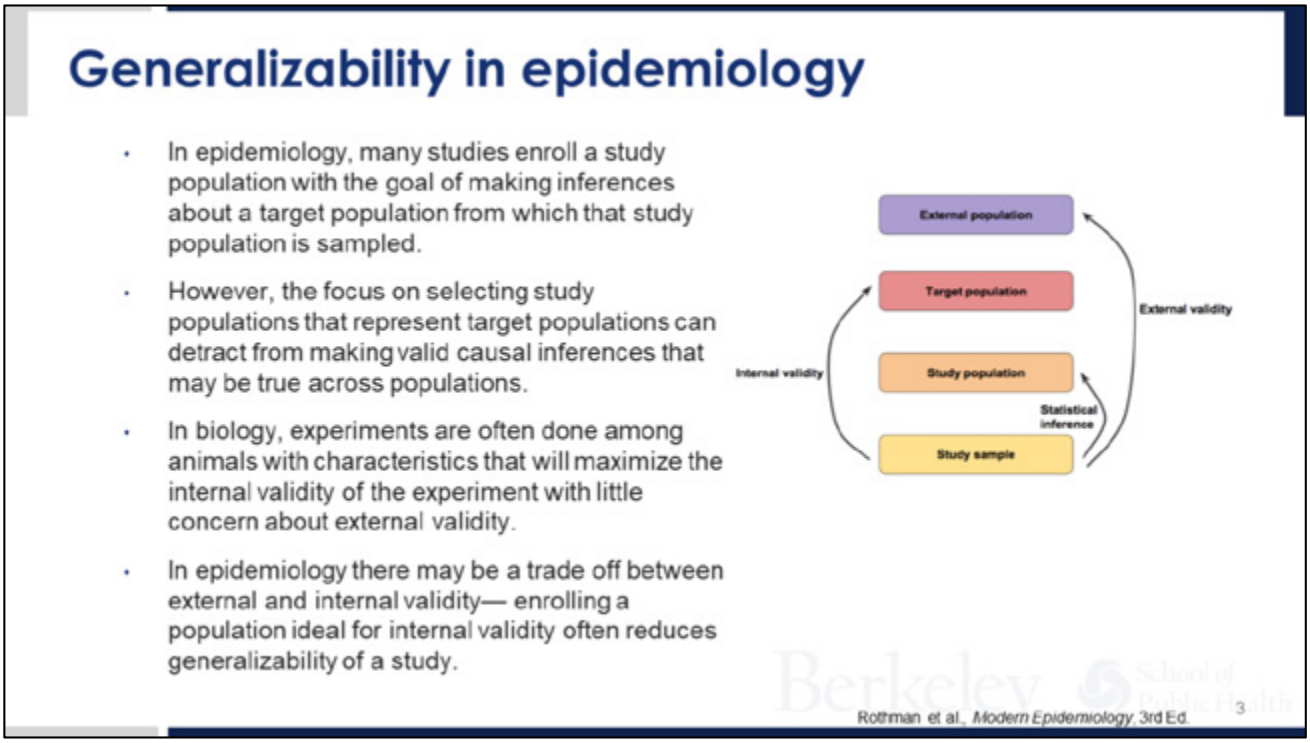
- In some fields, like physics, it's reasonable to assume that the laws of nature have universal applicability (ie, high generalizability)
- However, in biomedical science, we often restrict our inferences to specific populations.
- This stems from the assumption that biological effects can differ across populations



Rothman et al., *Modern Epidemiology*, 3rd Ed.

Now, what's generalizability in scientific disciplines? So in some fields like physics, it's reasonable to assume that the laws of nature have universal applicability, that is, high generalizability. If I measure something about the properties of water in the United States, then if I measure the properties of water in Europe, I should come up with the same measurements and ideas about what's in water and how it's composed, et cetera. It's a generalizable finding, a generalizable fact.

But in biomedical science, we often are forced to restrict our inferences or conclusions about relationships to specific populations, the ones we've studied. And then we're left with the question, how similar are the populations we've studied to other populations that we didn't study? Now, this, of course, arises from the assumption that biological effects can actually differ across populations. It's possible that the effect of smoking in men is different than the effect of smoking in women. And similarly, in different populations, different exposures could have different effects. That could be the actual truth.



So in epidemiology, we often enroll a study population with the goal of making inferences about a target population from which that study population is sampled. So this is a helpful schematic on the right to get to know very well. Starting at the bottom, we have our study sample, and we hope that it has been representative of a broader study population. However, the focus on selecting study populations that represent target populations can detract us from making valid causal inferences that may be true across populations.

And we talk about two different types of validity in these sorts of hierarchy of populations. So again, we have our study sample that's a small sampling of the broader population. And we use statistical tools to try to see that our sampling hasn't introduced statistical differences between the study sample and the study population. So we use statistical inference, statistical methods, to try to make sure that that is correct.

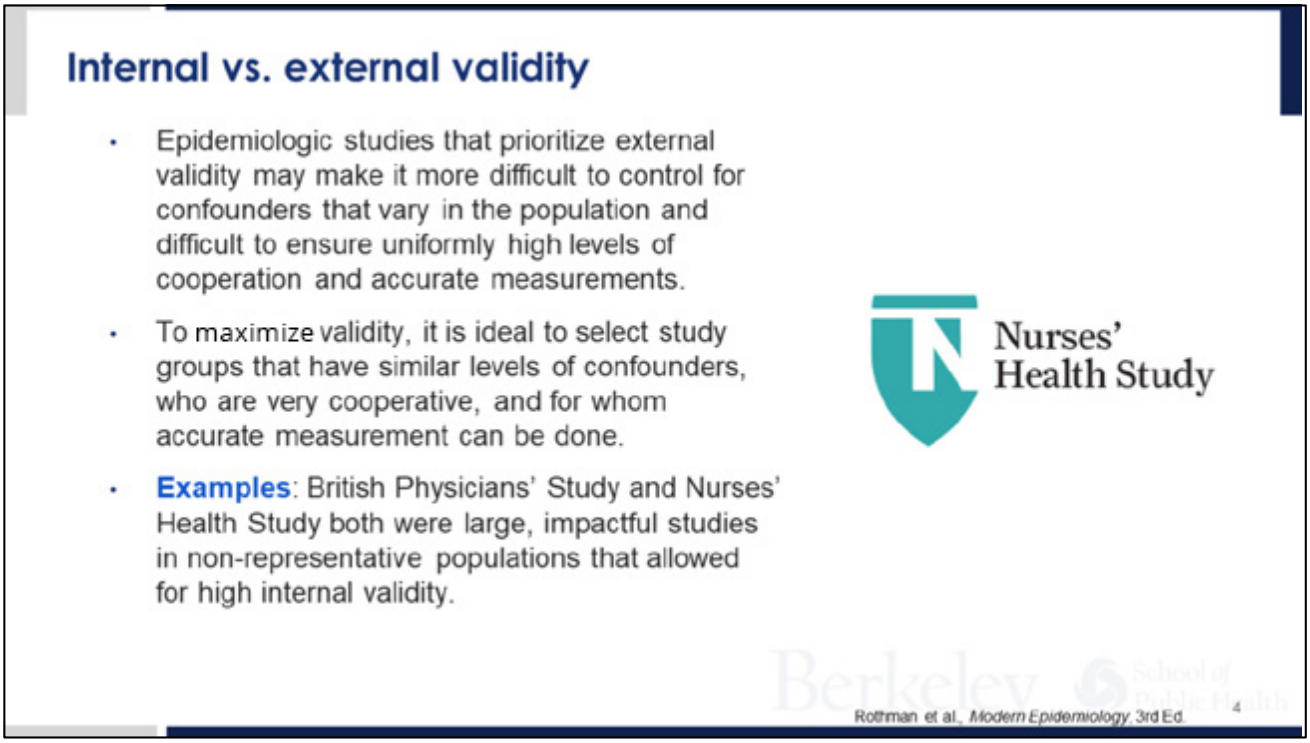
Now, if we've done things correctly and our study population is properly representative of our target population, then we believe we have what's called internal validity-- that is, that the study sample is representative of the target population to which we hope we are trying to extend our results.

And finally, if our study sample is a good representation of our study population, and our study population is a good representation of the target population, and the target population is, finally, a good representation of the external population, then we

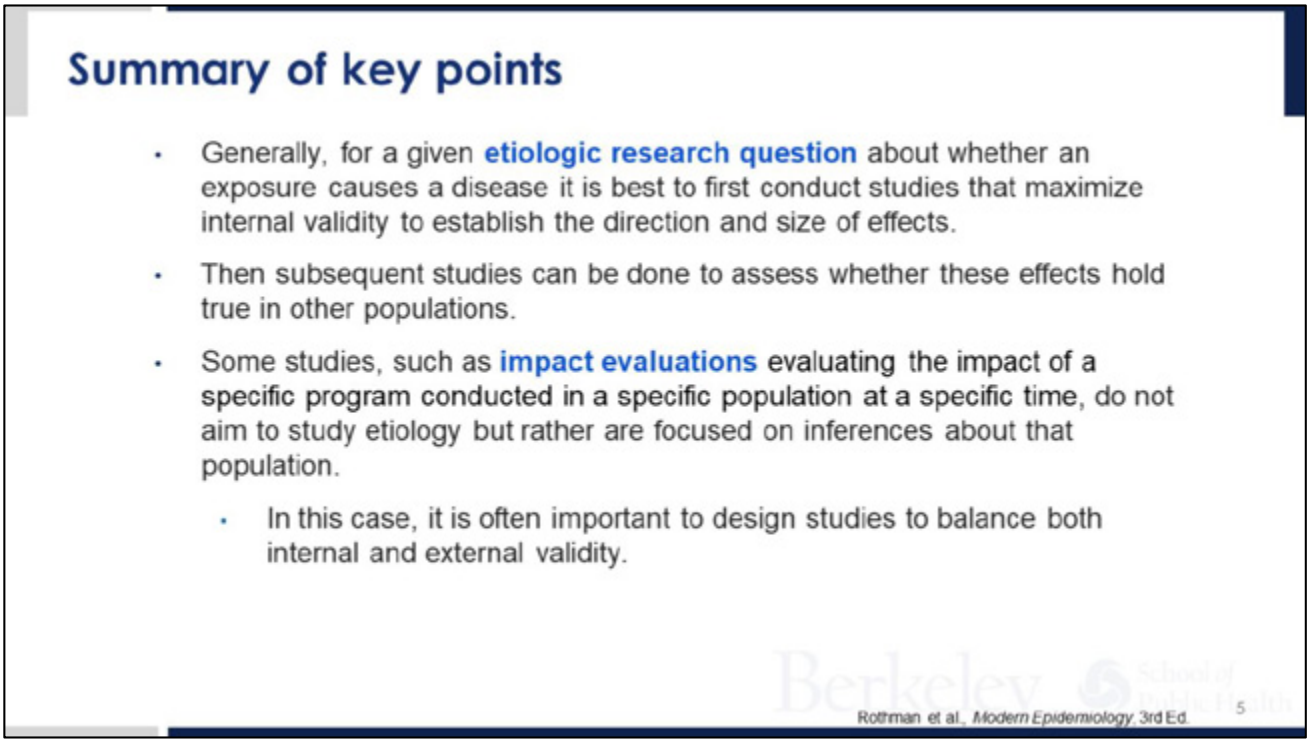
believe we have external validity.

So in biology, experiments are often done among the animals with characteristics that maximize the internal validity of the experiment. But they have little concern about external validity. So I may just be studying a small sample and all I care about is getting that sample analyzed correctly, and I haven't perhaps paid enough attention to whether that study population from which my small sample has come is more broadly representative of higher levels, the target population, the external population.

So in epidemiology, we often have trade-offs between external and internal validity. We may enroll a population ideal for internal validity, but in doing so, we may not have captured a population that is ideal for external validity. So a good example of this is when we do clinical trials, it's well-known that the people who are enrolled in clinical trials are often not a very good representation of the larger population in general. So we may enroll a population in a clinical trial, study them very correctly, do our study properly, et cetera, but we're left with a result that, while it's correct for the population we sampled and studied, it's not correct for these larger target and external populations.



The contrast between internal and external validity is an important one in epidemiology. And epi studies that prioritize external validity may make it more difficult to control for confounders that vary in the population, and make it difficult to ensure uniformly high levels of cooperation and accurate measurements. To maximize validity, it's ideal to select study groups that have similar levels of confounders, who are very cooperative, and for whom accurate measurement can be done. For example, the British Physicians' Study and Nurses' Health Study both were large, impactful studies in non-representative populations that allowed for high internal validity, but then, of course, raised the question, were these nurses and were these physicians properly representative of larger populations to which we hope to apply the findings and the results?



Generally, for a specific etiologic, causal kind of research question about whether an exposure causes a disease, it's best for us first to conduct studies that maximize internal validity to establish the direction and the size of effects. And then, in subsequent studies, we try to assess whether these effects hold true in other populations. So we repeat our work in other populations.

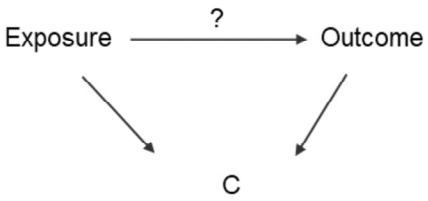
Now, some studies, such as a class of study called impact evaluations, evaluate the impact of a specific program conducted in a specific population at a specific time. Their goal is not to study etiology or causation, but rather focus on inferences about that larger population. And in that case, it's often important to design studies that balance both internal and external validity.

Lecture: Diagnosing Bias using DAGs

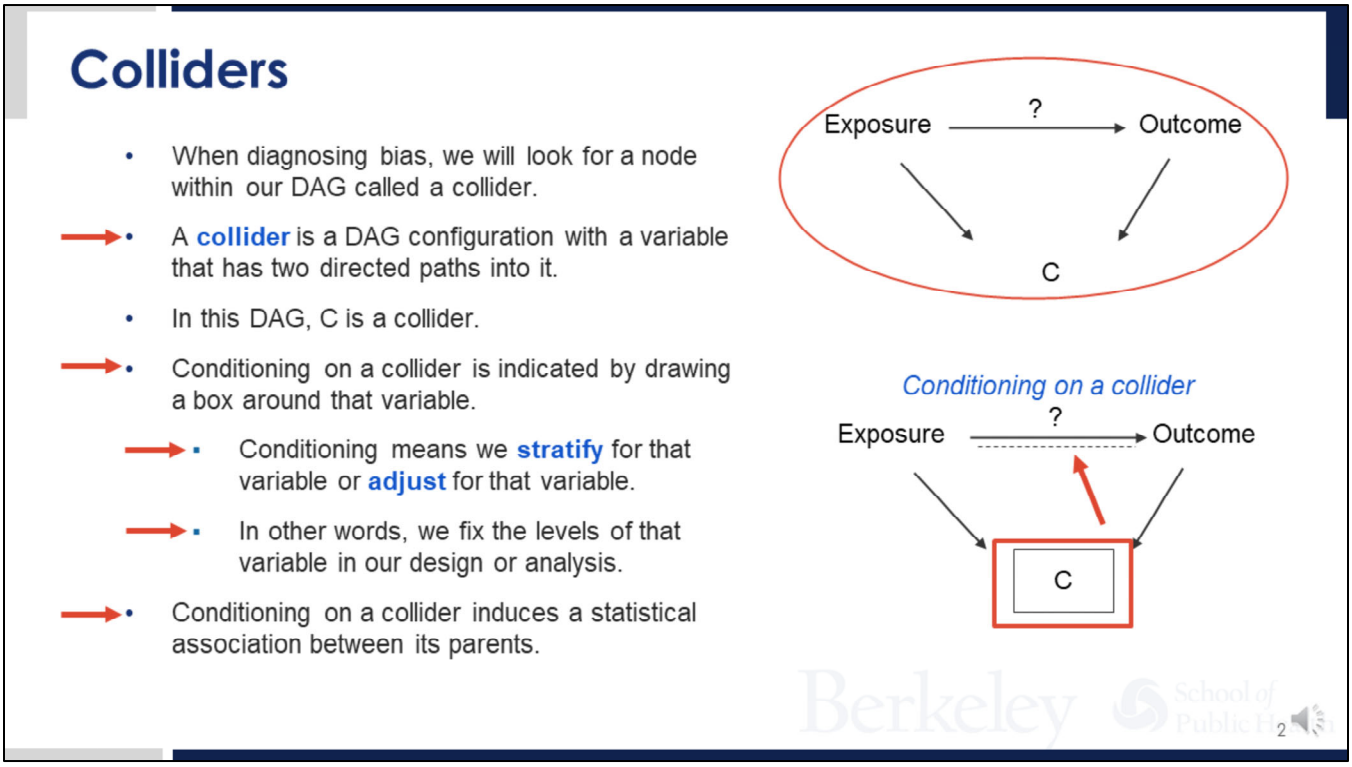
In this video, I'll talk about how to use Directed Acyclic Graphs, or DAGs, to diagnose bias in epidemiologic studies

Directed acyclic graphs (DAGs) and bias

- DAGs are a powerful tool in the modern epidemiology toolkit for diagnosing bias.
- They allow us to diagnose bias using our beliefs about the causal relationships between variables in our study.
- They can be used during the design or analysis phase to detect possible confounding.



*DAGs are a very powerful tool for diagnosing bias in our studies. When we make a DAG, we're able to encode our assumptions about how data is generated by different variables in our study and *what the causal relationships are between those different variables. *Then we can use a DAG that we designed particular to our study to assess bias or risk of bias in that study. This can be done during the design of a study. Or once the data is collected and analyzed, it can be done after that to assess if bias occurred. But of course, it's best to do this early on so that you can do everything you can to prevent further bias before all of the data is collected.



I'd like to start by introducing something called a collider. This is a particular structure or configuration within a DAG that we're going to be looking for in this video as we assess the presence of bias. *In this top right DAG here, we see an exposure that may lead to an outcome. And then we see that the exposure causes a variable or node, C, and the outcome also causes that node, C. C is a collider. *It's a DAG configuration where there is a variable that has two directed paths into it. In this video, we'll be talking about conditioning on a collider. *And when we condition on the collider, we indicate that by drawing a box a random variable. *In the bottom right DAG, we see a box around C.

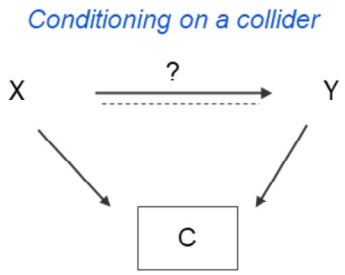
What we mean by conditioning? *Well, it's another way of saying that we stratify on that variable or that we adjust for that variable or restrict to that variable to a certain value of that variable. *In other words, we're fixing a level of that variable in our design or analysis. For example, if C was attending a medical clinic, and we only enrolled people who attended a medical clinic and didn't enroll people who did not attend a medical clinic, we would be conditioning on that collider. And that's because the study would only have people with one value of that variable for attending a medical clinic. Now, I won't go into great detail about why this is. *But what happens when we condition on a collider is we induce a statistical association between its parents. So in the bottom right DAG, the parents of C are exposure and outcome. *And I've been using this arrow moving from left to right with the question mark to indicate the causal effect that we're interested in. Well, when we condition on C, we induce an additional statistical association between exposure and outcome. And the reason I'm calling this a statistical association is that it's just a product of our study design or our study analysis or the way we collected the data. It's not really related to the true effect of the exposure on the outcome. And so that's why

it's denoted with a dashed line in addition to that solid arrow, which represents the fact that we're truly interested in. That dashed line represents bias. It's the influence of some other source, some other relationship in our dataset or in our study that causes our exposure outcome effect estimate to be distorted away from the truth.

Why does conditioning on a collider induce an association between its parents?

Example:

- C = X + Y
- If you know X = 3 you don't know the value of Y because X and Y are independent.
- If you know C = 10 and X = 3 then you know the value of Y.
- This is because X and Y are dependent **given that C = 10**.
- i.e., when we set C to a fixed value, knowing the value of X allows us to know the value of Y and vice versa.



Conditioning on a collider creates bias

3

So why does conditioning on a collider induce an association between its parents? Let's go for an example to build a little intuition around this. *So let's say C is equal to X plus Y. Now, it's important to keep in mind this is very unrealistic and epidemiology. So we would almost never have a formula like this for we know the relationship between X, Y, and C. But for the purpose of this little example, let's just say we know this formula is true.

*If you know that X is 3, and that's all you know, you don't know the value of Y. If all you have is C equals X plus Y and X equals 3, without knowing the value of C, you don't know the value of Y, and vise versa. And that's another way of saying that X and Y are independent.

*If you know C equals 10 and X equals 3, then you can calculate the value of Y, which is 7. And that's because, if you know that C is 10 and X is 3, X and Y are dependent on each other, given that C equals 10.

*This blue part here, given that C equals 10, this is what we're doing when we condition on a collider or really any other variable. What we're doing is we're fixing that variable to a single value.

*And when we fix it to a single value, that induces this dependency between its parents in the case of a collider. In other words, when we set C of a fixed value, knowing the value of X allows us to know the value of Y and vise versa. So in a real epidemiologic data set, again, we wouldn't know the formula C equals X plus Y. But there would be these statistical relationships. And if we didn't condition on C, X and Y could remain

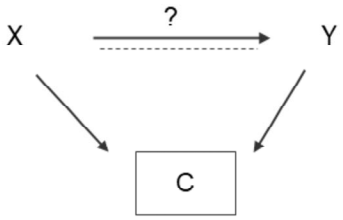
independent. Or they could be dependent on each other. X could cause Y, but that would have nothing to do with bias. That could be a true causal effect of X on Y. If we condition on that collider C, we're inducing dependency, statistically, between X and Y. And that induced dependency is termed bias in epidemiology. **So the key takeaway is that conditioning on a collider creates bias.

What we mean by "condition"

If C is hospitalization, conditioning on C could mean:

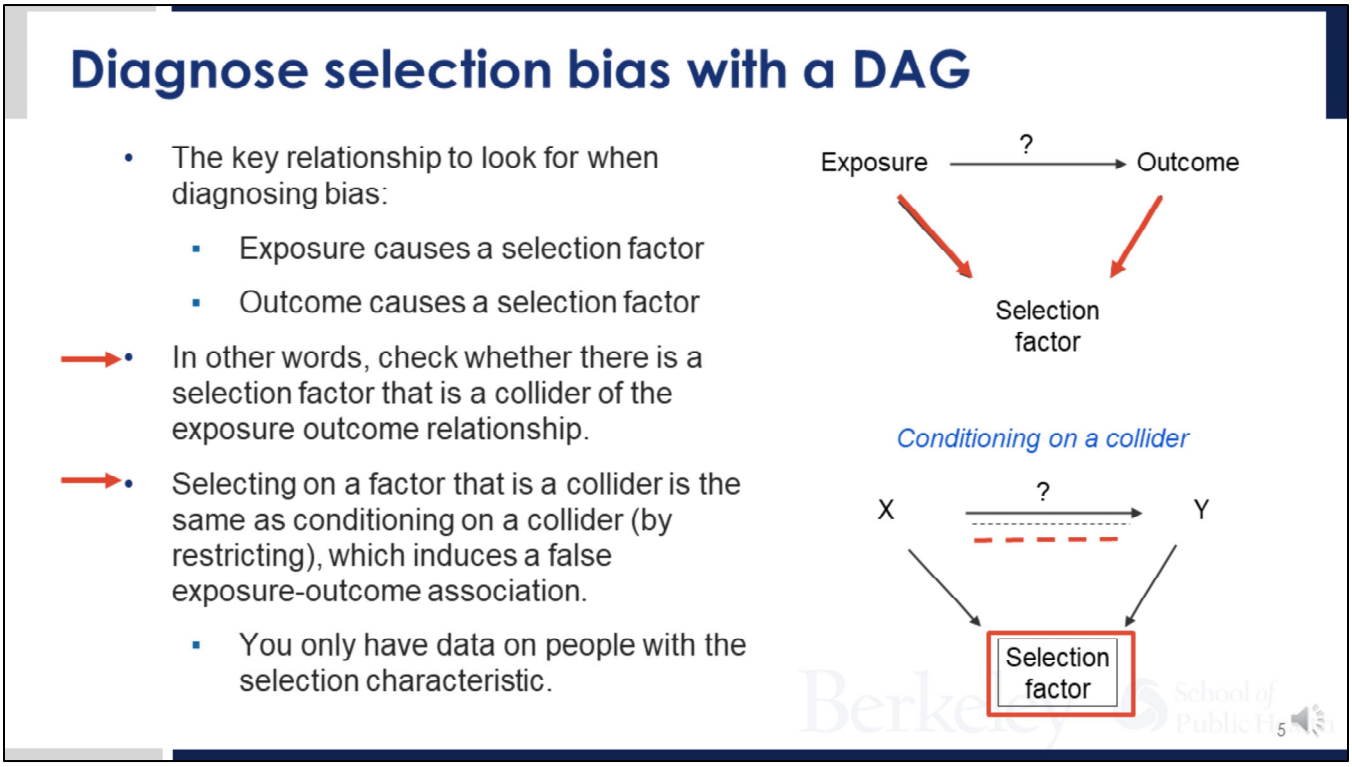
- **Restriction**: We restrict the value of C to a single number.
  - We restrict to only people who were hospitalized.
- **Stratification/ Adjustment**: We estimate the exposure-outcome association within each level of C.
  - We estimate the association separately among the hospitalized and among the not hospitalized.

Conditioning on a collider

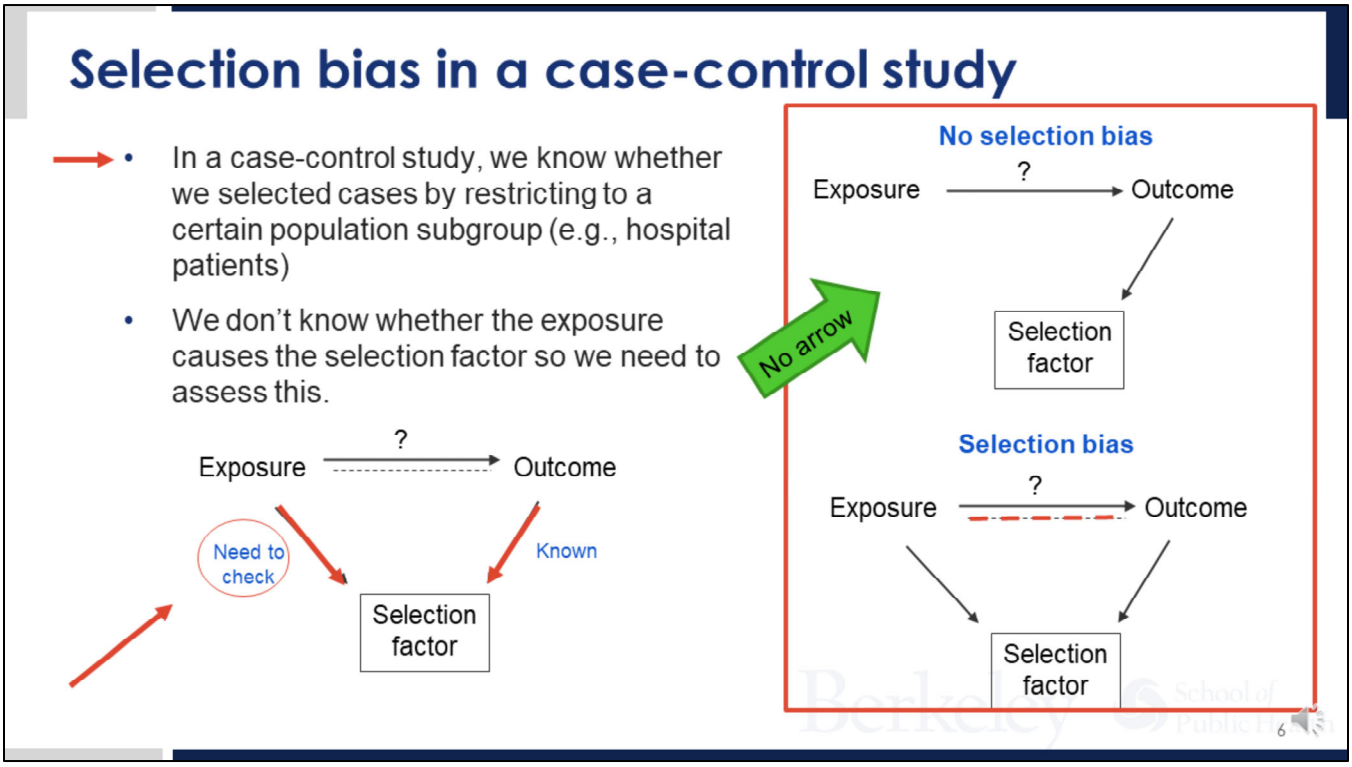


4

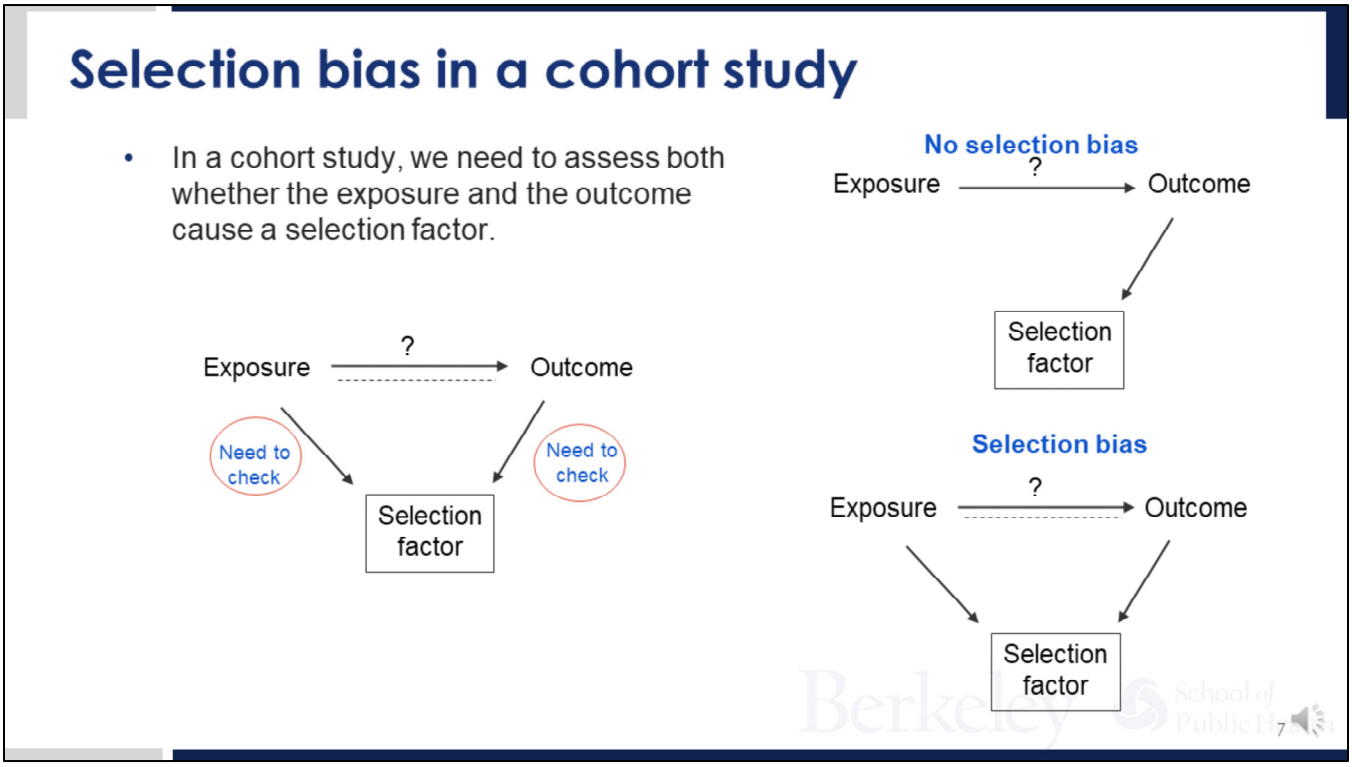
Here's another example. Let's say *C is hospitalization-- so whether someone was admitted to hospital or not. So what is conditioning on C mean? Well, it could mean *restriction. We restrict the value of C to a single number. We restrict people in our study to only those who were hospitalized. We don't enroll anyone who was not hospitalized. That's an example of restriction. And it's quite common, especially in case-control studies, for example. It could also mean *stratification or adjustment. So we might estimate our association between X and Y or the association between exposure and outcome within each level of C. So C may take on different levels-- hospitalized and not hospitalized, for example. If we estimate the association separately among those who are hospitalized, and separately among those who are not hospitalized, that is stratification, and it can also sometimes be referred to as adjustment. So that's also conditioning on C.



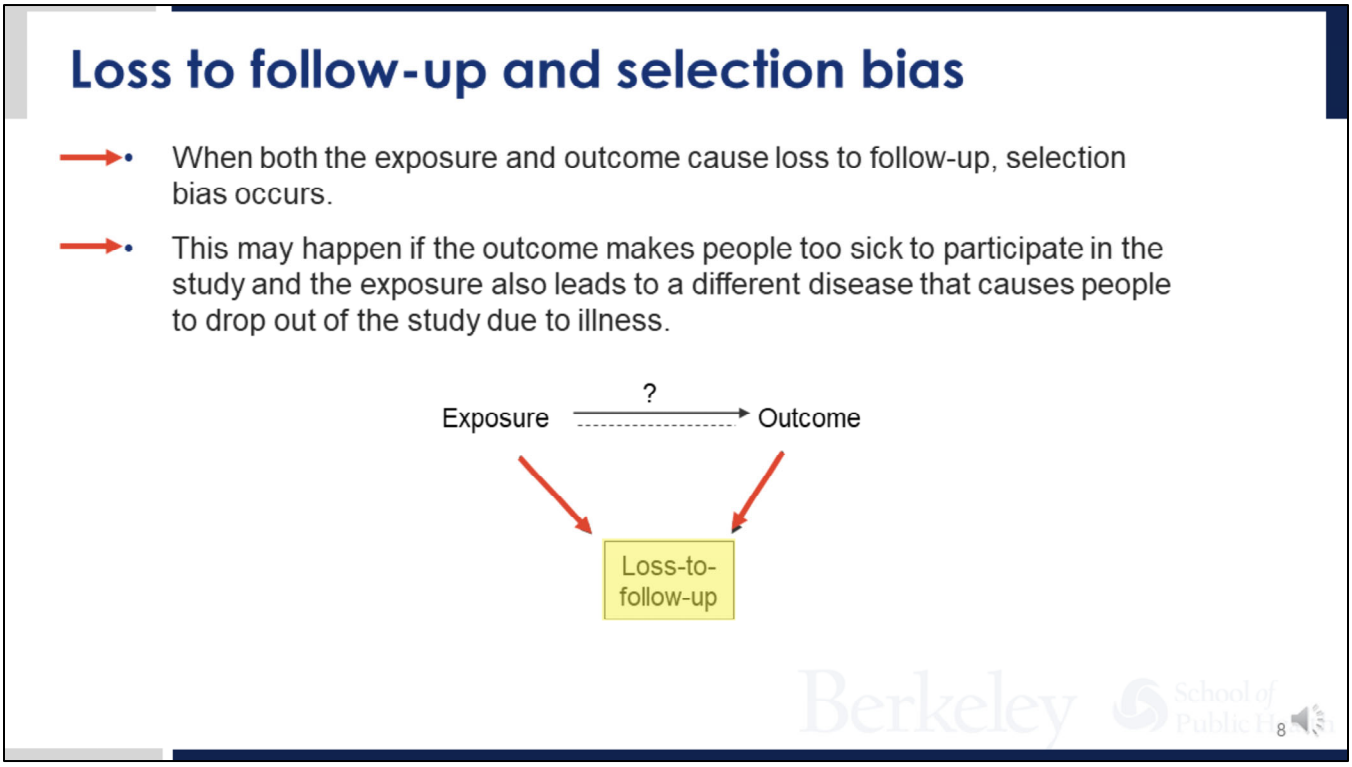
To diagnose selection bias with a DAG, we can then look for specific structures. So we want to look for whether *an exposure causes a selection factor and *the outcome causes the same selection factor in the DAG as shown on the upper right here. *So in other words, we're checking whether there's *a selection factor that is a collider of our exposure outcome relationship. And if that is the case, if we condition on that collider through restriction, stratification, adjustment, *we're going to induce a false exposure outcome association. *Now, in the case of selection bias, this is almost always through restriction, meaning, we're only selecting certain people into our study based on the value of a certain variable. This induces bias in our exposure-outcome association.



*In a case-control study, we know who our cases are, typically. So if we do a traditional, cumulative case-control study, we select our cases after they've all developed their disease of interest. And we may restrict to a certain population subgroup, like people who attended a certain hospital. *So if we look at this DAG here on the bottom left, we know that *the outcome causes *our selection factor if we are only enrolling people from a certain hospital. We don't necessarily know whether the exposure causes that selection factor. *If the exposure causes someone to go to the hospital, and we're only enrolling hospitalized patients, we likely have selection bias because we're conditioning on hospitalization through the design of our case-control study. So again, in a case control study, *we need to check whether the exposure causes the selection factor. And we usually know that the outcome does. We also need to assess whether we're conditioning on that selection factor. *And then, on the right, it shows you what scenarios look like when drawn as DAGs. If the exposure doesn't cause the selection factor, *there's no arrow from exposure to the selection factor. And even if we condition on selection factor, it doesn't induce bias, because it's not a collider. In the bottom right DAG, it's a classic example of selection bias. If the exposure and the outcome lead to the selection factor, which we condition on, *then that induces a bias between exposure and outcome.



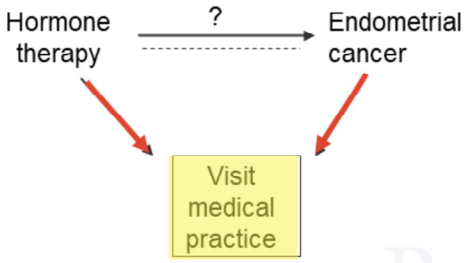
In a cohort study, we have to assess both arrows from exposure to selection factor and outcome to selection factor. In most cohort studies, we won't know for sure at the beginning of the study whether these arrows happen to be true. And then the DAGs on the right here are the exact same DAGs from the previous slide showing what it looks like if the arrow from exposure to selection factor is present or absent.



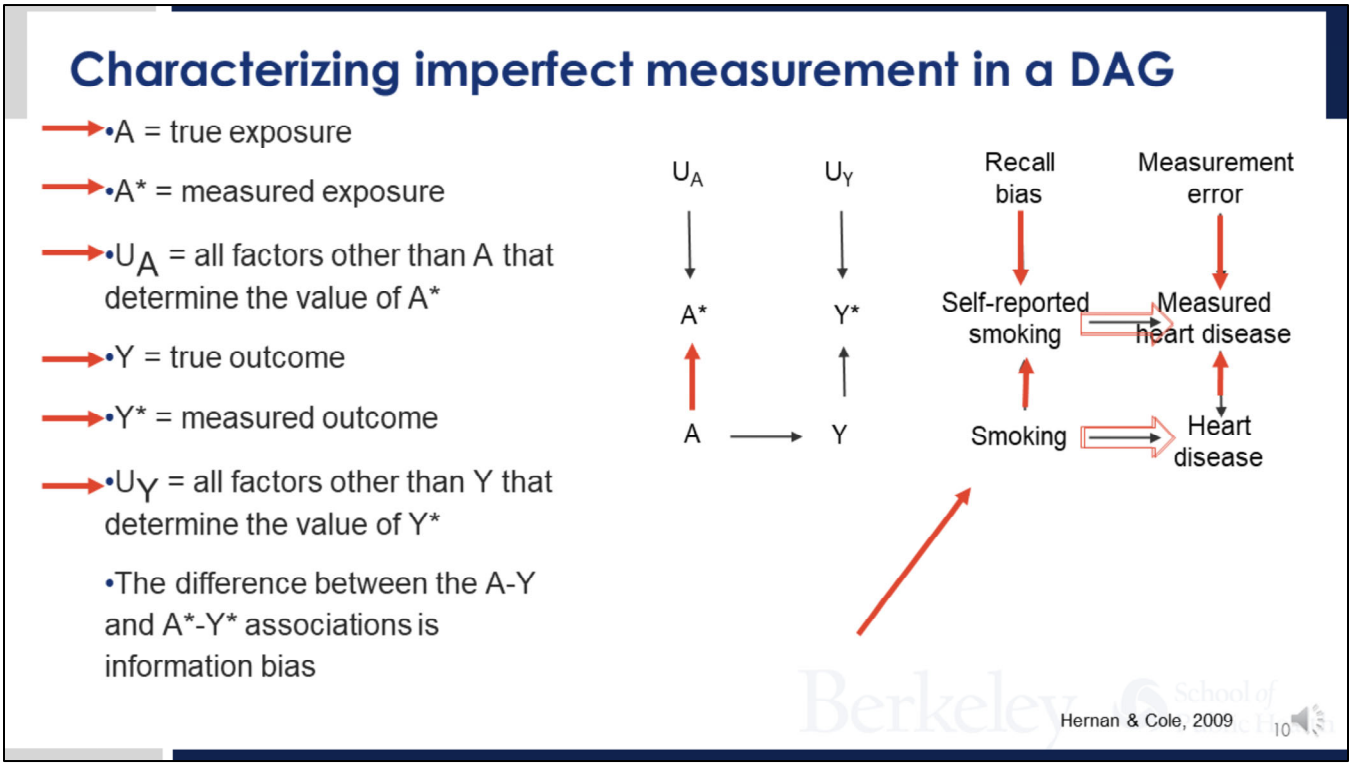
So we've been talking about enrollment into a study and selection bias. But selection bias can also take on a different form in loss to follow-up. *So when both the exposure and outcome cause loss to follow-up-- meaning people drop out of the study or people can't be contacted again for one reason or another-- we have differential loss to follow-up. *This could happen, for example, if the outcome makes people so sick that they just can't participate in the study anymore, and the exposure leads to a different disease that causes people to drop out of the study due to their severe illness. *So if the exposure causes loss to follow-up and *the outcome causes loss to follow up, we have a collider. *And then we're implicitly conditioning on that collider because we only have the people who weren't lost to follow-up in our dataset. So that's a collider that we're conditioning on. And that induces bias in our exposure-outcome relationship

Example

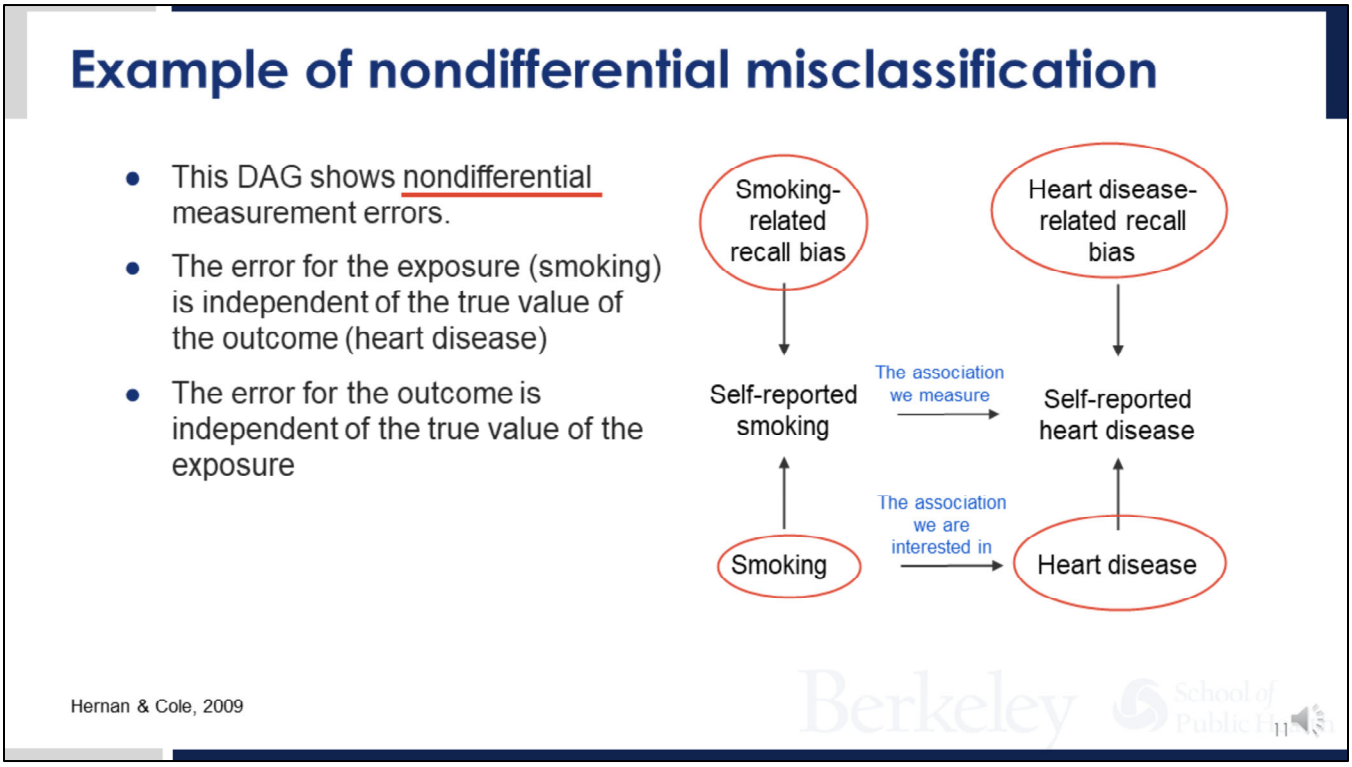
- A case-control study of postmenopausal hormone therapy and endometrial cancer
- Cases and controls enrolled from same medical practice
- Both women with endometrial cancer and those with postmenopausal hormone therapy are more likely to have bleeding and seek medical care.
- Visiting the medical practice is a collider that is conditioned on, so selection bias is present.



Let's go over an example. *A case control study focuses on whether or not post-menopausal hormone therapy leads to endometrial cancer. *And the study enrolls cases and controls from the same medical practice. *Let's say, in this study, both the women with endometrial cancer and those with post-menopausal hormone therapy are more likely to have bleeding as a result of the cancer and of the therapy. And as a result, they're both more likely to seek medical care. So that creates a collider structure because *hormone therapy leads to visiting medical practice and also *endometrial cancer leads to visiting a medical practice. When we condition on visiting a medical practice by only enrolling people who *visit a medical practice, we're introducing selection bias into our study



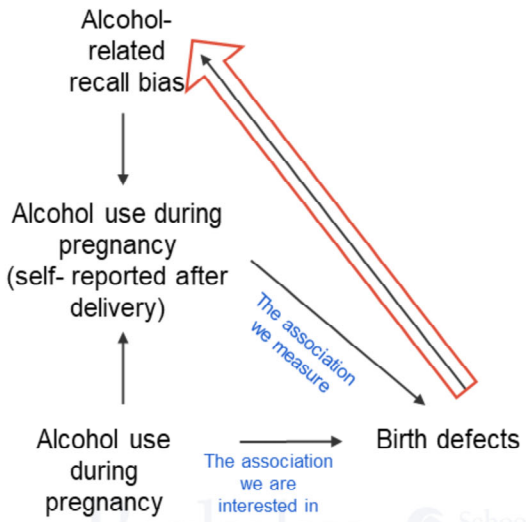
So far we've been focusing on selection bias, but now let's talk about how we can use DAGs to assess information bias. I'm only going to briefly cover this, just so that you've seen it before. This isn't something that we'll include an exams for this class or in problems sets for this class. But I just think it's helpful to have seen it before in case you encounter this in your future work. First, let me introduce the notation we're going to use. *So A can denote our true exposure. And that's our exposure that has no measurement error, no bias. *A star represents the measured exposure with any error or bias. *UA is all of the factors other than A that determine the value of A star. *Y indicates our true outcome. *Y star is our measured outcome including any error or bias. *And UY is all factors other than Y that determine the value of Y star. Now let's look at our DAGs. *A, the true exposure, causes the measured exposure. The exposure that we measure—and *let's say, in our example here, that we're interested in smoking. *When we measure self-reported smoking, it's affected by true smoking status. *But the self-reported smoking is also affected by recall bias and other forms of bias and error. Similarly, for heart disease, we can measure heart disease. And the measurement we get of whether someone has it or doesn't have it *depends on their true heart disease status, which might be unknown, and *any measurement error that affects the measured status of heart disease or not heart disease. *And then, we're interested in *the relationship between true smoking and true heart disease. But in our study, perhaps we only have *self-reported smoking and measured heart disease status.* This DAG incorporates measurement error and information bias into our causal relationships of interest. And it helps us clearly see that we're not estimating the association that we really care about. *This middle here, self-reported smoking to measure heart disease, is the estimate of the association that we'll get in our study. *But what we actually want is the association from the bottom row. And that disconnect is information bias.



Here's that same example again. So we can say that this is *nondifferential misclassification because the error for the exposure, smoking-- *and that's indicated by smoking-related recall bias-- that error is independent of the true value of the outcome, *heart disease. There is no arrow from smoking-related recall bias to heart disease. Similarly, the error for the outcome-- *heart disease-related recall bias, is independent of the true value of the exposure-- *smoking. So there's no arrow from smoking to heart disease-related recall bias. If those arrows were there, we would call it differential misclassification. Instead, it's nondifferential misclassification.* Now, you may be thinking to yourself, I see some colliders in this DAG. Is this something I need to assess? While self-reported smoking is a collider of the smoking and smoking-related recall bias relationship, and there's also a collider for self reported heart disease, but those colliders are not between smoking and heart disease. So because they are not colliders of the exposure-outcome relationship, this is a different DAG structure than the selection bias DAG structure. Similarly, we're not conditioning on either of these colliders, because we'll have values for self-reported smokers and self-reported nonsmokers. There's no restriction in this example or stratification necessarily. So this doesn't induce any additional other form of bias. But rather, the structure of the DAG itself illustrates to us that there is some nondifferential misclassification.

Example of differential misclassification

- This DAG shows differential measurement errors.
- The true value of the outcome affects the measurement of the exposure.
- Once a child is born, their birth defects status is likely to influence recall of alcohol use during pregnancy.



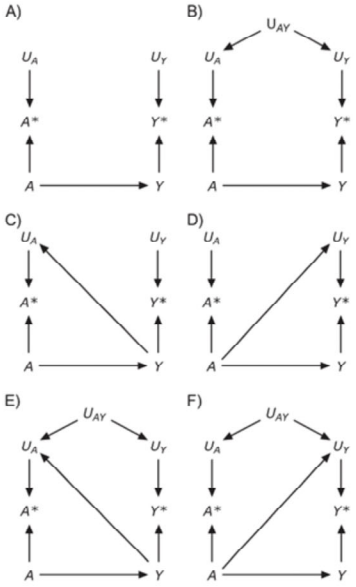
Hernan & Cole, 2009

12

Here's an example with differential misclassification. So this study is looking at alcohol use during pregnancy and birth defects. And let's say that we measure alcohol use during pregnancy through self-report after the child is born. So it's very likely that, once a child is born, their status of whether or not they have birth defects will affect the mother's recall of their alcohol use during pregnancy. So the outcome effects the exposure's error. That's the arrow from birth defects to alcohol-related recall bias. And that's why this is an example of differential misclassification because the classification of our exposure depends on the outcome status.

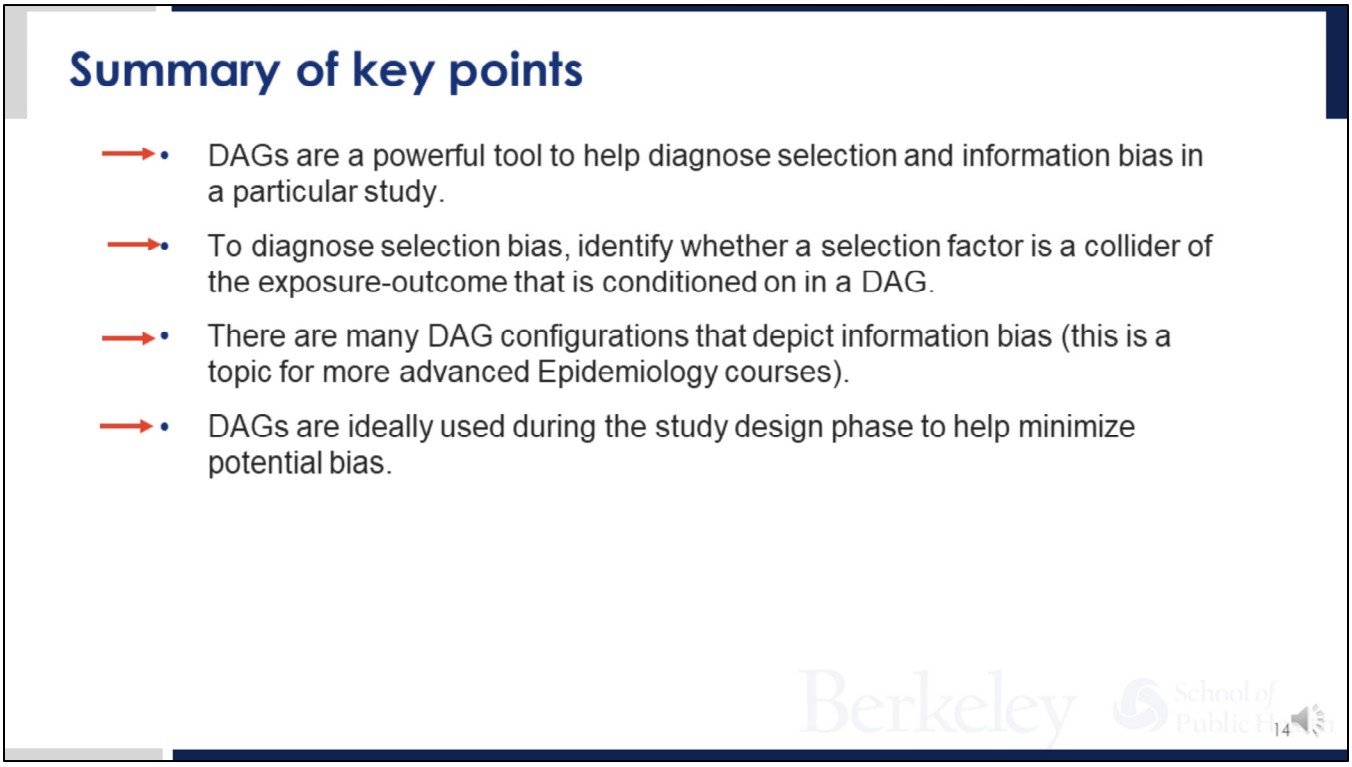
Diagnosing information bias using DAGs

Many possible DAG configurations exist that could reflect different forms of information bias.



Hernan & Cole, 2009 13

So I've just shown you a few examples. But there is a lot of different DAG configurations that could reflect different forms of information bias. So this is more of an advanced topic, but I think it's helpful to just have seen it before, so that if you're ever doing a study yourself, and you are worried about information bias, you can think about it, and try to make your own DAG to assess whether any information bias might be present.



To summarize, DAGs are a really powerful tool to help us diagnose selection and information bias in our own studies. To diagnose selection bias, we look for a collider that is our selection factor. And this is a collider of the exposure-outcome relationship. When we condition on that collider, we introduce bias. There are many different DAG configurations that can depict information bias. And they do so by including nodes for measurement error and for measured exposure and outcome status. And it's really ideal to try to use DAGs at the study design phase rather than after the study has been conducted to help you identify potential sources of bias and do your best to minimize these biases in the study

Epidemiology series

Bias and causal associations in observational research

David A Grimes, Kenneth F Schulz

Readers of medical literature need to consider two types of validity, internal and external. Internal validity means that the study measured what it set out to; external validity is the ability to generalise from the study to the reader's patients. With respect to internal validity, selection bias, information bias, and confounding are present to some degree in all observational research. Selection bias stems from an absence of comparability between groups being studied. Information bias results from incorrect determination of exposure, outcome, or both. The effect of information bias depends on its type. If information is gathered differently for one group than for another, bias results. By contrast, non-differential misclassification tends to obscure real differences. Confounding is a mixing or blurring of effects: a researcher attempts to relate an exposure to an outcome but actually measures the effect of a third factor (the confounding variable). Confounding can be controlled in several ways: restriction, matching, stratification, and more sophisticated multivariate techniques. If a reader cannot explain away study results on the basis of selection, information, or confounding bias, then chance might be another explanation. Chance should be examined last, however, since these biases can account for highly significant, though bogus results. Differentiation between spurious, indirect, and causal associations can be difficult. Criteria such as temporal sequence, strength and consistency of an association, and evidence of a dose-response effect lend support to a causal link.

Clinicians face two important questions as they read medical research: is the report believable, and, if so, is it relevant to my practice? Uncritical acceptance of published research has led to serious errors and squandered resources.[1] Here, we will frame these two questions in terms of study validity, describe a simple checklist for readers, and offer some criteria by which to judge reported associations.

Internal and external validity

Analogous to a laboratory test, a study should have internal validity—ie, the ability to measure what it sets out to measure.[2] The inference from participants in a study should be accurate. In other words, a research study should avoid bias or systematic error.[3] Internal validity is the sine qua non of clinical research; extrapolation of invalid results to the broader population is not only worthless but potentially dangerous.

A second important concern is external validity; can results from study participants be extrapolated to the reader's patients? Since a total enumeration or census approach to medical research is usually impossible, the customary tactic is to choose a sample, study it, and, hopefully, extrapolate the result to one's practice. Gauging external validity is necessarily more subjective than is assessment of internal validity.

Internal and external validity entail important trade-offs. For example, randomised controlled trials are more likely than observational studies to be free of bias,[4] but, because they usually enrol selected participants, external validity can suffer. This problem of unsuitable participants is also termed distorted assembly.[5] Participants in randomised controlled trials tend to be different (including being healthier[6–8]) from those who choose not to take part, a function of the restricted entry criteria. The filtering process for admission to randomised trials might, therefore, result in "a type of hothouse flower, which cannot bloom or be successfully removed beyond its special greenery."[5]

Lancet 2002; **359:** 248–52

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (D A Grimes MD, K F Schulz PhD)

Correspondence to: Dr David A Grimes
(e-mail: dgrimes@fhi.org)

Bias

Bias undermines the internal validity of research. Unlike the conventional meaning of bias—ie, prejudice—bias in research denotes deviation from the truth. All observational studies (and, regrettably, many badly done randomised controlled trials)[9,10] have built-in bias; the challenge for investigators, editors, and readers is to ferret these out and judge how they might have affected results. A simple checklist, such as that shown in panel 1, can be helpful.[11–14]

Several taxonomies exist for classification of biases in clinical research. Sackett's landmark compilation,[15] for example, included 35 different biases. By contrast Feinstein[5] consolidated biases into four categories that arise sequentially during research: susceptibility, performance, detection, and transfer. Susceptibility bias refers to differences in baseline characteristics, performance bias to different proficiencies of treatment, detection bias to different measurement of outcomes, and transfer bias to differential losses to follow-up. Another approach,[3,11,16,17] which is often used, is to group all biases into three general categories: selection, information, and confounding. The leitmotif for all three is "different".[17] Something "different" distorts the planned comparison.

Selection bias

Are the groups similar in all important respects?

Selection bias stems from an absence of comparability between groups being studied. For example, in a cohort study, the exposed and unexposed groups differ in some important respect aside from the exposure. Membership bias is a type of selection bias: people who choose to be members of a group—eg, joggers—might differ in important respects from others. For instance, both cohort and case-control studies initially suggested that jogging after myocardial infarction prevented repeat

Panel 1: **What to look for in observational studies**

Is selection bias present?

In a cohort study, are participants in the exposed and unexposed groups similar in all important respects except for the exposure?

In a case-control study, are cases and controls similar in all important respects except for the disease in question?

Is information bias present?

In a cohort study, is information about outcome obtained in the same way for those exposed and unexposed?

In a case-control study, is information about exposure gathered in the same way for cases and controls?

Is confounding present?

Could the results be accounted for by the presence of a factor—eg, age, smoking, sexual behaviour, diet—associated with both the exposure and the outcome but not directly involved in the causal pathway?

If the results cannot be explained by these three biases, could they be the result of chance?

What are the relative risk or odds ratio and 95% CI?[11,12]

Is the difference statistically significant, and, if not, did the study have adequate power to find a clinically important difference?[13,14]

If the results still cannot be explained away, then (and only then) might the findings be real and worthy of note.

infarction. However, a randomised controlled trial failed to confirm this benefit.[15] Those who chose to exercise might have differed in other important ways from those who did not exercise, such as diet, smoking, and presence of angina.

In case-control studies, selection bias implies that cases and controls differ importantly aside from the disease in question. Two types of selection bias have earned eponyms: Berkson and Neyman bias. Also known as an admission-rate bias, Berkson bias (or paradox) results from differential rates of hospital admission for cases and controls. Berkson initially thought that this phenomenon was due to presence of a simultaneous disease.[5] Alternatively, knowledge of the exposure of interest might lead to an increased rate of admission to hospital. For example, doctors who care for women with salpingitis were more likely to recommend hospital admission for those using an intrauterine device (IUD) than for those using a hormonal method of contraception.[18,19] In a hospital-based case-control study, this would stack the deck (or gynaecology ward) with a high proportion of IUD-exposed cases, spuriously increasing the odds ratio.

Neyman bias is an incidence-prevalence bias. It arises when a gap in time occurs between exposure and selection of study participants. This bias crops up in studies of diseases that are quickly fatal, transient, or subclinical. Neyman bias creates a case group not representative of cases in the community. For example, a hospital-based case-control study of myocardial infarction and snow shovelling (the exposure of interest) would miss individuals who died in their driveways and thus never reached a hospital; this eventuality might greatly lower the odds ratio of infarction associated with this strenuous activity.

Other types of selection bias include unmasking (detection signal) and non-respondent bias. An exposure might lead to a search for an outcome, as well as the outcome itself. For example, oestrogen replacement therapy might cause symptomless endometrial cancer patients to bleed, resulting in initiation of diagnostic tests.[20] In this instance, the exposure unmasked the subclinical cancer, leading to a spurious increase in the odds ratio. In observational studies, non-respondents are different from respondents. Cigarette smokers are a case in point: smokers are less likely to return questionnaires than are non-smokers or pipe and cigar smokers.[21]

Information bias

Has information been gathered in the same way?

Information bias, also known as observation, classification, or measurement bias, results from incorrect determination of exposure or outcome, or both. In a cohort study or randomised controlled trial, information about outcomes should be obtained the same way for those exposed and unexposed. In a case-control study, information about exposure should be gathered in the same way for cases and controls.

Information bias can arise in many ways. Some use the term ascertainment to describe gathering information in different ways. For example, an investigator might gather information about an exposure at bedside for a case but by telephone from a community control. Diagnostic suspicion bias implies that knowledge of a putative cause of disease might launch a more intensive search for the disease among those exposed, for example, preferentially searching for infection by HIV-1 in intravenous drug users. Conversely, the presence of a disease might prompt a search for the putative exposure of interest. Another type of bias is family history bias, in which medical information flows differently to affected and unaffected family members, as has been shown for rheumatoid arthritis.[22] To minimise information bias, detail about exposures in case-control studies should be gathered by people who are unaware of whether the respondent is a case or a control. Similarly, in a cohort study with subjective outcomes, the observer should be unaware of the exposure status of each participant.

In case-control studies that rely on memory of remote exposures, recall bias is pervasive. Cases tend to search their memories to identify what might have caused their disease; healthy controls have no such motivation. Thus, better recall among cases is common. For example, the putative association between induced abortion and subsequent development of breast cancer has emerged as a hot medical and political issue. Many case-control studies have reported an increase in cancer risk after abortion.[23] However, when investigators compared histories of prior abortions, obtained by personal interview, against centralised medical records, they documented systematic underreporting of abortions among controls (but not among cases) that accounted for a spurious association.[24] In Swedish and Danish cohort studies,[25,26] free from recall bias, induced abortion has had either a protective effect or no effect on risk of breast cancer.

Is the information bias random or in one direction?

The effect of information bias depends on its type. If information is gathered differentially for one group than for another, then bias results, raising or lowering the relative risk or odds ratio dependent on the direction of the bias. By contrast, non-differential misclassification—ie, noise in the system—tends to obscure real differences. For example, an ambiguous questionnaire might lead to errors in data collection among cases and controls, shifting the odds ratio toward unity, meaning no association.

Confounding

Is an extraneous factor blurring the effect?

Confounding is a mixing or blurring of effects. A researcher attempts to relate an exposure to an outcome, but actually measures the effect of a third factor, termed a confounding variable. A confounding variable is associated with the exposure and it affects the outcome, but it is not an intermediate link in the chain of causation between exposure and outcome.[27,28] More simply, confounding is a methodological fly in the ointment. Confounding is often easier to understand from examples than from definitions.

Oral contraceptives and myocardial infarction, and smoking

Early studies of the safety of oral contraceptives reported a pronounced increased risk of myocardial infarction. This association later proved to be spurious, because of the high proportion of cigarette smokers among users of birth control pills.[29–31] Here, cigarette smoking confounded the relation between oral contraceptives and infarction. Women who chose to use birth control pills also chose, in large numbers, to smoke cigarettes, and cigarettes, in turn, increased the risk of myocardial infarction. Although investigators thought they were measuring an effect of birth control pills, they were in fact measuring the hidden effect of smoking among pill users.

IUD insertion and salpingitis, and exposure to sexually transmitted disease

Results of a large case-control study of IUDs indicated a significant increase in salpingitis soon after insertion.[32] However, among married or cohabiting women with only one reported sex partner in the past 6 months, no significant increase in risk was evident.[33] In the study, exposure to sexually transmitted diseases apparently confounded the association. Even among women at low risk of salpingitis, frequent coitus might increase risk of infection,[34] and few studies have controlled for this variable.

Oral contraceptives and cervical cancer, and smoking

Reported associations between oral contraceptives and squamous cervical cancer[35] might be due to unsuspected confounding by cigarette smoking and human papillomavirus infection.[36] Control of confounding is inevitably limited by our meagre understanding of human biology; unsuspected confounding factors evade control in observational studies.[37]

Control for confounding

When selection bias or information bias exist in a study, irreparable damage results. Internal validity is doomed. By contrast, when confounding is present, this bias can be corrected, provided that confounding was anticipated and the requisite information gathered. Confounding can be controlled for before or after a study is done. The purpose of these approaches is to achieve homogeneity between study groups.

Restriction

The simplest approach is restriction (also called exclusion or specification).[28] For example, if cigarette smoking is suspected to be a confounding factor, a study can enrol only non-smokers. Although this tactic avoids confounding, it also hinders recruitment (and thus power) and precludes extrapolation to smokers. Restriction might increase the internal validity of a study at the cost of poorer external validity.

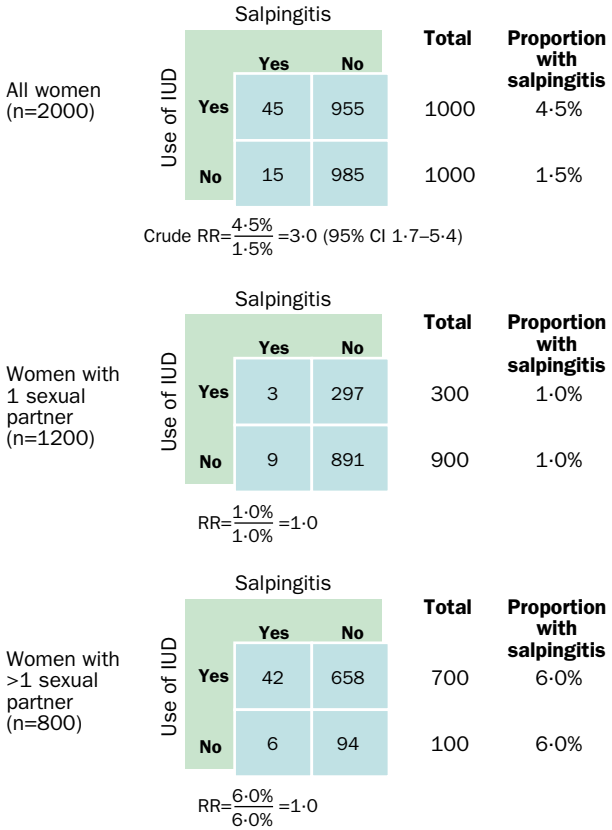
Matching

Another way to control for confounding is pairwise matching. In a case-control study in which smoking is deemed a confounding factor, cases and controls can be matched by smoking status. For each case who smokes, a control who smokes is found. This approach, although often used by investigators, has two drawbacks. If matching is done on several potential confounding factors, the recruitment process can be cumbersome, and, by definition, one cannot examine the effect of a matched variable.[28]

Stratification

Investigators can also control for confounding after a study has been completed. One approach is stratification. Stratification can be considered a form of post hoc restriction, done during the analysis rather than during the accrual phase of a study. For example, results can be stratified by levels of the confounding factor. In the smoking example, results are calculated separately for smokers and non-smokers to see if the same effect arises independent of smoking. The Mantel-Haenszel procedure[38] combines the various strata into a summary statistic that describes the effect. The strata are weighted inversely to their variance—ie, strata with larger numbers count more than those with smaller numbers. If the Mantel-Haenszel adjusted effect differs substantially from the crude effect, then confounding is deemed present. In this instance, the adjusted estimate of effect is considered the better estimate to use.

Confounding is not always intuitive, as shown by the fictitious example in the figure. In this hypothetical



Example of confounding in a hypothetical cohort study of intrauterine device use and salpingitis

When the crude relative risk is controlled for the confounding effect of number of sexual partners, the raised risk disappears.

cohort of 2000 women, use of an IUD was strongly related to development of salpingitis (relative risk 3·0; 95% CI 1·7–5·4). However, the number of sexual partners was related to women's choice of contraception and to their risk of upper-genital-tract infection. Here, a disproportionate number of women with more than one sexual partner chose to use an IUD (700 *vs* 300 women with only one partner). The number of partners was also related to the risk of infection (6% among those with >1 partner *vs* 1% among those with only one partner). In each stratum by number of partners, the relative risk is 1·0, indicating no association between the IUD and salpingitis. The Mantel-Haenszel weighted relative risk, which controls for this confounding effect, is 1·0 (95% CI 0·5–2·0). In this fictitious example, the apparent three-fold increase in risk associated with IUD use was all due to confounding bias.

Multivariate techniques

In multivariate techniques, mathematical modelling examines the potential effect of one variable while simultaneously controlling for the effect of many other factors. A major advantage of these approaches is that they can control for more factors that can stratification. For example, an investigator might use multivariate logistic regression to study the effect of oral contraceptives on ovarian cancer risk. In this way, they could simultaneously control for age, race, family history, parity, &c. Another example would be use of a proportional hazards regression analysis for time to death; this method could control simultaneously for age, blood pressure, smoking history, serum lipids, and other risk factors.[39] Disadvantages of multivariate approaches, for some researchers, include greater difficulty in understanding the results, and loss of hands-on feel for the data.[28]

Chance

If a reader cannot explain results on the basis of selection, information, or confounding bias, then chance might be another explanation. The reason for examination of bias before chance is that biases can easily cause highly significant (though bogus) results. Regrettably, many readers use the p value as the arbiter of validity, without considering these other, more important, factors.

The venerable p value measures chance. It advises the reader of the likelihood of a false-positive conclusion: a difference was seen in the study, although it does not exist in the broader population (type I error). Many clinicians are surprised to learn, however, that the p value of 0·05 as a threshold has no basis in medicine. Rather, it stems from agricultural and industrial experiments early in the 20th century.[40,41] Should a study not achieve significance at this level, one needs to see if the study had adequate power to find a clinically important difference. Many "negative" studies simply have too few participants to do the job.[13,14] Better yet, investigators should present measures of association with confidence intervals[41] in preference to hypothesis tests.

Judgment of associations

Bogus, indirect, or real?

When statistical associations emerge from clinical research, the next step is to judge what type of association exists. Statistical associations do not necessarily imply causal associations.[17] Although several classifications are available,[28] a simple approach includes just three types: spurious, indirect, and causal. Spurious associations are the result of selection bias, information bias, and chance. By contrast, indirect associations (which stem from confounding) are real but not causal.

Judgment of cause-effect relations can be tough. Few rules apply, though criteria first suggested by Hill have received the most attention (panel 2).[17,42,43] The only iron-clad criterion is temporality: the cause must antedate the effect. However, in many studies, especially with chronic diseases, answering this chicken-egg question can be daunting. Strong associations argue for causation. Whereas weak associations in observational studies can easily be due to bias, large amounts of bias would be necessary to produce strong associations. (This large bias is evident in reports that link IUD use with salpingitis.) Some suggest that relative risks more than 3 in cohort studies, or odds ratios greater than 4 in case-control studies, provide strong support for causation.[44] Consistent observation of an association in different populations and with different study designs also lends support to a real effect. For example, results of studies done around the world have consistently shown that oral contraceptives protect against ovarian cancer; a causal relation can, therefore, be argued. Evidence of a biological gradient supports a causal association too. For instance, protection against ovarian cancer is directly related to duration of use of oral contraceptives.[45] The risk of death from lung cancer is linearly related to years of cigarette smoking. In both of these examples, increasing exposure is associated with an increasing biological effect.

Other criteria of Hill's are less useful. Specificity is a weak criterion. With a few exceptions, such as the rabies virus, few exposures lead to only one outcome. Should an association be highly specific, this provides support for causality. However, since many exposures—eg, cigarette smoke—lead to numerous outcomes, lack of specificity does not argue against causation. Biological plausibility is another weak criterion, limited by our lack of knowledge. 300 years ago, clinicians would have rejected the suggestion that citrus fruits could prevent scurvy or that mosquitoes were linked with blackwater fever. Ancillary biological evidence that is coherent with the association might be helpful. For example, the effect of cigarette

Panel 2: **Criteria for judgment of causal associations**[17,42,43]

Temporal sequence
Did exposure precede outcome?

Strength of association
How strong is the effect, measured as relative risk or odds ratio?

Consistency of association
Has effect been seen by others?

Biological gradient (dose-response relation)
Does increased exposure result in more of the outcome?

Specificity of association
Does exposure lead only to outcome?

Biological plausibility
Does the association make sense?

Coherence with existing knowledge
Is the association consistent with available evidence?

Experimental evidence
Has a randomised controlled trial been done?

Analogy
Is the association similar to others?

smoke on the bronchial epithelium of animals is coherent with an increased risk of cancer in human beings. Finally, experimental evidence is seldom available, and reasoning by analogy has sometimes caused harm. Since thalidomide can cause birth defects, for instance, some lawyers (successfully) argued by analogy that Bendectin (an antiemetic widely used for nausea and vomiting in pregnancy) could also cause birth defects, despite evidence to the contrary.[46]

Conclusion

Studies need to have both internal and external validity: the results should be both correct and capable of extrapolation to the population. A simple checklist for bias (selection, information, and confounding) then chance can help readers decipher research reports. When a statistical association appears in research, guidelines for judgment of associations can help a reader decide whether the association is bogus, indirect, or real.

We thank Willard Cates and David L Sackett for their helpful comments on an earlier version of this report. Much of this material stems from our 15 years of teaching the Berlex Foundation Faculty Development Course.

References

1 Grimes DA. Technology follies: the uncritical acceptance of medical innovation. *JAMA* 1993; **269:** 3030–33.
2 Last JM, ed. A dictionary of epidemiology, 2nd edn. New York: Oxford University Press, 1988.
3 Ahlbom A, Norell S. Introduction to modern epidemiology, 2nd edn. Chestnut Hill, Massachusetts: Epidemiology Resources, 1990.
4 Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983; **309:** 1358–61.
5 Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia: WB Saunders Company, 1985.
6 Anon. The National Diet-Heart Study Final Report. *Circulation* 1968; **37:** I1–428.
7 Moinpour CM, Lovato LC, Thompson IM Jr, et al. Profile of men randomized to the prostate cancer prevention trial: baseline health-related quality of life, urinary and sexual functioning, and health behaviors. *J Clin Oncol* 2000; **18:** 1942–53.
8 Halbert JA, Silagy CA, Finucane P, Withers RT, Hamdorf PA. Recruitment of older adults for a randomized, controlled trial of exercise advice in a general practice setting. *J Am Geriatr Soc* 1999; **47:** 477–81.
9 Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998; **352:** 609–13.
10 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; **273:** 408–12.
11 Rothman KJ. Modern epidemiology. Boston: Little, Brown and Company, 1986.
12 Grimes DA. The case for confidence intervals. *Obstet Gynecol* 1992; **80:** 865–66.
13 Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 "negative" trials. *N Engl J Med* 1978; **299:** 690–94.
14 Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994; **272:** 122–24.
15 Sackett DL. Bias in analytic research. *J Chronic Dis* 1979; **32:** 51–63.
16 Wingo PA, Higgins JE, Rubin GL, Zahniser SC, eds. An epidemiologic approach to reproductive health. Geneva: WHO, 1994.
17 Hennekens CH, Buring JE. Epidemiology in medicine. Boston: Little, Brown and Company, 1987.
18 Burkman RT. Association between intrauterine device and pelvic inflammatory disease. *Obstet Gynecol* 1981; **57:** 269–76.
19 Kronmal RA, Whitney CW, Mumford SD. The intrauterine device and pelvic inflammatory disease: the Women's Health Study reanalyzed. *J Clin Epidemiol* 1991; **44:** 109–22.
20 Feinstein AR, Horwitz RI. Oestrogen treatment and endometrial carcinoma. *BMJ* 1977; **2:** 766–67.
21 Seltzer CC, Bosse R, Garvey AJ. Mail survey response by smoking status. *Am J Epidemiol* 1974; **100:** 453–57.
22 Schull WJ, Cobb S. The intrafamilial transmission of rheumatoid arthritis: 3, the lack of support for a genetic hypothesis. *J Chronic Dis* 1969; **22:** 217–22.
23 Bartholomew LL, Grimes DA. The alleged association between induced abortion and risk of breast cancer: biology or bias? *Obstet Gynecol Surv* 1998; **53:** 708–14.
24 Lindefors-Harris BM, Eklund G, Adami HO, Meirik O. Response bias in a case-control study: analysis utilizing comparative data concerning legal abortions from two independent Swedish studies. *Am J Epidemiol* 1991; **134:** 1003–08.
25 Harris BM, Eklund G, Meirik O, Rutqvist LE, Wiklund K. Risk of cancer of the breast after legal abortion during first trimester: a Swedish register study. *BMJ* 1989; **299:** 1430–32.
26 Melbye M, Wohlfahrt J, Olsen JH, et al. Induced abortion and the risk of breast cancer. *N Engl J Med* 1997; **336:** 81–85.
27 Abramson JH. Making sense of data. New York: Oxford University Press, 1988.
28 Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman RB, eds. Designing clinical research: an epidemiologic approach, 2nd edn. Baltimore: Lippincott Williams and Wilkins, 2001.
29 Ory HW. Association between oral contraceptives and myocardial infarction: a review. *JAMA* 1977; **237:** 2619–22.
30 Schwingl PJ, Ory HW, Visness CM. Estimates of the risk of cardiovascular death attributable to low-dose oral contraceptives in the United States. *Am J Obstet Gynecol* 1999; **180:** 241–49.
31 Jain AK. Cigarette smoking, use of oral contraceptives, and myocardial infarction. *Am J Obstet Gynecol* 1976; **126:** 301–07.
32 Lee NC, Rubin GL, Ory HW, Burkman RT. Type of intrauterine device and the risk of pelvic inflammatory disease. *Obstet Gynecol* 1983; **62:** 1–6.
33 Lee NC, Rubin GL, Borucki R. The intrauterine device and pelvic inflammatory disease revisited: new results from the Women's Health Study. *Obstet Gynecol* 1988; **72:** 1–6.
34 Lee NC, Rubin GL, Grimes DA. Measures of sexual behavior and the risk of pelvic inflammatory disease. *Obstet Gynecol* 1991; **77:** 425–30.
35 Schlesselman JJ. Cancer of the breast and reproductive tract in relation to use of oral contraceptives. *Contraception* 1989; **40:** 1–38.
36 Lacey JV Jr, Brinton LA, Abbas FM, et al. Oral contraceptives as risk factors for cervical adenocarcinomas and squamous cell carcinomas. *Cancer Epidemiol Biomarkers Prev* 1999; **8:** 1079–85.
37 Kjellberg L, Hallmans G, Ahren AM, et al. Smoking, diet, pregnancy and oral contraceptive use as risk factors for cervical intra-epithelial neoplasia in relation to human papillomavirus infection. *Br J Cancer* 2000; **82:** 1332–38.
38 Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959; **22:** 719–48.
39 Lang TA, Secic M. How to report statistics in medicine. Philadelphia: American College of Physicians, 1997.
40 Rothman KJ. A show of confidence. *N Engl J Med* 1978; **299:** 1362–63.
41 Sterne JA, Smith GD. Sifting the evidence: what's wrong with significance tests? *BMJ* 2001; **322:** 226–31.
42 Hill AB. The environment and disease association or causation. *Proc R Soc Med* 1965; **58:** 295–300.
43 Streiner DL, Norman GR, Munroe Blum H. PDQ epidemiology. Toronto: BC Decker, 1989.
44 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine, 2nd edn. Boston: Little, Brown and Company, 1991.
45 Grimes DA, Economy KE. Primary prevention of gynecologic cancers. *Am J Obstet Gynecol* 1995; **172:** 227–35.
46 McKeigue PM, Lamm SH, Linn S, Kutcher JS. Bendectin and birth defects: 1, a meta-analysis of the epidemiologic studies. *Teratology* 1994; **50:** 27–37.

PRESENTER: Welcome to epidemiology case studies podcast. This is a continuation of our second episode where you heard Dr. Jade Benjamin-Chung interview Dr. Jack Colford about the WASH Benefits study. In this second part, you will learn about the potential sources of bias and confounding, the analysis of the trials, and what the trials found.

JADE BENJAMIN-CHUNG: Welcome back. We're going to continue talking about the WASH Benefits trials, and in this part we're going to focus on the threats to validity, the analysis of the trials, and the results. So with trials, we're a little bit less concerned with different threats to validity than in observational studies. But there could still be different forms of bias. So let's just start with one in particular. Were you concerned about selection bias in this trial?

JACK COLFORD: Well, we were concerned about selection bias, and to that end, we specifically chose to enroll children before birth as similar as we could. And we did that by enrolling women who were pregnant, but in early trimesters of pregnancy, but of course, hadn't delivered yet. So a threat to validity from selection bias would have arisen from having children of all sorts of different ages in the study, for example, but all our children were-- this was done pre-birth for enrollment.

JADE BENJAMIN-CHUNG: And what about information bias? So who was blinded in the trial?

JACK COLFORD: Well, so the participants, of course, couldn't be blinded to which water sanitation hygiene or nutrition interventions they were receiving. But what we did do is we rigorously blinded all the analysis in the trial. So the data were collected and processed, and then when the analysis was done, and we were very rigorous about nobody who was touching the data knew what intervention group a particular family was in. So the analysis was done, completed, and the results declared, and then we had what we called unblinding parties, where as you'll recall, we all got together at a party and opened the envelope, basically, to be told who was in each group and how that group did.

JADE BENJAMIN-CHUNG: So that means when we were analyzing the data the treatment assignment was not known to the analysts. It was using a fake treatment.

JACK COLFORD: Correct. That's right.

JADE BENJAMIN-CHUNG: And so basically, all the results before the unblinding would sort of be fake and scrambled results.

JACK COLFORD: They were all scrambled.

JADE BENJAMIN-CHUNG: And one of the primary outcomes was diarrhea. So how is that measured, and were you concerned about any bias related to that measurement?

JACK COLFORD: Well, one of the big concerns in our field is about how difficult it is to measure diarrhea and whether people who are in a study might be reporting preferentially more diarrhea or less diarrhea based on what they think you want them to be telling you. So one of the key reasons WASH Benefits was designed as it was was to not just measure diarrhea with all its problems, but to rigorously measure growth, because we hypothesized that if we greatly improve diarrhea, we were going to hopefully improve growth as well.

Growth is not easy to measure in a young child who can be squirming and moving a lot, but there are rigorous ways to, in a standardized fashion, with repeated measurements, measure the growth of a child and record that numerically. So that one of these secondary goals of WASH Benefits in a sense was to be able to compare growth measurements to diarrhea measurements in a large scale rigorous way.

JADE BENJAMIN-CHUNG: So what you're essentially saying is that self-reported diarrhea is a bit subjective, based on the person who is answering the question.

JACK COLFORD: Very subjective.

JADE BENJAMIN-CHUNG: Is there no objective measurement of diarrhea that could have been used?

JACK COLFORD: Well, not specifically of diarrhea, but there are-- you could, for example, measure causes of diarrhea, particular pathogens. But those are, for each individual cause, so infrequent generally, that there are never enough cases to be large enough to show a difference between groups. So diarrhea is a good measurement because it's frequent, but it's a difficult measurement because it's self-reported and subjective.

JADE BENJAMIN-CHUNG: And presumably the pathogens causing diarrhea would differ--

JACK COLFORD: In different settings, in different countries, even within Bangladesh, even within different times of the year. The particular pathogens in the rainy season can be different than they are during the summer.

JADE BENJAMIN-CHUNG: So what about confounding? This is a trial. So we're less concerned about that than usual. But were there any steps you took to minimize confounding or account for confounding in the analysis?

JACK COLFORD: Great. So as the students know that we do in class, we adjust for potential confounding variables through multivariate adjustment and other techniques. So in WASH Benefits, we had a pre-specified analysis plan where we adjusted for potential confounding variables with rigorous multivariate models. And then, as is true in all of epidemiology, we compare the results of those adjusted models to the results without adjustment.

And if they differ very little or not at all, then we conclude there's no confounding. And essentially, that's what we found was little to no confounding in the results.

JADE BENJAMIN-CHUNG: And another check for different kinds of bias and confounding would be to look at the table one, the classic first table of the paper summarizing a trial. So in your table looking at characteristics across the different treatment arms in both countries. Did they appear to be pretty similar across the arms?

JACK COLFORD: Great point that we try to create in a trial a counter factual by doing randomization to make groups that are equal to each other in all possible ways. They differ only by random chance. And when we compared to table one for our villages in our different arms, they were extremely well-balanced, so randomization worked in our cluster randomized design.

JADE BENJAMIN-CHUNG: So then given that, the adjustment for confounding could also be not just to adjust for confounding, but to account for-- or to increase the efficiency of the analysis, right?

JACK COLFORD: That's right.

JADE BENJAMIN-CHUNG: You want the standard errors to be a little bit smaller.

JACK COLFORD: If the randomization hadn't perfectly balanced everything, which it never does or can, the adjusted analyses would further help to account for that and adjust for the variance there.

JADE BENJAMIN-CHUNG: Can you tell me just really briefly about the statistical methods they can use to analyze the trial?

JACK COLFORD: Well, so in addition to the kind of basic simple analysis, which is just sort of proportions with different outcomes, we used advanced multivariate models to adjust for all these covariates that we had, as you'll recall from causal inference, we don't want to adjust for covariates that might be on the causal pathway, but we adjust for covariates that might be confounders.

And so these generalized estimating equation models and other types of advanced models that you'll see in the paper were used to carry out this adjustment.

JADE BENJAMIN-CHUNG: Earlier, we talked about compliance and sort of whether or not the participants actually use the interventions. Can you tell me really briefly what was found in each country and how that affected the results?

JACK COLFORD: So in Bangladesh, compliance was really tremendous and excellent through basically both years of the study. And of course, we're talking about multiple different interventions here. So this conversation would have to cover each of the different interventions for water, sanitation, hygiene, and nutrition. But as I said, in Bangladesh, they were quite good across both years.

In Kenya, sanitation, of course, here we're constructing something, so it tends not to go away, but that compliance was quite good across both years. Compliance with the water interventions was very good in the first year and then decreased a bit in the second year. But we also analyzed the data at year one and year two to see if that was having any impact on the trends we were seeing. And it didn't seem to.

JADE BENJAMIN-CHUNG: Is that challenge of kind of getting high compliance and a wash or nutrition trial common? And what are some examples of reasons why it might be difficult to achieve that?

JACK COLFORD: So this is a big problem in wash studies. And one of the issues in the literature and one of the reasons WASH Benefits we think was an important trial was so many wash studies are much shorter than WASH Benefits. They're three months long or six months long. WASH benefits was a to full years of intervention. So that was entirely because of the fact that we knew that people are quite good about taking up a new intervention, but then it can drop off rapidly.

Often, that's not measured. We measured it rigorously. Often, it drops off a lot. It didn't drop off a lot in WASH Benefits.

JADE BENJAMIN-CHUNG: If there is imperfect compliance-- as you mentioned, there is a little bit of this in Kenya, would you then expect that there would be less of a health impact of those interventions?

JACK COLFORD: Theoretically you would expect there potentially be less of a benefit from any intervention that isn't complied with. If the intervention actually works, and it's not complied with, you'd expect to see a decrement in its effectiveness.

JADE BENJAMIN-CHUNG: So in terms of the results, can you give me sort of a high level summary of what was found and what portion of the results conformed with your expectations, and also what was surprising?

JACK COLFORD: So let's talk about Bangladesh first. In Bangladesh, there was a reduction in diarrhea in many of the arms, but not a statistically significant reduction in the water arm. So that was a little surprising. But their reduction was occurring, surprisingly in Bangladesh, there was less baseline diarrhea than we expected. So that was a surprising finding. But we saw a reduction in diarrhea, but it was from a level that was a little bit lower than we expected to see from our preliminary work and from other studies in the field.

But importantly in Bangladesh, we did not see an impact on growth of the children. Growth was not improved except in the arms that had nutrition. So the arm that had nutrition, the nutrition supplement or the arm that had wash plus nutrition, we did see, as expected, about the same amount as expected an improvement in growth in the children.

In Kenya, we didn't see an impact on diarrhea. And we also didn't see an impact on growth in the non nutrition arms. The nutrition arms also impacted growth in Kenya. So the results were felt to be very consistent with respect to growth across both countries and actually, frankly, quite surprising.

JADE BENJAMIN-CHUNG: Was it a disappointment to people in the wash and nutrition communities?

JACK COLFORD: I'd say it wasn't disappointment in the sense that we believe the study was done very rigorously and followed all its pre-specified protocols quite strongly. So the lack of impact on growth was surprising and sent us back to the drawing board to think of all the different pathways by which growth occurs and where it can be intervened on to make a difference in the growth of children.

JADE BENJAMIN- So in reflecting on these findings, sort of thinking forward about policy implications and

CHUNG: research implications, how generalizable do you think the study's findings are?

JACK COLFORD: Well, this is an important question, generalizability, when you talk about epidemiology studies all the time is do they apply only to the countries where they're done, or do they apply to similar countries? So that's a question we still are trying to grapple with. But I will share that yet a very similar study to WASH Benefits is being done by another group funded by the Gates Foundation in Zimbabwe. And they're going to release their results quite soon, and I'll just, at this point, say those are going to be quite interesting as well.

But all in all, I'd say the there's going to have to be a lot more thought about ways to impact growth in children. And I know this from conversations with the World Bank and USAID that this is causing a rethinking of how WASH should be integrated into nutrition programs, for example.

JADE BENJAMIN-CHUNG: And then, as epidemiologists, we always hope to be able to make causal inferences from our research. In this particular case, do you feel comfortable making causal inferences about the findings in these trials?

JACK COLFORD: Well, I would say I only feel comfortable saying we don't quite understand the full pathway of growth. So I feel like we've really nailed down the lack of impact, the lack of causal impact of these interventions on our ultimate target, but we don't yet understand the interaction of all the different pathways that go together to form growth.

So I think this is going to advance the field in terms of how to do the next study. But we're not there yet.

JADE BENJAMIN-CHUNG: So how do you think the results from this trial will impact future research on this research question?

JACK COLFORD: Well, there's no doubt that the clean water and a sanitary environment is important for kind of basic life functions. But in terms of impacting growth, the impact this is going to have is-- my own opinion here is that what's going to have to happen now are packages of interventions that are much broader than just water sanitation, hygiene, and nutrition. I mean, there are many other interventions being discussed and talked about being integrated into these kinds of settings, such as cement flooring, for example, to help clean up the local environment for young children as they play on the floor.

Perhaps there's going to be much more discussion of more intensive kind of wash interventions and just the self-treatment of water. You know, when we look at how countries have evolved over time, it's really only when water systems have been brought to the countries with, you know, centralized treatment of the water, and that may be where we have to head in these settings, as well, sort of super wash.

JADE BENJAMIN-CHUNG: So this super wash idea is really suggesting that the interventions focused on wash and wash benefits weren't enough to critically reduce environmental contamination.

JACK COLFORD: That's the question we're left with is were they not strong enough, even though we believe that the WASH Benefits interventions were as good as is generally done in the field, and were really intensive and measured compliance and so forth that it may be that even stronger WASH interventions are going to be needed to solve this problem.

JADE BENJAMIN-CHUNG: My last question for you, if you could go back in time, is there anything you would change about the design or the analysis of these trials?

JACK COLFORD: I think we would have paid much more attention to the long term follow up of the participants in the studies. I think we would have sought out more support for banking of specimens to do more work. Those would be my initial two reactions there.

JADE BENJAMIN-CHUNG: And the banking of specimens is valuable. You're talking about blood, stool, saliva, so that in the future, you can measure different specific pathogens?

JACK COLFORD: Correct. Once you've done a randomized trial, that randomization is available to you forever. And so you're only limited by what you've collected and able to study going forward.

JADE BENJAMIN-CHUNG: And it must be very expensive to collect all those things and store them securely for a long periods of time.

JACK COLFORD: That's correct. Yeah. The logistics of that and the costs of that are large.

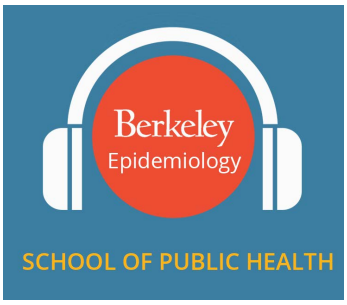
JADE BENJAMIN-CHUNG: Well, good food for thought for the future.

JACK COLFORD: Yes.

JADE BENJAMIN-CHUNG: Thanks so much.

JACK COLFORD: Thank you, Jade.

JADE BENJAMIN-CHUNG: Thanks for listening to us talk about the WASH Benefits trials. We hope you enjoy learning about them.



Epidemiology Case Studies Podcast: Episode 5 – Acute Changes in community violence and increases in hospital visits and deaths from stress-responsive diseases, Part 2

MICHELLE RUIZ: Hello welcome once again to Epidemiology case studies Podcast. I am Michelle Ruiz, instructional designer at the school of public health. This is a continuation of our forth episode where you heard Dr. Jade Benjamin Chung interview Dr. Jennifer Ahern about the ecological study she conducted. In This second part Dr. Benjamin Chung and Dr. Ahern will dive into analysis methods, threats to validity, and the results of this study.

JADE BENJAMIN-CHUNG: **W**e're going to talk about the threats to validity and the analysis and results from this study. So tell me what your principal concerns are in this analysis with respect to confounding and bias.

JENNIFER AHERN: So the biggest threat would be some sort of time varying factor that is affecting both violence and the health outcomes that we're studying. It has to be something that is not following any sort of predictable temporal pattern because our method has sort of pulled all the predictable temporal patterning out of both our exposure and our outcome before looking at how the two are related.

So it would be something that's following no particular predictable pattern, that's just sort of happening, and when it happens, it's instigating both violence and health problems. But it's not the violence causing the health outcome. So we had a hard time thinking about what something like that could be in this setting.

We did go through news records to identify events of civil unrest, because we thought those are the sort of thing that could be a problem. So for example, if something happens that causes outrage in the community and it leads to both protesting that may turn violent, as well as then people who are upset about whatever the underlying event was who are then experiencing an asthma attack or whatever and showing up at the hospital.

And so that's the category. That's why we controlled for events of civil unrest. But if there were some other thing like that that we didn't think through, that's what we would worry about.

JADE BENJAMIN-CHUNG: So you kind of referred to removing seasonality and temporal trends. Can give me a high level overview of the statistical approach used in this analysis to do that?

JENNIFER AHERN: Definitely. So there's sort of a category of methods that are called time series, and they're intended to remove predictable temporal patterning from any sort of series of data points in time. So we can think about there a lot of health events, including violence, that tend to be higher in the summer and lower in the winter.

And so that sort of seasonal patterning in things. For example, one of our outcomes, asthma, is extremely strongly seasonally patterned as well. So you could think about how if we just looked at the correlation between monthly violence and monthly asthma without worrying about temporal patterning, they're probably very strongly correlated, not because of any real causal connection, but because they both share seasonality in patterning.

And so we want to remove that sort of patterning from our series before we look at how they're related to each other. And so seasonality is one form of temporal patterning we remove. The other types of things are trends. So if a series has been generally increasing over time, you would predict that it continue to do that, and you'd want to remove that sort of trend.

And so the general category of time series methods is intended to remove anything predictable from this that you could have predicted about this series from how it's behaved in the past before you then look at how things are related to each other.

JADE BENJAMIN-CHUNG: Is this kind of method increasingly being used by epidemiologists? Because it sounds like it comes more from the economics literature, but is it gaining speed in epidemiology?

JENNIFER AHERN: I don't know if it's gaining speed, generally, in epidemiology. We've certainly found it useful for these questions about acute changes in things, where you just really want to avoid this pitfall that you could easily fall into if you don't worry enough about the temporal patterning. And I know economists who do time series methods think that epidemiologists don't worry enough about temporal patterning.

If your study's over a couple of years and you're not looking at something that's varying at a fine grain time scale, you're probably fine. But if you're talking about 20 years of follow up, we do ignore it probably more than we ought to know. And they worry about it in a more upfront way. So the disciplines are maybe helping each other--

JADE BENJAMIN-CHUNG: That's great.

JENNIFER --think different things through more carefully.

AHERN:

JADE BENJAMIN-CHUNG: So tell me about the results, sort of at a high level. What did you find?

JENNIFER AHERN: So we found from some of the outcomes, where we'd hypothesized that acute changes in violence would relate to these acute changes and disease outcomes, we did find some associations. We found increases in anxiety-related disorders, visits to the hospital for those. Similarly for asthma events and substance use as well.

And then, in the realm of cardiovascular things, we found increases in fatal heart attacks. So not everything we had hypothesized was there an association, but for a variety of the things we did see an effect.

JADE BENJAMIN-CHUNG: Thinking about the magnitude of the effects, are you able to give us a sense of how big or how small they were relative to some standard difference in risk that we think about?

JENNIFER AHERN: Sure. I mean, so one of the sort of biggest caveats about this study is that these are quite small associations. So whether you should worry about them from sort of a public health perspective, I think, is a question. We sort of view them as a tip of the iceberg. These are people who have had such a bad asthma attack that they're in the hospital.

If you've seen an uptick in who's in the hospital, you could imagine, well, out in the community, there may have been a much larger group of people who had exacerbations and were using rescue medicine or whatever. Things might have been going on but maybe it didn't get to the point that they had to go to the hospital.

And so I think if we can see these-- really, we're looking at pretty severe outcomes. If we can think of them as being just sort of a documentation of a signal that may represent a broader underlying health impact, then I think we can think of them as meaningful. But others have looked at the work and just said, well, I see that there's something here, but it doesn't look big enough to really worry about. I wouldn't feel like I had to intervene on this exposure in response to this magnitude.

So I think it's open for interpretation. We tried to be really clear. In the paper we translated into literally numbers of deaths that this would represent, to try to just be really transparent about the magnitude. And I think as sort of an aside, I think papers don't necessarily always do that. We worked on the additive scale, so you can really see what this is as a difference in the rate. If we just reported ours, you wouldn't necessarily-- it might not jump as quickly to your

attention this didn't represent a huge number of additional cases.

JADE BENJAMIN-CHUNG: So building on that, I know it's recently published or recently accepted for publication, but presumably you've shared the results with folks in the social epi community, and I'm curious what the reaction has been and if you expect that this will influence future research or public health programming.

JENNIFER AHERN: Well, I think the reaction has been really interesting. The first journal we sent it to the reaction was, it's an ecological design. And I think I know who that was reading it, and I was a little disappointed that they couldn't see beyond that. But that is still a chronic problem with this kind of work, is that people have certain assumptions about what the limitations of that design and aren't thinking through, OK, do those limitations apply here?

On the other side, though, we had-- hopefully I can get you a version of the commentary. The journal asked for a commentary to be written on the paper, and it raised some interesting ideas or different directions we could have gone with things, but seemed to really view it as a pretty strong contribution in this area of looking at violence and health that's been plagued by a lot of-- what is out there to date has been plagued by a lot of limitations.

JADE BENJAMIN-CHUNG: So what would be the ideal study design for this question? And is this the beginning of a series of studies you're going to lead on this question? For example, would you hope to, in the future, lead a cohort study on this question, or is this sort of the final study design that you want to use for this?

JENNIFER AHERN: That's a good question. I mean, I think for this type of work, for work on violence and health to be actionable, I think that the direction it needs to go is in looking at whether-- you know, we do a lot of work in different cities to try to reduce violence, and sometimes it's effective. And I think we would want to be adding to evaluations of those an assessment of whether that's actually improving health in general.

Because it's one thing to show that there's this relationship, but if when we intervene to reduce violence we don't actually see benefits for health, then that's sort of taking the evidence to a more actionable level. And that would be the next step from my perspective. And so we are, as, I would say, early groundwork for that, we're trying to just figure out where violence reduction efforts have actually been effective, which is using the same data resources. And that's sort of in itself been a lot more complicated than-- I don't know, than maybe I anticipated

initially.

As an example, there are places where we're seeing an intervention intended to reduce a particular form of violence being perhaps successful in doing that, but then seeing upticks in other forms of violence. So there's a lot of complexity that's coming out of our attempts to do some things that we thought would be more straight forward. And then, of course, you'd want ultimately to be able to do a trial of some kind.

But the tricky thing is that the reality of violence prevention is it's done by cities. It's always going on. It's probably a little tricky to randomize it unless it's just in time, with not a lot of lag. And so it may be that we can't rely on what we think of as the gold standard design for a lot of this because we need to understand how violence prevention efforts that are just being carried out, how they're affecting health.

Are they effective in reducing violence, and is that better for the community health overall? It may not be a situation where we're going to get a lot of trial work in. But I think it's really important nonetheless.

JADE BENJAMIN-CHUNG: And I think it's a great example of how observational studies can be really useful and helpful when we can't necessarily do good trials. Well, thank you so much. It's been a pleasure talking to you.

JENNIFER AHERN: Yeah. Thank you.

JADE BENJAMIN-CHUNG: And that concludes our interview. Hope you enjoyed learning about this study.

American Journal of Epidemiology

Vol. 186, No. 7
DOI: 10.1093/aje/kwx019
Advance Access publication:
May 11, 2017

Original Contribution

Acute Illness Among Surfers After Exposure to Seawater in Dry- and Wet-Weather Conditions

Benjamin F. Arnold*, Kenneth C. Schiff, Ayse Ercumen, Jade Benjamin-Chung, Joshua A. Steele, John F. Griffith, Steven J. Steinberg, Paul Smith, Charles D. McGee, Richard Wilson, Chad Nelsen, Stephen B. Weisberg, and John M. Colford, Jr.

* Correspondence to Dr. Benjamin F. Arnold, Division of Epidemiology, School of Public Health, University of California, Berkeley, 101 Haviland Hall, MC #7358, Berkeley, CA 94720-7358 (e-mail: benarnold@berkeley.edu).

Initially submitted September 8, 2016; accepted for publication January 23, 2017.

Rainstorms increase levels of fecal indicator bacteria in urban coastal waters, but it is unknown whether exposure to seawater after rainstorms increases rates of acute illness. Our objective was to provide the first estimates of rates of acute illness after seawater exposure during both dry- and wet-weather periods and to determine the relationship between levels of indicator bacteria and illness among surfers, a population with a high potential for exposure after rain. We enrolled 654 surfers in San Diego, California, and followed them longitudinally during the 2013–2014 and 2014–2015 winters (33,377 days of observation, 10,081 surf sessions). We measured daily surf activities and illness symptoms (gastrointestinal illness, sinus infections, ear infections, infected wounds). Compared with no exposure, exposure to seawater during dry weather increased incidence rates of all outcomes (e.g., for earache or infection, adjusted incidence rate ratio (IRR) = 1.86, 95% confidence interval (CI): 1.27, 2.71; for infected wounds, IRR = 3.04, 95% CI: 1.54, 5.98); exposure during wet weather further increased rates (e.g., for earache or infection, IRR = 3.28, 95% CI: 1.95, 5.51; for infected wounds, IRR = 4.96, 95% CI: 2.18, 11.29). Fecal indicator bacteria measured in seawater (*Enterococcus* species, fecal coliforms, total coliforms) were strongly associated with incident illness only during wet weather. Urban coastal seawater exposure increases the incidence rates of many acute illnesses among surfers, with higher incidence rates after rainstorms.

diarrhea; *Enterococcus*; rain; seawater; waterborne diseases; wound infection

Abbreviations: CI, confidence interval; IRR, incidence rate ratio.

Freshwater runoff after rainstorms increases levels of fecal indicator bacteria measured in seawater (1), but little is known about whether persons who participate in ocean recreation have a higher risk of acute illness after rainstorms. Absent epidemiologic studies to inform beach management guidelines after rainstorms, California beach managers post advisories at beaches that discourage contact with seawater for 72 hours after rainfall—a practice that is based on fecal indicator bacteria profiles in storm water outflows, which typically decline to prerainstorm levels within 3–5 days (2, 3).

In prospective cohorts in California, investigators have found increased incidence of gastrointestinal illness and other acute symptoms (e.g., eye and ear infections) associated with seawater exposure during dry summer months (4–8). In the same studies, researchers found that levels of fecal indicator bacteria in seawater were positively associated with incident gastrointestinal illness if there was a well-defined source of human fecal contamination impacting the seawater (4–8). Individual cases of acute infections and deaths associated with waterborne pathogens have been reported among surfers in southern California who surfed during or after rainstorms (9), and 2 cross-sectional studies of surfers found that seawater exposure after heavy rainfall increased reported illness (10, 11). To our knowledge, there have been no prospective studies to determine whether rainstorms increase illness among persons who participate in ocean recreation and no studies that have evaluated whether levels of fecal indicator bacteria are associated with incident illness during wet weather periods.

We conducted a longitudinal cohort study among surfers in San Diego, California. We focused on surfers because they are a well-defined population that regularly enters the ocean year-round, even during and immediately after rainstorms, given that surfing conditions often improve during storms (12). Our objectives were to determine whether exposure to seawater increased rates of incident illness among surfers compared with periods when they did not surf in order to determine whether exposure during or immediately after rainstorms increased rates more than did exposure during dry weather. We also sought to evaluate the relationship between levels of fecal indicator bacteria in seawater and incident illness rates during dry and wet weather.

METHODS

Setting

Southern California has one of the most urbanized coastlines in the world, and it receives nearly all of its annual rainfall during the winter months (November–April). San Diego County beaches have some of the best water quality in California based on levels of fecal indicator bacteria, but water quality deteriorates after rainstorms (13). The most heavily used beaches in the region are affected by urban runoff after storms, and local beach managers post advisories that discourage water contact within 72 hours of rainfall. In the present study, we focused enrollment and conducted extensive water quality measurement at 2 monitored beaches within San Diego city limits—Ocean Beach and Tourmaline Surfing Park. Both monitored beaches have storm-impacted drainage, attract surfers year-round, and have water quality levels similar to those of other beaches in the county (13). Ocean Beach is adjacent to the San Diego river, which drains a 1,088-km^2 varied land-use watershed with many flow-control structures; Tourmaline Surfing Park is adjacent to Tourmaline Creek and a storm drain, which together drain an urban, largely impervious, 6-km^2 watershed (Figure 1). The study's technical report includes additional details (14).

Study design and enrollment

We conducted a longitudinal cohort study of surfers recruited in San Diego over 2 winters, with enrollment and follow-up periods chosen to capture most rainfall events in the region. During the first winter (open enrollment from January 14, 2014, to March 18, 2014; end of follow-up on June 4, 2014), we enrolled surfers through in-person interviews at the 2 monitored beaches and through targeted online advertising on Surfline.com, a popular website on which surf conditions are reported. We enrolled participants at monitored beaches and online to assess whether individuals enrolled through these 2 modes were similar in their exposures and other characteristics. Participants enrolled on the beach were very similar to those enrolled online (Table 1), so we exclusively enrolled participants through the study's website during the second winter (open enrollment from December 1, 2014,

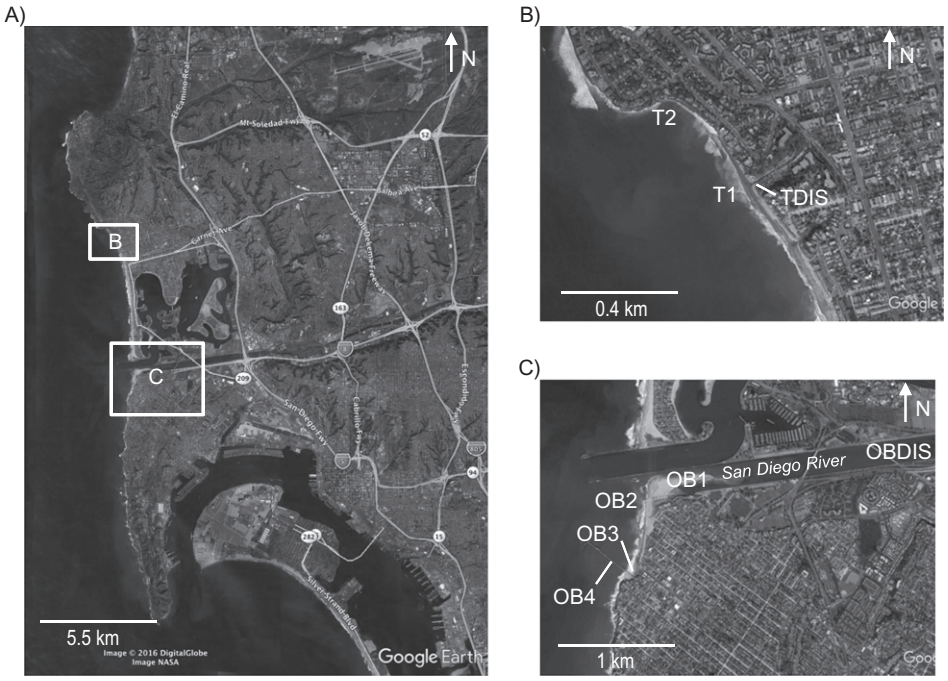


Figure 1. Monitoring beach water quality sampling locations in San Diego, California, winters of 2013–2014 and 2014–2015. Shown are the locations of the 2 monitored beaches along the San Diego coastline (A) and the water quality sampling sites at Tourmaline Surfing Park (B) and Ocean Beach (C). Samples were only collected at Ocean Beach and Tourmaline Surfing Park discharge locations (OBDIS and TDIS, respectively) during wet weather. Wet weather was defined as 0.25 cm or more of rain in 24 hours. T1 and T2, Tourmaline Surfing Park sampling sites 1 and 2; OB1–OB4, Ocean Beach sampling sites 1–4. Map Data: Google, Digital Globe, NASA.

Table 1. Characteristics of the Study Population by Mode of Enrollment, San Diego, California, 2013–2015

Characteristic	Beach[a]		Online[a]		Total	
	No.	%	No.	%	No.	%
No. of participants	89		565		654	
Participants with background survey	72	100	535	100	607	100
Age, years[b]						
18–30		35		35		35
31–40		22		26		26
41–50		11		16		16
≥51		29		13		15
Unreported		3		9		8
Female sex		19		21		21
College educated		68		63		63
Currently employed		74		76		75
Household income[b]						
<$15,000		11		6		7
$15,000–$35,000		15		10		11
$35,001–$50,000		11		7		7
$50,001–$75,000		8		13		12
$75,001–$100,000		17		14		14
$100,001–$150,000		17		14		14
>$150,000		7		13		12
Unreported		14		23		22
Days of surfing per week[b]						
≤1		11		15		14
2		12		18		17
3		26		26		26
4		26		20		21
≥5		24		18		19
Unreported		1		3		3
Chronic health conditions						
Ear problems		12		14		14
Sinus problems		7		8		8
Gastrointestinal condition		0		3		2
Respiratory condition		4		3		3
Skin condition		1		6		5
Allergies		10		16		15
Total days of observation	2,623	100	30,754	100	33,377	100
Days of observation by exposure						
Unexposed		46		47		47
Dry-weather exposure		48		43		43
Wet-weather exposure		6		10		10

[a] Beach enrollment only took place during the first winter (2013–2014); online enrollment spanned both winters (2013–2014 and 2014–2015). The study enrolled 73 individuals online during the first winter.

[b] Percentages within categories might not sum to 100 because of rounding.

to March 22, 2015; end of follow-up on April 16, 2015). We recruited surfers through postcards distributed at the monitored beaches and through an electronic newsletter distributed by the Surfrider Foundation's San Diego County chapter. Surfers were eligible if they were 18 years of age or older, could speak and read English, planned to surf in southern California

during the study period, had a valid e-mail address or mobile telephone number, and could access the internet with a computer or smartphone.

Participants completed a brief enrollment questionnaire, and each Tuesday they received a text message or e-mail reminder to complete a short weekly survey. Participants reported daily surf activity (location, date, and times of entry and exit) and illness symptoms (details below) for the previous 7 days using the study's web or smartphone (iOS or Android) application. We used an open cohort design in which participants were allowed to enter and exit the cohort over the follow-up period. We excluded follow-up time during which participants reported surfing outside of southern California. The study protocol was reviewed and approved by the institutional review board at the University of California, Berkeley, and all participants provided informed consent. Participants received a modest incentive for participation ($20 gift certificate per 4 weekly surveys completed). Web Table 1 (available at https://academic.oup.com/aje) includes a Strengthening the Reporting of Observational Studies in Epidemiology checklist.

Outcome definition and measurement

In weekly surveys, participants reported daily records of the following symptoms: diarrhea (defined as ≥3 loose/watery stools in 24 hours), sinus pain or infection, earache or infection, infection of an open wound, eye infection, skin rash, and fever. During the second winter, we added sore throat, cough, and runny nose. We created composite outcomes from the symptoms, including: gastrointestinal illness, which was defined as 1) diarrhea, 2) vomiting, 3) nausea and stomach cramps, 4) nausea and missed daily activities due to gastrointestinal illness, or 5) stomach cramps and missed daily activities due to gastrointestinal illness (15); and upper respiratory illness, which was defined as any 2 of the following: 1) sore throat, 2) cough, 3) runny nose, and 4) fever (16). We created a composite outcome of "any infectious symptom" defined as having any 1 of the following: gastrointestinal illness, diarrhea, vomiting, eye infection, infection of open wounds or fever. Our rationale was that it would exclude outcomes that could potentially have noninfectious causes (earache or infection, sinus pain or infection, skin rash, upper respiratory illness) and would capture a broad spectrum of sequelae associated with waterborne pathogens. We defined incident episodes as the onset of symptoms preceded by 6 or more symptom-free days to increase the likelihood that separate episodes represented distinct infections (17, 18).

Exposure definition and measurement

We classified the 3 days after each seawater exposure as exposed periods and all other days of observation as unexposed periods. We defined wet-weather exposure as exposure to seawater within 3 days of 0.25 cm or more of rainfall in a 24-hour period, which is the rainfall criterion used by San Diego County for posting wet-weather beach advisories; we classified all other seawater exposure as dry-weather exposure. We used rainfall measurements from the National Oceanic and Atmospheric Administration Lindbergh Field Station. Among surfers, most exposure took place during the morning hours, so if a storm's precipitation started after 12:00 PM, we did not classify that day as wet weather (only the following day) to reduce exposure misclassification.

Staff collected daily water samples from January 15, 2014, to March 5, 2014, and from December 2, 2014, to March 31, 2015, at 6 sites across the 2 monitored beaches (Figure 1). Staff collected 1-liter water samples in the morning (08:30 AM ± 2 hours) just below the water surface (0.5–1.0 meters) in sterilized, sample-rinsed bottles. We sampled discharges during 6 rainstorms immediately upstream from where Tourmaline Creek and the San Diego River discharge to the sea (Figure 1). We tested samples for culturable *Enterococcus* (US Environmental Protection Agency method 1600), fecal coliforms (standard method 9222D), and total coliforms (standard method 9222B). All laboratory analyses met quality-control objectives for absence of background contamination (blanks) and precision (duplicates).

Statistical analysis

We prespecified all analyses (19). Web Appendices 1 and 2 contain statistical details and sample size calculations. In the seawater exposure analysis, we calculated incidence rates by dividing incident episodes by person-days in unexposed and exposed periods during follow-up. If participants missed weekly surveys during follow-up, we did not include those periods in the analysis. We measured the association between seawater exposure and subsequent illness using an incidence rate ratio, which we estimated using a log-linear rate model with robust standard errors to account for repeated observations within individuals (20, 21). To examine illness rates separately for dry- and wet-weather exposures, we created a 3-level categorical exposure that classified each participant's follow-up time into unexposed, dry-weather exposure, and wet-weather exposure periods. We calculated a log-linear test of trend in the incidence rate ratios for dry- and wet-weather exposures (22).

In the fecal indicator association analysis, we estimated the association between levels of fecal indicator bacteria and illness using the subset of surf sessions matched to water-quality indicator measurements at the monitored beaches. We matched daily geometric mean indicator levels to surfers by beach and date (weighted by time in water if recent exposure included multiple days).We modeled the relationship between indicator levels and illness using a log-linear model and estimated the incidence rate ratio associated with a $1-\log_{10}$ increase in indicator level. We also estimated the incidence rate ratio associated with exposures to water above versus below US Environmental Protection Agency regulatory guidelines (geometric mean *Enterococcus* >35 colony-forming units per 100 mL) (23) or, in a second definition, if any single sample on the exposure day exceeded 104 colony-forming units per 100 mL. We hypothesized that the relationship between fecal indicator bacteria and illness could be modified by dry- or wet-weather exposure and allowed the exposure-response relationship to vary during dry and wet weather by including an indicator for wet-weather periods and a term for the interaction between indicator bacteria levels and the indicator of wet weather. We controlled for potential confounding (24) from demographic,

exposure-related, and baseline health characteristics (Web Appendix 1). In Web Appendices 3–6 we describe additional analyses, including conversion of estimates to the absolute risk scale, sensitivity analyses, and negative control exposure analyses (25, 26).

RESULTS

Study population

We enrolled 654 individuals who contributed on average 51 days of follow-up (range, 6–139 days). The study population's median age was 34 years (interquartile range, 27–45), and the majority of participants were male (73%), college-educated (63%), and employed (75%) (Table 1). Follow-up included 33,377 person-days of observation after excluding time spent outside of southern California (623 person-days). We excluded from adjusted analyses 47 individuals (1,599 person-days of observation) who provided outcome and exposure information but failed to complete a background questionnaire and thus had missing covariate information.

Water quality and surfer exposure

There were 10 rainstorms with 0.25 cm or more of rain during the study. Field staff collected 1,073 beach water samples and 92 wet-weather discharge samples for fecal indicator bacteria analysis. Median *Enterococcus* levels were higher during wet weather than during dry weather (Figure 2). During follow-up, surfers entered the ocean twice per week on average and experienced 10,081 total days of seawater exposure, including 1,327 days of wet-weather exposure. Surfers were less likely to enter the ocean during or within 1 day of rain. The median ocean entry time was 08:00 AM (interquartile range, 06:45–10:30 AM), and the median time spent in the water was 2 hours (interquartile range, 1–2 hours) (Web Figure 1). Of the 10,081 exposure days, surfers reported wearing a wetsuit during 95%, immersing their head during 96%, and swallowing water during 38%. The most frequented surf locations were the 2 monitored beaches: Tourmaline Surfing Park (25% of surf days) and Ocean Beach (16% of surf days), which reflected targeted enrollment at those beaches (Web Figure 2). There were 5,819 days of observation matched to water-quality measurements at monitored beaches, including 1,358 days during wet weather.

Illness associated with seawater exposure

Seawater exposure in the past 3 days was associated with increased incidence rates of all outcomes except for upper respiratory illness (Web Table 2). Unadjusted and adjusted incidence rate ratio estimates were similar, and for most outcomes, adjusted incidence rate ratios were slightly attenuated toward the null (Web Table 2). With the exception of fever and skin rash, incidence rates increased from unexposed to dry-weather exposure to wet-weather exposure periods (Table 2), a pattern also present on the risk scale (Web Figure 3). Compared with unexposed periods, wet-weather exposure led to the largest relative increase in earaches/infections (Table 3; adjusted incidence rate ratio (IRR) = 3.28, 95% confidence interval (CI): 1.95, 5.51) and infection of open wounds (Table 3; adjusted IRR: 4.96, 95% CI: 2.18, 11.29). Sensitivity analyses that shortened the wet-weather window increased the difference between dry- and wet-weather incidence rates for most outcomes (Web Figure 4).

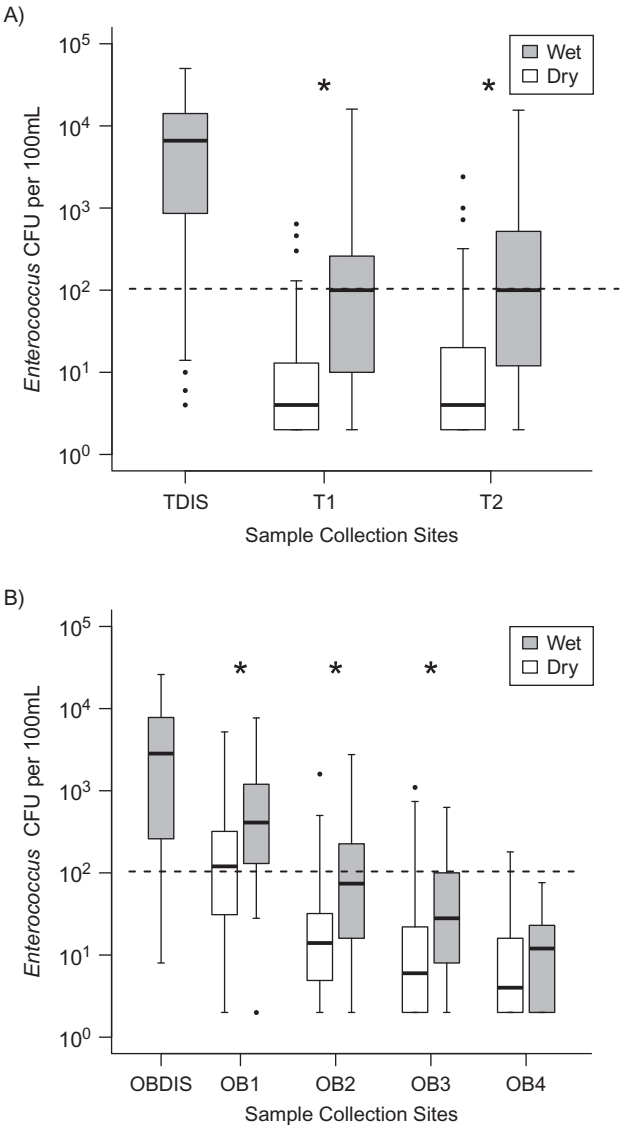


Figure 2. *Enterococcus* levels during dry and wet weather at the sampling locations at Tourmaline Surfing Park (A) and Ocean Beach (B) mapped in Figure 1. Boxes mark interquartile ranges, vertical lines mark 1.5 times the interquartile range, and points mark outliers. Horizontal dashed lines mark the single-sample California recreational water quality guideline (104 CFU/100 mL). Asterisks (*) identify sampling locations with levels that differ between wet and dry periods based on a 2-sample, 2-sided t-test ($P < 0.05$) assuming unequal variances. Samples were only collected at Ocean Beach and Tourmaline Surfing Park discharge locations (OBDIS and TDIS, respectively) during wet weather. Wet weather was defined as 0.25 cm or more of rain in 24 hours. CFU, colony-forming units; T1 and T2, Tourmaline Surfing Park sampling sites 1 and 2; OB1–OB4, Ocean Beach sampling sites 1–4.

Table 2. Incidence Rates Among Surfers by Type of Seawater Exposure, San Diego, California, 2013–2015

Outcome	Unexposed Periods			Dry-Weather Exposure			Wet-Weather Exposure[a]		
	No. of Episodes	No. of Days at Risk	Rate per 1,000	No. of Episodes	No. of Days at Risk	Rate per 1,000	No. of Episodes	No. of Days at Risk	Rate per 1,000
Gastrointestinal illness	90	14,884	6.0	116	13,769	8.4	31	3,037	10.2
Diarrhea	75	15,086	5.0	88	13,909	6.3	27	3,061	8.8
Sinus pain or infection	109	14,475	7.5	139	13,391	10.4	37	2,998	12.3
Earache or infection	59	14,931	4.0	111	13,618	8.2	37	3,008	12.3
Infection of open wound	14	15,456	0.9	30	14,080	2.1	11	3,119	3.5
Skin rash	42	15,024	2.8	66	13,750	4.8	15	3,007	5.0
Fever	51	15,156	3.4	69	14,138	4.9	6	3,152	1.9
Upper respiratory illness[b]	117	12,001	9.7	111	11,025	10.1	31	2,543	12.2
Any infectious symptom[c]	138	14,445	9.6	181	13,176	13.7	47	2,926	16.1

[a] Defined as entering the sea within 3 days of 0.25 cm or more of rain in 24 hours.
[b] Only measured in year 2 of the study.
[c] Includes gastrointestinal illness, eye infections, infected wounds, and fever.

Illness associated with fecal indicator bacteria levels

Enterococcus, total coliform, and fecal coliform levels were positively associated with increased incidence of almost all outcomes during the study (Web Table 3). Rainfall was a strong effect modifier of the association (Table 4). During dry weather, there was no association between *Enterococcus* levels and illness except for infected wounds, but *Enterococcus* was strongly associated with illness after wet-weather exposure (e.g., for each $\log_{10}$ increase, gastrointestinal illness IRR = 2.17, 95% CI: 1.16, 4.03; Table 4, Web Figure 5, and Web Table 4). Associations were attenuated in adjusted analyses, but relationships were similar (e.g., for gastrointestinal illness, wet-weather IRR = 1.75, 95% CI: 0.80, 3.84; Table 4). There was evidence for excess risk of gastrointestinal illness at higher *Enterococcus* levels only during wet-weather periods (Web Figure 6): The predicted excess risk that corresponded to the current US Environmental Protection Agency regulatory guideline of 35 colony-forming units per 100 mL was 16 episodes per 1,000 (95% CI: 5, 27). Negative control analyses showed no consistent association between fecal indicator bacteria and illness among participants during periods in which they had no recent seawater contact (Web Table 5).

DISCUSSION

Key results

To our knowledge, this is the first prospective cohort study in which the association between incident illness and exposure to seawater in wet weather has been measured, and the findings represent novel empirical measures of incident illness associated with storm water discharges. There was a consistent increase in acute illness incidence rates between unexposed, dry-weather, and wet-weather exposure periods (Tables 2 and 3). Rainstorms led to higher levels of fecal indicator bacteria (Figure 2), and a sensitivity analysis illustrated that a 2–3 day window after rainstorms captured the majority of excess incidence associated with wet-weather exposure (Web Figure 4). Fecal indicator bacteria matched to individual surf sessions were strongly associated with illness only during wet weather periods (Table 4, Web Figure 5).

Interpretation

Swimmers are more rare during the winter months, and surfers' frequent and intense exposure made them an ideal population in which to study the relationship between illness and exposure to seawater in wet weather (27). The associations estimated in this study may not reflect those of the general population, but among a highly exposed subgroup of athletes, our results measure the illness associated with seawater exposure after rainstorms in southern California. Enrolling surfers led to some important differences between the present study population and most swimmer cohorts. We enrolled adults because we could not guarantee adequate consent for minors through online enrollment, whereas swimmer cohorts have historically enrolled predominantly families with children (28); children are more susceptible and have greater risk than do adult swimmers (15). Participants surfed twice per week for 2 hours each session, with nearly universal head immersion (96% of exposures) and frequent water ingestion (38% of exposures). This far exceeds exposure levels recorded in swimmer cohorts. Likely because of surfers' repeated exposures to pathogens in seawater, studies have found higher levels of immunity to hepatitis A and more frequent gut colonization by antibiotic-resistant *Escherichia coli* among surfers than among the general population (29, 30).

Despite surfers' intense and frequent exposures, gastrointestinal illness rates observed in the present study were similar to those measured among beachgoers California cohorts in the summer (Web Appendix 6, Web Figure 7), and the increase in gastrointestinal illness rates associated with seawater exposure (adjusted IRR = 1.33, 95% CI: 0.99, 1.78; Web Table 2) was similar to estimates measured in marine swimmer cohorts in California and elsewhere in the United States (15, 31). However, the 3-fold increase in rates of

Table 3. Incidence Rate Ratios for Surfer Illnesses Within 3 Days of Dry- and Wet-Weather Seawater Exposure Compared With Unexposed Periods, San Diego, California, 2013–2015

Outcome	Unadjusted[a]				Adjusted[a,b]			
	Dry Weather		Wet Weather[c]		Dry Weather		Wet Weather[c]	
	IRR	95% CI	IRR	95% CI	IRR	95% CI	IRR	95% CI
Gastrointestinal illness	1.39	1.05, 1.86	1.69	1.10, 2.59	1.30	0.95, 1.76	1.41	0.92, 2.17
Diarrhea	1.27	0.92, 1.76	1.77	1.11, 2.83	1.22	0.86, 1.73	1.51	0.95, 2.41
Sinus pain or infection	1.38	1.05, 1.80	1.64	1.12, 2.40	1.23	0.93, 1.64	1.51	1.01, 2.26
Earache or infection	2.06	1.47, 2.90	3.11	1.94, 4.98	1.86	1.27, 2.71	3.28	1.95, 5.51
Infection of open wound	2.35	1.27, 4.36	3.89	1.83, 8.30	3.04	1.54, 5.98	4.96	2.18, 11.29
Skin rash	1.72	1.16, 2.54	1.78	0.98, 3.24	1.64	1.11, 2.41	1.80	0.97, 3.35
Fever	1.45	0.99, 2.12	0.57	0.24, 1.31	1.56	1.04, 2.34	0.64	0.27, 1.52
Upper respiratory illness[d]	1.03	0.79, 1.35	1.25	0.84, 1.86	1.04	0.79, 1.36	1.17	0.79, 1.74
Any infectious symptom[e]	1.44	1.14, 1.82	1.68	1.19, 2.38	1.50	1.17, 1.92	1.62	1.14, 2.30

Abbreviations: CI, confidence interval; IRR, incidence rate ratio.

[a] Unadjusted and adjusted incidence rate ratios compare incidence rates in the 3 days after seawater exposure during dry or wet weather with incidence rates during unexposed periods. Table 2 includes the underlying data. Tests of trend in the IRR between exposure categories are significant ($P < 0.05$) if the confidence interval for wet-weather exposure excludes 1.0 (22).

[b] We controlled for the following time-invariant potential confounders: age, sex, educational level, employment status, household income, years the individual had surfed, reported behavior of typically avoiding the ocean after wet weather, surfboard length, mode of enrollment (beach vs. Internet). We controlled for chronic health conditions only for the corresponding outcomes: ear problems, sinus problems, gastrointestinal conditions, respiratory conditions, skin conditions. We also controlled for the following time-varying potential confounders: entered the ocean for an activity other than surfing, any illness symptoms in the week preceding the risk window, day of recall, day of the week, and rainfall total during the past 3 days.

[c] Defined as entering the sea within 3 days of 0.25 cm or more of rain in 24 hours.

[d] Only measured in year 2 of the study.

[e] Includes gastrointestinal illness, eye infections, infected wounds, and fever.

earache/infection and 5-fold increase in infected open wounds associated with exposure after rainstorms (Table 3) are stronger associations than have been reported in previous studies, and they provide evidence for increased incidence of a broad set of infectious symptoms after seawater exposure within 3 days of rain.

Fecal indicator bacteria were a reliable marker of human illness risk in this setting only within 3 days of rainfall (Table 4). Our results are consistent with summer studies in California in which investigators found associations between *Enterococcus* levels and illness only if there was a well-defined source of human fecal contamination (4–8). Our findings are also consistent with model predictions of higher gastrointestinal illness risk among southern California surfers after storms (32). Molecular testing for pathogens in storm water discharge to study monitored beaches identified near-ubiquitous presence of norovirus and *Campylobacter* species, and models parameterized with pathogen measurements predicted higher illness risk after rainstorms (14). The association between fecal indicator bacteria measured during wet weather and a range of nonenteric illnesses, such as sinus pain or infection and fever (Table 4), suggests that fecal indicator bacteria may mark broader bacterial or viral pathogen contamination in seawater after rainstorms.

Some study outcomes could have noninfectious causes associated with surfing. Earache and sinus pain can result from physical incursion of saltwater through surfing's high-intensity exposure, ingestion of saltwater can cause gastrointestinal symptoms, and wetsuit use could cause skin rashes. If the association between surf exposure and symptoms resulted from noninfectious causes, we would expect similar incidence rates after wet- and dry-weather exposures. This was observed for skin rash, but incidence rates for sinus, ear, and gastrointestinal illnesses were higher after wet-weather exposure (Table 2), and the strong association between fecal indicator bacteria and fever during wet-weather conditions was consistent with an infectious etiology (Table 4).

It is also possible that some infections acquired during surfing could result from nonanthropogenic sources. The ocean was warmer than usual during the second winter because of a weak El Niño, which caused conditions favorable to naturally occurring *Vibrio parahaemolyticus* and toxin-producing marine algae that can cause human illness (33). Wound infection was the single outcome strongly associated with fecal indicator bacteria measured during dry weather (Table 4), an observation consistent with a pathogen source like *V. parahaemolyticus* that covaries with fecal indicator bacteria even in nonstorm conditions. Yet, the consistently higher rates of infected wounds and other symptoms after wet-weather exposure compared with dry-weather exposure (Tables 2 and 3) suggests that storm water runoff impacted by anthropogenic sources constitutes an important pathogen source in this setting.

Table 4. Surfer Illness Associated With a $\log_{10}$ Increase in Fecal Indicator Bacteria Levels, Stratified by Exposure During Dry and Wet Weather, Tourmaline Surfing Park and Ocean Beach, San Diego, California, 2013–2015

Fecal Indicator Bacteria and Illness Symptom	Dry Weather		Wet Weather		Unadjusted					Adjusted[a]				
					Dry Weather		Wet Weather		*P* Value[b]	Dry Weather		Wet Weather		*P* Value[b]
	Episodes	Days at Risk	Episodes	Days at Risk	IRR	95% CI	IRR	95% CI		IRR	95% CI	IRR	95% CI	
Enterococcus														
Gastrointestinal illness	30	4,251	10	1,297	0.86	0.47, 1.58	2.17	1.16, 4.03	0.04	0.85	0.46, 1.56	1.75	0.80, 3.84	0.16
Diarrhea	24	4,285	9	1,305	1.13	0.62, 2.07	2.38	1.27, 4.46	0.11	1.16	0.63, 2.14	2.00	0.92, 4.32	0.31
Sinus pain or infection	44	4,130	19	1,262	1.34	0.79, 2.26	1.93	1.17, 3.19	0.33	0.96	0.53, 1.76	1.61	0.96, 2.69	0.22
Earache or infection	38	4,233	14	1,274	0.74	0.37, 1.47	1.23	0.50, 3.02	0.38	0.70	0.35, 1.40	1.32	0.51, 3.41	0.31
Infection of open wound	19	4,360	6	1,332	2.69	1.05, 6.90	2.24	0.65, 7.69	0.83	2.79	1.12, 6.95	2.94	0.79, 10.97	0.95
Skin rash	19	4,230	5	1,267	1.46	0.68, 3.14	0.89	0.21, 3.82	0.56	1.09	0.42, 2.80	0.51	0.06, 4.04	0.50
Fever	22	4,366	2	1,342	1.33	0.69, 2.56	3.29	2.35, 4.59	0.01	1.29	0.66, 2.52	3.53	2.37, 5.24	0.01
Upper respiratory illness[c]	37	3,679	15	1,090	0.89	0.55, 1.45	1.94	0.85, 4.42	0.10	0.74	0.44, 1.25	1.89	0.87, 4.11	0.06
Any infectious symptom[d]	50	4,080	17	1,264	1.12	0.69, 1.83	2.51	1.49, 4.24	0.04	1.06	0.64, 1.76	2.52	1.41, 4.50	0.03
Fecal coliforms														
Gastrointestinal illness	30	4,251	10	1,297	0.82	0.42, 1.61	2.96	1.50, 5.83	0.01	0.76	0.38, 1.54	2.59	1.02, 6.56	0.04
Diarrhea	24	4,285	9	1,305	1.04	0.53, 2.04	3.34	1.72, 6.47	0.02	1.05	0.51, 2.16	3.20	1.31, 7.85	0.08
Sinus pain or infection	44	4,130	19	1,262	1.57	0.87, 2.84	2.18	1.11, 4.26	0.48	0.75	0.35, 1.58	1.52	0.62, 3.73	0.22
Earache or infection	38	4,233	14	1,274	0.83	0.39, 1.76	1.46	0.63, 3.39	0.29	0.99	0.51, 1.92	1.59	0.84, 3.01	0.32
Infection of open wound	19	4,360	6	1,332	2.76	0.91, 8.36	2.67	0.85, 8.41	0.97	3.21	1.03, 10.03	4.12	0.95, 17.91	0.79
Skin rash	19	4,230	5	1,267	1.69	0.72, 3.99	1.03	0.24, 4.43	0.56	1.18	0.39, 3.56	0.54	0.09, 3.06	0.42
Fever	22	4,366	2	1,342	1.15	0.49, 2.70	4.99	3.19, 7.79	0.00	1.16	0.49, 2.73	6.22	3.88, 9.96	0.00
Upper respiratory illness[c]	37	3,679	15	1,090	0.97	0.50, 1.89	2.33	0.75, 7.23	0.19	0.73	0.38, 1.40	2.03	0.70, 5.89	0.11
Any infectious symptom[d]	50	4,080	17	1,264	1.17	0.69, 1.97	3.21	1.84, 5.58	0.01	1.11	0.65, 1.91	3.42	1.76, 6.66	0.01
Total coliforms														
Gastrointestinal illness	30	4,251	10	1,297	0.77	0.40, 1.47	2.62	1.63, 4.24	0.01	0.83	0.42, 1.63	1.96	1.22, 3.15	0.08
Diarrhea	24	4,285	9	1,305	0.66	0.29, 1.51	2.59	1.53, 4.38	0.02	0.78	0.35, 1.70	1.99	1.19, 3.35	0.09
Sinus pain or infection	44	4,130	19	1,262	1.52	0.84, 2.77	2.02	1.04, 3.93	0.55	1.08	0.54, 2.19	1.79	0.93, 3.44	0.33
Earache or infection	38	4,233	14	1,274	1.03	0.54, 1.96	1.67	0.63, 4.41	0.40	0.92	0.46, 1.82	1.72	0.64, 4.61	0.32
Infection of open wound	19	4,360	6	1,332	3.46	0.79,15.20	2.16	0.46, 10.16	0.69	4.02	0.91, 17.67	2.38	0.60, 9.43	0.63
Skin rash	19	4,230	5	1,267	1.58	0.73, 3.40	1.14	0.34, 3.81	0.65	1.30	0.48, 3.53	1.11	0.28, 4.41	0.86
Fever	22	4,366	2	1,342	1.59	0.78, 3.22	7.48	4.28, 13.08	0.00	1.62	0.77, 3.37	9.24	4.64, 18.41	0.00
Upper respiratory illness[a]	37	3,679	15	1,090	0.87	0.49, 1.52	2.04	0.84, 4.96	0.12	0.72	0.40, 1.30	1.87	0.84, 4.19	0.08
Any infectious symptom[d]	50	4,080	17	1,264	1.35	0.78, 2.34	3.26	1.76, 6.01	0.06	0.69	0.23, 2.07	3.02	1.56, 5.38	0.10

Abbreviations: CI, confidence interval; IRR, incidence rate ratio.

[a] We controlled for the following time-invariant potential confounders: age, sex, educational level, employment status, household income, years the individual had surfed, reported behavior of typically avoiding the ocean after wet weather, surfboard length, mode of enrollment (beach vs. Internet). We controlled for chronic health conditions only for the corresponding outcomes: ear problems, sinus problems, gastrointestinal conditions, respiratory conditions, skin conditions. We also controlled for the following time-varying potential confounders: entered the ocean for an activity other than surfing, any illness symptoms in the week preceding the risk window, day of recall, day of the week, and rainfall total during the past 3 days.

[b] *P* value for multiplicative effect modification of dry versus wet weather.

[c] Only measured in year 2 of the study.

[d] Includes gastrointestinal illness, eye infections, infected wounds, and fever.

Limitations

The use of self-reported symptoms could bias the association between seawater exposure and illness away from the null if surfers overreported illness after exposure; conversely, random (nondifferential) errors in exposures or outcomes could bias associations toward the null (34). The survey measured daily exposure and outcomes in separate modules—an intentional decision to separate the measurements and inhibit systematic reporting bias. Adjusted analyses controlled for day of recall and day of the week to reduce nondifferential bias from recall errors but would not control for systematic bias. Negative control exposure analyses found no association between *Enterococcus* levels and illness on days with no recent water exposure (Web Table 5), which suggests that unmeasured confounding or reporting bias is unlikely to explain the association between *Enterococcus* levels and illness. Moreover, the use of daily average levels of fecal indicator bacteria could bias the association between water quality and illness toward the null if the averaging resulted in nondifferential misclassification error (35).

We measured incident outcomes within 3 days of seawater exposure because the population regularly entered the ocean, a 3-day period captures the incubation period for the most common waterborne pathogens (e.g., norovirus, *Campylobacter* species, *Salmonella* species) (36), and past studies found that most excess episodes of gastrointestinal illness associated with seawater exposure occurred in the first 1–2 days (15). Illness caused by waterborne pathogens with longer incubation periods (e.g., *Cryptosporidium* species) (37) could have been misclassified in this study, which could bias results toward the null by artificially increasing incidence rates in unexposed periods and decreasing rates in exposed periods.

Conclusions

Surfing was associated with increased incidence of several categories of symptoms, and associations were stronger if surfing took place shortly after rainstorms. Higher levels of fecal indicator bacteria were strongly associated with fever, sinus pain/infection, wound infection, and gastrointestinal symptoms within 3 days of rainstorms. The internal consistency between water-quality measurements, patterns of illness after dry- and wet-weather exposures, and incidence profiles with time since rainstorms lead us to conclude that seawater exposure during or close to rainstorms at beaches impacted by urban runoff in southern California increases the incidence rates of a broad set of acute illnesses among surfers. These findings provide strong evidence to support the posting of beach warnings after rainstorms and initiatives that would reduce pathogen sources in urban runoff that flows to coastal waters.

ACKNOWLEDGMENTS

Author affiliations: Division of Epidemiology, School of Public Health, University of California, Berkeley, California (Benjamin F. Arnold, Ayse Ercumen, Jade Benjamin-Chung, John M. Colford, Jr.); Southern California Coastal Water Research Project, Costa Mesa, California (Kenneth C. Schiff, Joshua A. Steele, John F. Griffith, Steven J. Steinberg, Paul Smith, Stephen B. Weisberg); Orange County Sanitation District, Fountain Valley, California (Charles D. McGee; retired); and Surfrider Foundation, San Clemente, California (Richard Wilson, Chad Nelsen).

The study was funded by the city and county of San Diego, California.

We thank the field team members who enrolled participants at the beach and collected water samples throughout the study. We also thank Laila Othman, Sonji Romero, Aaron Russell, Joseph Toctocan, Laralyn Asato, Zaira Valdez, and the staff at City of San Diego Marine Microbiology Laboratory who generously provided laboratory space to test water specimens, and Jeffrey Soller, Mary Schoen, and members of the study's external advisory committee for earlier comments on the results.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest: none declared.

REFERENCES

1. Noble RT, Weisberg SB, Leecaster MK, et al. Storm effects on regional beach water quality along the southern California shoreline. *J Water Health*. 2003;1(1):23–31.
2. Leecaster MK, Weisberg SB. Effect of sampling frequency on shoreline microbiology assessments. *Mar Pollut Bull*. 2001; 42(11):1150–1154.
3. Ackerman D, Weisberg SB. Relationship between rainfall and beach bacterial concentrations on Santa Monica bay beaches. *J Water Health*. 2003;1(2):85–89.
4. Haile RW, Witte JS, Gold M, et al. The health effects of swimming in ocean water contaminated by storm drain runoff. *Epidemiology*. 1999;10(4):355–363.
5. Colford JM, Wade TJ, Schiff KC, et al. Water quality indicators and the risk of illness at beaches with nonpoint sources of fecal contamination. *Epidemiology*. 2007;18(1): 27–35.
6. Colford JM, Schiff KC, Griffith JF, et al. Using rapid indicators for Enterococcus to assess the risk of illness after exposure to urban runoff contaminated marine water. *Water Res*. 2012; 46(7):2176–2186.
7. Arnold BF, Schiff KC, Griffith JF, et al. Swimmer illness associated with marine water exposure and water quality indicators: impact of widely used assumptions. *Epidemiology*. 2013;24(6):845–853.
8. Yau VM, Schiff KC, Arnold BF, et al. Effect of submarine groundwater discharge on bacterial indicators and swimmer health at Avalon Beach, CA, USA. *Water Res*. 2014;59:23–36.
9. Taylor K. Contagion Present. Surfer Magazine. http://www.surfermag.com/features/contagion-present. Published July 20, 2016. Accessed August 17, 2016.
10. Dwight RH, Baker DB, Semenza JC, et al. Health effects associated with recreational coastal water use: urban versus rural California. *Am J Public Health*. 2004;94(4):565–567.
11. Harding AK, Stone DL, Cardenas A, et al. Risk behaviors and self-reported illnesses among Pacific Northwest surfers. *J Water Health*. 2015;13(1):230–242.

12. Stormsurf. Weather basics. http://www.stormsurf.com/page2/tutorials/weatherbasics.shtml. Published September 26, 2003. Accessed October 27, 2016.
13. Heal the Bay. Heal the Bay's 2014-2015 Annual Beach Report Card. Santa Monica, CA: Heal the Bay; 2015. http://www.healthebay.org/sites/default/files/BRC_2015_final.pdf. Accessed December 5, 2016.
14. Schiff K, Griffith J, Steele J, et al. The Surfer Health Study: A Three-Year Study Examining Illness Rates Associated With Surfing During Wet Weather. Costa Mesa, CA: Southern California Coastal Water Research Project; 2016. http://ftp.sccwrp.org/pub/download/DOCUMENTS/TechnicalReports/943_SurferHealthStudy.pdf. Published September 20, 2016. Accessed December 5, 2016.
15. Arnold BF, Wade TJ, Benjamin-Chung J, et al. Acute gastroenteritis and recreational water: highest burden among young US children. *Am J Public Health*. 2016;106(9):1690–1697.
16. Wade TJ, Sams E, Brenner KP, et al. Rapidly measured indicators of recreational water quality and swimming-associated illness at marine beaches: a prospective cohort study. *Environ Health*. 2010;9:66.
17. Colford JM, Wade TJ, Sandhu SK, et al. A randomized, controlled trial of in-home drinking water intervention to reduce gastrointestinal illness. *Am J Epidemiol*. 2005;161(5):472–482.
18. Colford JM, Hilton JF, Wright CC, et al. The Sonoma Water Evaluation Trial: a randomized drinking water intervention trial to reduce gastrointestinal illness in older adults. *Am J Public Health*. 2009;99(11):1988–1995.
19. Arnold B, Ercumen A. The Surfer Health Study. Open Science Framework. https://osf.io/hvn7s. Published July 29, 2015. Updated July 29, 2016. Accessed December 5, 2016.
20. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2008.
21. Zou G. A modified poisson regression approach to prospective studies with binary data. *Am J Epidemiol*. 2004;159(7):702–706.
22. Vittinghoff E, Glidden DV, Shiboski SC, et al. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2nd ed. New York, NY: Springer Science & Business Media; 2012.
23. United States Environmental Protection Agency. *Recreational Water Quality Criteria*. Washington, DC: United States Environmental Protection Agency; 2012. (Office of Water publication no. 820-F-12-058). https://www.epa.gov/sites/production/files/2015-10/documents/rwqc2012.pdf. Accessed January 24, 2017.
24. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. *Biometrics*. 2011;67(4):1406–1413.
25. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*. 2010;21(3):383–388.
26. Arnold BF, Ercumen A, Benjamin-Chung J, et al. Brief report: negative controls to detect selection bias and measurement bias in epidemiologic studies. *Epidemiology*. 2016;27(5):637–641.
27. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol*. 2013;42(4):1012–1014.
28. Wade TJ, Pai N, Eisenberg JN, et al. Do US Environmental Protection Agency water quality guidelines for recreational waters prevent gastrointestinal illness? A systematic review and meta-analysis. *Environ Health Perspect*. 2003;111(8):1102–1109.
29. Gammie A, Morris R, Wyn-Jones AP. Antibodies in crevicular fluid: an epidemiological tool for investigation of waterborne disease. *Epidemiol Infect*. 2002;128(2):245–249.
30. Leonard A. *Are Bacteria in the Coastal Zone a Threat to Human Health?* [dissertation]. Exeter, UK: University of Exeter; 2016. https://ore.exeter.ac.uk/repository/handle/10871/22805. Accessed October 14, 2016.
31. Fleisher JM, Fleming LE, Solo-Gabriele HM, et al. The BEACHES Study: health effects and exposures from non-point source microbial contaminants in subtropical recreational marine waters. *Int J Epidemiol*. 2010;39(5):1291–1298.
32. Tseng LY, Jiang SC. Comparison of recreational health risks associated with surfing and swimming in dry weather and post-storm conditions at Southern California beaches using quantitative microbial risk assessment (QMRA). *Mar Pollut Bull*. 2012;64(5):912–918.
33. Van Dolah FM. Marine algal toxins: origins, health effects, and their increased occurrence. *Environ Health Perspect*. 2000;108(suppl 1):133–141.
34. Copeland KT, Checkoway H, McMichael AJ, et al. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol*. 1977;105(5):488–495.
35. Fleisher JM. The effects of measurement error on previously reported mathematical relationships between indicator organism density and swimming-associated illness: a quantitative estimate of the resulting bias. *Int J Epidemiol*. 1990;19(4):1100–1106.
36. Widdowson MA, Sulka A, Bulens SN, et al. Norovirus and foodborne disease, United States, 1991-2000. *Emerg Infect Dis*. 2005;11(1):95.
37. Jokipii L, Jokipii AM. Timing of symptoms and oocyst excretion in human cryptosporidiosis. *N Engl J Med*. 1986;315(26):1643–1647.

Implications of WASH Benefits trials for water and sanitation

Authors' reply

We appreciate the thoughtful comments from Oliver Cumming and Val Curtis and from Diane Coffey and Dean Spears regarding the Kenya and Bangladesh WASH Benefits trials.[1,2] Since both trials took place in populations with relatively low levels of open defecation at enrolment, we agree that the global health community should be cautious about transporting effect estimates from the trials to populations with high levels of open defecation, or to populations in urban environments with vastly different conditions. These trials were done in populations that were similar to much of rural Bangladesh and Kenya, and were chosen explicitly because of the high burden of both linear growth faltering and diarrhoea.[3] We anticipate that the results will be generalisable to many similar rural populations with persistent growth faltering. Forthcoming trial results from Zimbabwe[4] and Mozambique[5] will complement the WASH Benefits trials by providing evidence of the effect of water, sanitation, and handwashing (WASH) interventions on growth in populations with high baseline levels of open defecation, and—in the case of Mozambique—in a high-density, urban setting.

The letters propose that WASH interventions could have possibly improved child growth had we fielded the trials in populations with higher levels of open defecation or in populations with worse drinking water sources. Yet, linear growth faltering prevailed: average length-for-age z-scores (LAZ) in control groups at the final endpoint were –1·54 in Kenya and –1·79 in Bangladesh. Low average LAZ despite low levels of open defecation and access to improved water sources for the majority of people in both populations show that prenatal or postnatal exposures, or both, beyond open defecation and water source are important determinants of linear growth faltering.[6] In Kenya, water supply remained a challenge because few participants had piped water to their homes, but in Bangladesh tube wells within household compounds were ubiquitous, suggesting that adequate water supply alone will be insufficient to prevent growth faltering.

Both letters suggest that a more comprehensive, community-level approach to improving the environment might be necessary to influence child growth. In theory, this is certainly possible, but the trials delivered compound-level interventions because formative research in rural Bangladesh and sub-Saharan Africa showed that, among children younger than 18 months, exposure to faecal contamination occurs primarily within the compound.[7,8] Despite delivering intensive compound-level WASH interventions, it remains possible that the trials did not reduce faecal exposure among children enrolled in the study sufficiently to influence growth through the hypothesised subclinical pathways,[9] despite improving many other outcomes. High diarrhoea prevalence in the Kenya trial[2] and widespread enteric pathogen infection in Bangladesh[10] and Kenya (Pickering AJ, Tufts University, personal communication) reflect high levels of transmission. Environmental measurements in the Bangladesh trial documented widespread faecal contamination that was strongly associated with the presence of animals and their faeces.[11] Forthcoming results from both trials will summarise intervention effects on enteric pathogens and on faecal contamination throughout the children's environment, including complementary foods.

Well designed and conducted randomised trials answer specific

Published Online
April 26, 2018
http://dx.doi.org/10.1016/S2214-109X(18)30229-8

	Population (n)	Mean LAZ (SD)	Difference (95% CI)	p value	Adjusted* difference (95% CI)	p value
Kenya trial control group†						
No improved latrine	1737	–1·58 (1·08)	ref		ref	
Access to improved latrine	364	–1·33 (1·08)	0·25 (0·12–0·37)	<0·001	0·15 (0·02–0·28)	0·02
Bangladesh trial control group†						
No latrine	513	–1·89 (0·98)	ref		ref	
Latrine with no water seal	391	–1·86 (1·00)	..		..	
Latrine has functional water seal	199	–1·37 (1·01)	0·52 (0·34–0·70)	<0·001	0·22 (0·03–0·40)	0·02

Median age 25 months for Kenya trial and 22 months for Bangladesh trial. LAZ=length-for-age z-scores. *Adjusted by use of ensemble machine learning with double-robust, targeted maximum likelihood estimation following the same methods from the prespecified adjusted analyses in the trials. Prespecified, baseline covariates included: child age, child sex, household food insecurity, birth order, maternal age, maternal education, maternal height, number of children and total individuals living in the compound, distance to water, and a broad set of household characteristics and assets. The computational notebook that created the table includes additional analysis details, plus adjusted effects using generalised linear models that resulted in similar estimates (https://osf.io/qkgp8). Data used to make the table are available on the Open Science Framework website for Bangladesh (https://osf.io/wvyn4) and Kenya (https://osf.io/uept9). †In the Kenya trial, improved sanitation was defined as the presence of a latrine with a slab following the standard WHO/UNICEF Joint Monitoring Program definition. In the Bangladesh trial, improved sanitation was defined as a toilet with a functional water seal. These definitions mirrored those reported in the original trials.

Table: **LAZ among children in the control groups of the WASH Benefits trials in Kenya and Bangladesh, stratified by whether the child's household had improved sanitation at enrolment**

For more on the **UN Sustainable Development Goals** see https://www.un.org/sustainabledevelopment/

questions with high validity—a feature that is at once valuable and limiting. It will never be possible to do randomised trials in every setting, and fielding a randomised trial that delivers even more intensive environmental interventions than WASH Benefits to entire communities rather than compounds would probably be logistically and financially prohibitive. Observational analyses could potentially help fill the evidence gap.

Yet, a re-analysis of the trials leads us to urge the global community to be cautious when interpreting observational analyses of the effects of sanitation on child growth, similar to those presented by Coffey and Spears. Inspired by an analysis that the SHINE investigators[4] presented at the American Society for Tropical Medicine and Hygiene 2017 conference, we re-analysed data from the WASH Benefits trials to estimate the difference in LAZ associated with improved sanitation access at enrolment among children born into the control group—creating an observational, prospective cohort nested within each trial. Among children in the control group, improved sanitation was associated with 0·15 LAZ increase in Kenya (p=0·02) and 0·22 LAZ increase in Bangladesh (p=0·02) in adjusted, double-robust analyses (table). The inconsistency between the observational analyses and null effects in the trials, estimated in the same study populations, illustrates the danger of bias from unmeasured confounding in observational studies, which has been shown in many other examples.[12] It also calls into question whether the observed associations between sanitation conditions and linear growth in India are causal. Sanitation facilities and open defecation practices are inextricably tied to many improvements in overall wellbeing. This cautionary example highlights the value of randomised trials for measuring the effects of exposure–outcome relationships that are deeply entwined with broader socioeconomic development. Nevertheless, we feel strongly that these findings should not diminish ongoing, ambitious efforts to achieve the UN Sustainable Development Goals (SDGs): myriad health, equity, and ethical arguments motivate elimination of open defecation and ample supply of microbiologically safe water, even in the absence of a strong link to child growth.

We declare no completing interests.

*Benjamin F Arnold, Clair Null, Stephen P Luby, John M Colford Jr
benarnold@berkeley.edu

Division of Epidemiology & Biostatistics, School of Public Health, University of California, Berkeley, CA 94720, USA (BFA, JMC); Center for International Policy Research and Evaluation, Mathematica Policy Research, Washington, DC, USA (CN); and Division of Infectious Diseases and Geographic Medicine, Stanford University, Stanford, CA, USA (SPL)

1 Luby SP, Rahman M, Arnold BF, et al. Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural Bangladesh: a cluster randomised controlled trial. *Lancet Glob Health* 2018; **6:** e302–15.

2 Null C, Stewart CP, Pickering AJ, et al. Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural Kenya: a cluster-randomised controlled trial. *Lancet Glob Health* 2018; **6:** e316–29.

3 Arnold BF, Null C, Luby SP, et al. Cluster-randomised controlled trials of individual and combined water, sanitation, hygiene and nutritional interventions in rural Bangladesh and Kenya: the WASH Benefits study design and rationale. *BMJ Open* 2013; **3:** e003476.

4 Sanitation Hygiene Infant Nutrition Efficacy (SHINE) Trial Team. The Sanitation Hygiene Infant Nutrition Efficacy (SHINE) trial: rationale, design, and methods. *Clin Infect Dis* 2015; **61** (suppl 7)**:** S685–702.

5 Brown J, Cumming O, Bartram J, et al. A controlled, before-and-after trial of an urban sanitation intervention to reduce enteric infections in children: research protocol for the Maputo Sanitation (MapSan) study, Mozambique. *BMJ Open* 2015; **5:** e008215.

6 Black RE, Victora CG, Walker SP, et al. Maternal and child undernutrition and overweight in low-income and middle-income countries. *Lancet* 2013; **382:** 427–51.

7 Kwong LH, Ercumen A, Pickering AJ, Unicomb L, Davis J, Luby SP. Hand- and object-mouthing of rural Bangladeshi children 3–18 months old. *Int J Environ Res Public Health* 2016; **13:** 563.

8 Mbuya MNN, Tavengwa NV, Stoltzfus RJ, et al. Design of an intervention to minimize ingestion of fecal microbes by young children in rural Zimbabwe. *Clin Infect Dis* 2015; **61:** S703–09.

9 Humphrey JH. Child undernutrition, tropical enteropathy, toilets, and handwashing. *Lancet* 2009; **374:** 1032–35.

10 Lin A, Ercumen A, Benjamin-Chung J, et al. Effects of water, sanitation, handwashing, and nutritional interventions on child enteric protozoan infections in rural Bangladesh: a cluster-randomized controlled trial. *Clin Infect Dis* 2018; published online April 13. DOI:10.1093/cid/ciy320.

11 Ercumen A, Pickering AJ, Kwong LH, et al. Animal feces contribute to domestic fecal contamination: evidence from *E. coli* measured in water, hands, food, flies, and soil in bangladesh. *Environ Sci Technol* 2017; **51:** 8725–34.

12 Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? *N Engl J Med* 2000; **342:** 1907–09.

Risk Factors for Menstrual Toxic Shock Syndrome: Results of a Multistate Case-Control Study

Arthur L. Reingold,* Claire V. Broome, Suzanne Gaventa, Allen W. Hightower, and the Toxic Shock Syndrome Study Group†

From the Meningitis and Special Pathogens Branch and the Statistical Services Activity, Division of Bacterial Diseases, Center for Infectious Diseases, Centers for Disease Control, Atlanta, Georgia

For assessment of current risk factors for developing toxic shock syndrome (TSS) during menstruation, a case-control study was performed. Cases with onset between 1 January 1986 and 30 June 1987 were ascertained in six study areas with active surveillance for TSS. Age-matched controls were selected from among each patient's friends and women with the same telephone exchange. Of 118 eligible patients, 108 were enrolled, as were 185 "friend controls" and 187 telephone exchange-matched controls. Tampon use was a risk factor for developing TSS during menstruation (odds ratio = 29; 95% confidence interval = 7–120), and risk increased with increasing tampon absorbency (odds ratio = 1.34 per gram increase in absorbency; 95% confidence interval = 1.2–1.6). The role of tampon chemical composition could not be assessed because the number of cases was inadequate. Neither use of birth control pills for contraception nor use of medications for premenstrual or menstrual symptoms protected against or was a risk factor for the development of menstrual TSS.

Case-control studies conducted in the early 1980s demonstrated that tampon use was the major risk factor for the development of toxic shock syndrome (TSS) during menstruation and that risk varied with the brand and style of tampon used [1–6]. One of these studies further demonstrated that a tampon's absorbency and/or chemical composition was important in determining the risk associated with its use, although the relative importance of these two tampon characteristics remained uncertain [3]. Subsequent in vitro studies have suggested that the chemical composition of tampons may be the major determinant of risk because of differences in the binding of magnesium and hence in the production of TSS toxin 1 [7–9]. However, a recent assessment of cases reported through a passive national-surveillance system suggests that both absorbency and chemical composition are important independent determinants of the risk of menstrual TSS [10].

In response to these findings and in an effort to minimize or eliminate the risk of menstrual TSS, manufacturers have both substantially altered the chemical composition and dramatically lowered the absorbency of the tampons they sell. As a result, the tampons that are available and being used today differ markedly from those in use in the early 1980s. In order to evaluate the risk of menstrual TSS associated with currently available tampons and to shed more light on the relative importance of tampon absorbency and chemical composition in determining that risk, we undertook a case-control study of menstrual TSS cases occurring in 1986–1987.

Methods

Patients with TSS and age-matched controls were sought in six study areas (Los Angeles County and the states of Missouri, New Jersey, Oklahoma, Tennessee and Washington) where active surveillance for TSS had been established. Details of the active surveillance methods used are presented elsewhere [11]. In brief, educational materials concerning TSS and

This study was supported by an interagency agreement of the Centers for Disease Control, the National Institute of Child Health and Human Development, and the U.S. Food and Drug Administration.

* Present address: Department of Biomedical and Environmental Health Sciences, School of Public Health, University of California at Berkeley, Berkeley, California 94720.

† The Toxic Shock Syndrome Study Group includes S. Waterman and C. Hoppe (Los Angeles County); M. Spurrier and S. Sitze (Missouri); R. McCready, D. Cundiff, and M. Farrell (New Jersey); G. Istre and S. Makintubee (Oklahoma); L. Lefkowitz and J. Taylor (Tennessee); W. Lafferty and J. Harwell (Washington); Drs. M. Donawa and C. Gaffey (U.S. Food and Drug Administration); and Drs. J. Perlman and P. Wolf (National Institute of Child Health and Human Development).

Please address requests for reprints to the Meningitis and Special Pathogens Branch, Division of Bacterial Diseases, Center for Infectious Diseases, Centers for Disease Control, Atlanta, Georgia 30333.

a request for reports of all suspected cases were distributed repeatedly to health care providers, infection control nurses, and medical records departments in the study areas. These materials stressed that TSS occurs in a variety of settings in patients of both sexes and all ages. Active surveillance for patients hospitalized with TSS was maintained by biweekly telephone calls to all hospitals in the study areas to ascertain the presence or absence of suspected cases.

All suspected cases in women 10–54 years of age with onset between 1 January 1986 and 30 June 1987 were assessed with regard to the case definition for TSS established by the Centers for Disease Control [12]. Cases meeting all of the criteria were considered definite cases, those lacking a single criterion were considered probable cases, and those lacking two or more criteria or having evidence of another cause of illness were considered not to be cases. All medical records were reviewed a second time by an individual blinded to the menstrual status and tampon use history of the patient. The few minor discrepancies in classification of cases were resolved by a second person blinded to menstrual status and tampon use history. Probable and definite cases with onset of symptoms during menstruation (i.e., during active bleeding) were eligible for inclusion in the study unless a focal site of infection outside the vagina was identified or a barrier contraceptive was used during the menstrual period.

For each patient who agreed to participate, two friends matched for age (± 3 years if <25 years of age; ± 5 years if ≥ 25 years of age) and two women matched for age and neighborhood of residence were sought as controls. Controls matched for neighborhood of residence were sought by taking the first five digits of the patient's phone number and randomly ordering the 99 other possible phone numbers with the same first five digits. These households matched by telephone exchange (and hence by neighborhood of residence) were called until two age-matched women were enrolled. Women with TSS and controls were interviewed by telephone concerning use of tampons and other catamenial products on each day of the menstrual period, use of medications for menstrual and premenstrual symptoms on each day for the 3 days before onset of menstruation and during menstruation, and use of contraceptives. Patients with TSS were asked about the menstrual period when they became ill (index menstrual period) and the preceding menstrual period; controls were asked about the two menstrual periods that coincided in time with those of the respective case. While the interviewer was aware of the study hypotheses, she was blinded to the case/control status of participants at the time of the interviews. Tampon-using study participants were asked to find the box of tampons used during the most recent menstrual period and answer questions about its labeling and color.

Results were analyzed with conditional multivariate logistic regression models that took the matching into account [13]. Information concerning the chemical composition, oxygen content, and in vivo and in vitro absorbency of various tampon brands and styles was obtained from tampon manufacturers.

Results

Altogether, 118 patients with TSS were eligible for enrollment in the study, and 108 of these patients were enrolled. Reasons for which patients were not enrolled included refusal (two patients) and loss to follow-up or inability to locate (eight patients). None of the 118 patients died. Of the 108 patients enrolled, 71 were classified as having definite and 37 as having probable TSS. Among the 37 probable cases, fever of $\geq 102°F$ was the criterion most often lacking (15 cases); desquamation was lacking in 14 cases, multisystem involvement in four, and hypotension in four. The characteristic rash of TSS was present in all probable cases. Onset of illness occurred most often on the third or fourth day of the menstrual cycle (day 1, 9%; day 2, 14%; day 3, 17%; day 4, 29%; day 5, 12%; day 6, 13%, day 7, 2%; and day 8, 4%).

Altogether, 372 age-matched controls were enrolled, including 185 friends of patients and 187 neighborhood residents. Four controls were enrolled for each of 71 cases (66%), three controls for each of 15 cases (14%), two controls for each of 21 cases (19%), and only one control for one case (1%). As expected, the patients and controls were similar in age, race, and marital status (table 1). "Friend controls" were somewhat more similar to patients than were "neighborhood controls" with regard to race and marital status, but these differences were not significant.

Of the 108 women with TSS, 106 (98%) were using tampons at the time of onset of illness; 88 women had been using a single brand and style of tampon during that menstrual period, whereas 18 had been using multiple brands and/or styles (table 2). Of the 372 control women, 244 (66%) had used tampons

Table 1. Characteristics of patients and controls enrolled in a multistate study of risk factors for menstrual toxic shock syndrome.

Characteristic (unit)	Value for indicated group			
	Patients	Friend controls	Neighborhood controls	Combined controls
Mean age (y)*	24.3 ± 8.1 (13–46)	24.8 ± 8.4 (11–48)	24.5 ± 8.1 (13–48)	24.6 ± 8.2 (11–48)
White (%)	94	94	89	91
Married (%)	44	39	36	37
Interval from onset of index menstrual period to interview (d)*	88 ± 50 (25–249)	. . .	. . .	87 ± 51 (17–281)
Interviews successfully completed with blinding to case/control status (%)	82	91	87	89

* Values given are mean ± SD (range).

during their index menstrual period. Friend controls were more likely to have used tampons than were neighborhood controls (71% vs. 60%; odds ratio = 1.7; 95% confidence interval = 1.02–2.7; two-tailed $P = .04$, conditional logistic regression). Altogether, 44% of tampon-using patients and 62% of tampon-using controls were able during their telephone interview to find the box of tampons used.

Tampon use was associated with an increased risk of developing TSS during menstruation, regardless of which control group was used as a basis for comparison (friends, neighbors, or combined; table 3). Women who used multiple brands and/or styles were at greater risk than women who used a single brand and style (odds ratio = 2.3; 95% confidence interval = 1.2–4.6; $P = .02$). However, this difference was due to the fact that users of multiple brands and/or styles tended to use more absorbent tampons. With control for absorbency, there was no difference in risk between users of a single brand and users of multiple brands and/or styles.

Because there were overall no significant differences between friend and neighborhood controls regarding the brand or style of tampon used, these control groups were combined in studies of the risk of menstrual TSS associated with individual brands and individual brand/style combinations. The use of all major tampon brands was associated with an increased risk of developing TSS during menstruation, with odds ratios for individual brands ranging from 15 to 59 (table 4). Odds ratios for individual styles of each tampon brand were calculated in two ways; in comparison with the risk of TSS in women not using tampons and in comparison with the risk of TSS in users of Tampax Original Regular tampons. In comparison with women using no tampons, users of all assessed individual brands and styles (except Tampax Slender Regular and Tampax Original Regu-

Table 2. Tampon use during the index menstrual period.

Pattern of tampon use	No. (%) in indicated group with pattern of use							
	Patients		Friend controls		Neighborhood controls		Combined controls	
None	2 (2)		54 (29)		74 (40)		128 (34)	
Single brand and style	88 (81)		115 (63)		104 (56)		219 (59)	
Multiple brands and/or styles	18 (17)	98	15 (8)	71*	7 (4)	60	22 (6)	66
Unknown brand	. . .		1 (<1)		2 (1)		3 (1)	
Total	108		185		187		372	

* Significant difference between friend and neighborhood controls (odds ratio = 1.7; 95% confidence interval = 1.02–2.7; two-tailed = $P = .04$).

Table 3. Association between tampon use and risk of menstrual toxic shock syndrome.

Tampon use	Odds ratio*/95% confidence interval for patients vs. indicated control group		
	Friend	Neighborhood	Combined
Any tampon	19/5-78	48/7-362	29/7-120
Single brand and style	. . .	. . .	27/7-111
Multiple brand and/or style	. . .	. . .	62/13-291

* Vs. no tampon use.

lar) were at increased risk of menstrual TSS (table 5). In comparison with users of Tampax Original Regular tampons, users of some but not all other brand/style combinations were demonstrated to be at increased risk.

We next analyzed risk of menstrual TSS as a function of various tampon characteristics, including measured in vitro and in vivo absorbency, weight, oxygen content, and chemical composition. There was a significant association between measured in vitro tampon absorbency and risk of menstrual TSS: the risk increased by 34% for every 1-g increase in absorbency (odds ratio per gram increase = 1.34; 95% confidence interval = 1.2-1.6). Tampon weight and in vivo absorbency were equally good predictors of the risk of menstrual TSS, while oxygen content correlated somewhat less well. After taking in vitro absorbency into account, we could detect no influence of oxygen content or of chemical compositon (categorized either as the presence or absence of a given material or as the percentage comparison by weight) on the risk of menstrual TSS.

Analysis of tampon users revealed that patterns of tampon use differed between patients and controls (table 6). Tampon-using women with TSS used tampons on more days of the menstrual cycle, were more likely to use tampons continuously for at least 1 day, used tampons continuously on more days and on a higher percentage of days of the menstrual cycle, and left a single tampon in place for a longer mean maximum time. Patients and controls were similar, however, in the average number of tampons used per day and the total number of tampons used per menstrual period. Because many of these characteristics of tampon use were correlated with the absorbency of the tampon used, we also examined their effect on the risk of menstrual TSS after adjustment for absorbency. Using tampons continuously on at least 1 day of the menstrual cycle remained strongly correlated with the risk of menstrual TSS after adjustment for absorbency (odds ratio = 6.5; 95% confidence interval = 2.5-17.2). Once absorbency and continuous use of tampons were taken into account, none of the other tampon-use variables remained significantly associated with risk of menstrual TSS.

Neither increased nor decreased risk of menstrual TSS in association with the use of birth control pills or barrier contraception was found (table 7). Use of condoms for contraception was commoner, however, among women with TSS (odds ratio = 2.6; 95% confidence interval = 1.1-6.1). The use of medications for premenstrual and menstrual syndromes was not associated with either an increased or a decreased risk of developing TSS, whether examined by individual brand, by active ingredient, or by overall use/nonuse (table 8).

Table 4. Association between tampon brand and risk of menstrual toxic shock syndrome.

Tampon brand*	No. using brand in indicated group		Matched odds ratio	95% confidence interval
	Patients	Combined controls		
None	2	128	1	. . .
Tampax	23	128	15	3-64
OB	9	15	56	9-330
Playtex	46	63	59	13-265
Kotex	10	12	54	10-302
Other	0	1	0	. . .
Total	90	347		

* Single brand and style use only.

Table 5. Risk of menstrual toxic shock syndrome among users of selected individual tampon brands and styles.

Brand and style of tampon	No. (%) using brand/style in indicated group		Odds ratio/95% confidence interval vs. indicated category	
	Patients	Controls	No tampon use	Use of Tampax Original Regular
No tampon	. . .	. . .	1/. . .	. . .
Tampax Original Regular	2 (2)	39 (18)	7/0.8–58	1/. . .
Tampax Slender Regular	4 (5)	27 (13)	6/1–35	0.98/0.1–8
Tampax Petal Soft Regular	2 (2)	11 (5)	22/2–212	3.2/0.4–30
Tampax Super	9 (11)	38 (18)	26/4–149	3.7/0.6–22
Tampax Super Plus	3 (4)	13 (6)	25/3–207	3.8/0.5–30
OB Regular	3 (4)	9 (4)	28/3–268	4.2/0.5–38
OB Super	4 (5)	5 (2)	86/9–862	13/1.4–122
OB Super Plus	2 (2)	1 (<1)	144/7–2,857	22/1.1–422
Playtex Slender Regular (D/ND)*	4 (5)	5 (2)	78/8–789	11/1.2–110
Playtex Regular (D/ND)	20 (24)	27 (13)	76/13–441	13/2.4–66
Playtex Super (D/ND)	16 (19)	25 (12)	74/13–429	11/2–58
Playtex Super Plus (D/ND)	6 (7)	6 (3)	79/10–612	12/1.6–83
Kotex Security Regular	2 (2)	6 (3)	21/1.7–253	2.9/0.2–40
Kotex Security Super	7 (8)	4 (2)	122/15–971	18/2.5–133

* Deodorant and nondeodorant, combined.

Discussion

The results presented here suggest that, despite marked changes in the absorbency and chemical composition of tampons in recent years, the use of many if not all tampons available in 1986–1987 is associated with an increased risk of menstrual TSS. Furthermore, while the measured absorbency of tampons has been reduced dramatically, there continues to be a direct correlation between measured tampon absorbency and risk of menstrual TSS. Continuous use of tampons on at least 1 day of the menstrual cycle appears to increase a tampon user's risk of developing TSS, as has been noted previously [5]. We were unable to confirm the results of earlier studies that suggested a protective effect of oral contraceptive pills with regard to menstrual TSS [14].

Table 6. Univariate analyses of patterns of tampon use among toxic shock syndrome patients and controls who used tampons.

Variable	Mean ± SD for indicated group		Odds ratio	95% confidence interval
	Patients (n = 106)	Controls (n = 244)		
Mean average no. of tampons used per day	4.7 ± 4.1	4.3 ± 2.3	1.04/tampon	0.97–1.13
Mean total no. of tampons used per menstrual period	21.9 ± 21.6	18.3 ± 12.2	1.02/tampon	1.0–1.03
Mean no. of days on which tampons were used	4.5 ± 1.6	4.2 ± 1.5	1.22/day of use	1.03–1.44
Mean no. of days on which tampons were used continuously	4.0 ± 2.1	2.3 ± 2.3	1.46/day of continuous use	1.27–1.67
Mean percentage of days on which tampons were used continuously	83.8 ± 8	52.9 ± 47	1.02/percentage of days	1.01–1.03
Mean maximum time a single tampon was left in place (hours)	7.8 ± 2.1	6.6 ± 2.4	1.46/hour	1.21–1.75
Any day(s) of continuous tampon use	95 (90)*	141 (58)*	9.4	3.9–22.3

* Values indicate number (percentage) of women.

Table 7. Use of contraceptives and risk of toxic shock syndrome.

Type of contraception	No. (%) using method in indicated group: Patients (n = 108)	Controls (n = 372)	Matched odds ratio	95% confidence interval
Condoms	10 (9)	15 (4)	2.6	1.1–6.1
Birth control pills	27 (25)	89 (24)	1.1	0.6–1.8
Any barrier contraception*	3 (3)	19 (5)	0.6	0.2–2.1
Diaphragm*	2 (2)	16 (4)	0.5	0.1–2.1
Contraceptive sponge*	1 (1)	2 (<1)	. . .	. . .
Any spermicide	6 (6)	22 (6)	. . .	. . .
Intrauterine device	2 (2)	7 (2)	. . .	. . .
Tubal ligation	6 (6)	31 (8)	. . .	. . .
Hysterectomy	1 (1)	1 (<1)	. . .	. . .
Rhythm	2 (2)	0	. . .	. . .
Withdrawal	2 (2)	1 (<1)	. . .	. . .
Cervical cap*	0	1 (<1)	. . .	. . .

* All cases of menstrual and nonmenstrual toxic shock syndrome associated with the use of a diaphragm, contraceptive sponge, or cervical cap were excluded from this study.

The magnitude of the risk associated with tampon use in our study remains somewhat ill defined because of the different frequencies of tampon use observed among the two types of controls enrolled. Thus, depending on whether friend or neighborhood controls were used as the standard for comparison, the estimate of the risk varied between 19 and 48. While combining of the two control groups for this particular comparison is not valid because of their heterogeneity, it is likely that the resultant estimate of the frequency of tampon use among control women (66%) would yield a more accurate estimate of the risk associated with tampon use (odds ratio = 29) than does an analysis of either control group alone. Data from national surveys conducted in 1985 suggest that ~65% of women with menstrual periods use tampons [10].

Table 8. Use of medications for premenstrual and menstrual symptoms and risk of toxic shock syndrome.

Medication	No. (%) taking medication in indicated group: Patients (n = 108)	Controls (n = 372)	Odds ratio	95% confidence interval
Any	40 (37)	138 (37)	1.0	0.7–1.6
Midol	4 (4)	18 (5)	0.7	0.2–2.2
Aspirin	5 (5)	22 (6)	0.8	0.3–2.3
Tylenol	10 (9)	32 (9)	1.1	0.5–2.4
Motrin	3 (3)	14 (4)	0.7	0.2–2.6
Advil	7 (6)	13 (3)	2.1	0.7–6.1
Nuprin	0 (0)	8 (2)	. . .	. . .
Pamprin	4 (4)	12 (3)	1.1	0.3–3.6
Premesyn	3 (3)	2 (1)	5.0	0.8–30
Other	10 (9)	31 (8)	. . .	. . .

Two limitations to this study warrant discussion in an assessment of the results. First, it is possible that, despite all of our educational efforts and publicity, medical care providers were more likely to diagnose and/or report a case of menstrual TSS if the patient was a tampon user. Bias of this type would have resulted in overestimation of the risk associated with tampon use vs. no tampon use. We currently are reviewing ~12,000 medical records for all women 10–54 years of age who were discharged from hospitals in the study areas in 1986 with TSS or diagnoses likely to be confused with TSS in an effort to determine how many of these women had TSS that was undiagnosed and/or unreported. By ascertaining the menstrual status and pattern of tampon use for women with TSS that was unreported and/or misdiagnosed, we hope to assess the impact of diagnostic and reporting biases on our results. It should be noted, however, that these biases would not have affected our analysis of the risk associated with use of individual brands and styles of tampons vs. use of Tampax Original Regular tampons. Similarly, these biases would not have affected our analysis of the relation between measured tampon absorbency or tampon use patterns and risk of menstrual TSS.

The second limitation is the paucity of cases available for study. Because of the small number of cases studied, the confidence intervals around our point estimates are very wide; that is, our estimates of var-

ious risks are imprecise. Furthermore, despite our efforts, there are insufficient cases to permit a meaningful assessment of the independent contributions of tampon absorbency, chemical composition, and other characteristics to the risk of menstrual TSS. Thus, it remains possible that one or more tampon characteristics other than measured in vitro absorbency could play an important role in determining the risk of menstrual TSS. Given the enormous effort and the size of the surveillance population required for the collection of the cases studied here, it seems unlikely that a prospective study that is based on active surveillance and is large enough to answer questions about the impact of tampon characteristics will be feasible.

While the observed incidence of nonmenstrual TSS in the study areas was approximately that predicted on the basis of findings from earlier studies, the incidence of menstrual TSS was substantially lower than that predicted from data gathered in other states during previous years [11]. Thus, while incidence rates in the range of 5–15 cases/100,000 menstruating women per year were observed in Wisconsin, Minnesota, Utah, and Colorado in 1980, the incidence rate of menstrual TSS observed in our six study areas in 1986 ranged between 1 and 2.5/100,000 menstruating women. Whether the incidence of menstrual TSS we observed was lower than expected because the incidence has dropped in recent years, because the areas under study always had lower incidences, because cases now are being recognized and treated earlier, or because other unknown factors are involved is unclear. However, even if the incidence of menstrual TSS has decreased in recent years, our data suggest that there is still a need for a uniform standard of tampon labeling with regard to measured absorbency.

References

1. Davis JP, Chesney PJ, Wand PJ, LaVenture M, the Investigation and Laboratory Team. Toxic-shock syndrome: epidemiologic features, recurrence, risk factors, and prevention. N Engl J Med 1980;**303**:1429–35
2. Helgerson SD, Foster LR. Toxic shock syndrome in Oregon: epidemiologic findings. Ann Intern Med 1982;**96**(Part 2):909–11
3. Osterholm MT, Davis JP, Gibson RW, Mandel JS, Wintermeyer LA, Helms CM, Forfang JC, Rondeau J, Vergeront JM, and the Investigation Team. Tri-state toxic-shock syndrome study. I. Epidemiologic findings. J Infect Dis 1982;**145**:431–40
4. Schlech WF III, Shands KN, Reingold AL, Dan BB, Schmid GP, Hargrett NT, Hightower A, Herwaldt LA, Neill MA, Band JD, Bennett JV. Risk factors for the development of toxic shock syndrome: association with a tampon brand. JAMA 1982;**248**:835–9
5. Shands KN, Schmid GP, Dan BB, Blum D, Guidotti RI, Hargrett NT, Anderson RL, Hill DL, Broome CV, Band JD, Fraser DW. Toxic-shock syndrome in menstruating women: its association with tampon use and *Staphylococcus aureus* and the clinical features in 52 cases. N Engl J Med 1980;**303**:1436–42
6. Kehrberg MW, Latham RH, Haslam BR, Hightower A, Tanner M, Jacobson JA, Barbour AG, Noble V, Smith CB. Risk factors for staphylococcal toxic-shock syndrome. Am J Epidemiol 1981;**114**:873–9
7. Kass EH, Kendrick MI, Tsai Y-C, Parsonnet J. Interaction of magnesium ion, oxygen tension, and temperature in the production of toxic-shock-syndrome toxin-1 by *Staphylococcus aureus*. J Infect Dis 1987;**155**:812–5
8. Mills JT, Parsonnet J, Kass EH. Production of toxic-shock-syndrome toxin-1: effect of magnesium ion [letter]. J Infect Dis 1986;**153**:993–4
9. Mills JT, Parsonnet J, Tsai Y-C, Kendrick M, Hickman RK, Kass EH. Control of production of toxic-shock-syndrome toxin-1 (TSST-1) by magnesium ion. J Infect Dis 1985; **151**:1158–61
10. Berkley SF, Hightower AW, Broome CV, Reingold AL. The relationship of tampon characteristics to menstrual toxic shock syndrome. JAMA 1987;**258**:917–20
11. Gaventa S, Reingold AL, Hightower AW, Broome CV, Schwartz B, Hoppe C, Harwell J, Lefkowitz LK, Mackintubee S, Cundiff D, Sitze S, the Toxic Shock Syndrome Study Group. Active surveillance for toxic shock syndrome in the United States, 1986. Rev Infect Dis 1989;**11**(Suppl 1):S28–34
12. Reingold AL, Hargrett NT, Shands KN, Dan BB, Schmid GP, Strickland BY, Broome CV. Toxic shock syndrome surveillance in the United States, 1980 to 1981. Ann Intern Med 1982;**92**:875–80
13. Breslow NE, Day NE. Statistical methods in cancer research. Lyon: International Agency for Research on Cancer, 1980
14. Shelton JD, Higgins JE. Contraception and toxic-shock syndrome: a reanalysis. Contraception 1981;**24**(6):631–4

Discussion

DR. EDWARD KASS. Dr. Reingold, I find it difficult to match your second conclusion with your data. The only data that show a clear relation are those dealing with polyacrylate rayon. All of the rest are not statistically significant. Now, the same thing was true in the Tri-State Study. I do not understand how you can say there is a linear relation between risk and absorbency if all of the excess statistically significant cases occur in relation to only one fiber. This is particularly important because, as you know, there is a question of national policy. There is a question of labeling absorbency. Representations have been made to the U.S. Food and Drug Administration. I find it difficult to make national policy recommen-

dations based on data that seem to me not secure, and, by your own statement, the numbers other than those dealing with polyacrylate rayon are not secure.

DR. ARTHUR REINGOLD. This study was done in 1986–1987, and none of these tampons contained polyacrylate rayon. Polyacrylate rayon was removed from Playtex tampons in the spring of 1985. Therefore, we are not able to look at the risk associated with polyacrylate in these data. I am the first to admit that the numbers here are very sparse. The question of whether there is any increased risk associated with various brands and styles compared with no tampon use depends on how many cases of TSS in non-tampon-using women went undiagnosed. We hope to get at least some assessment of that through this enormous chart review. To the extent that there has been a lot of diagnostic bias and those cases have been missed, it is possible that the increased risk in comparison to non-tampon use is, in fact, erroneous. The real problem then comes in terms of comparing other tampons with the Tampax Original Regular in that we have few cases relative to what we would like to have. I am, in fact, somewhat pleased that we were able to find so few cases because it indicates to me that we have been going in the right direction in the last few years and that this disease has really decreased in incidence. On the other hand, it makes for difficulties in interpreting the results of the study.

DR. JAMES TODD. I hope your conclusion is correct. As you say, you will only know whether the incidence has decreased once you have ascertained your reporting bias and what effect it has on your statistics. Certainly, your data from California do not suggest that the incidence has decreased significantly in that area. To speculate a bit, let us assume that there is a direct risk associated with absorbency. It has been said that this risk is not a function of leaving tampons in longer, although from seeing cases clinically I am convinced that it is. My own experience suggests that the severity of illness seems to relate directly to how long the tampon was left in. What are the data to convince us that the increase in absorbency in tampons is not directly related to an increase in the length of time that the tampon is left in?

DR. REINGOLD. The data are not good. In this study we did look at the number of tampons used per day (as the best indicator we could come up with because we were interviewing between 1 and 2 months after the illness), and there is not a substantial difference between the patients and the controls, which is what has been found in similar case-control studies. As to the other point you raise, I do not understand the biologic way in which absorbency could affect risk. We have looked at the data, substituting oxygen content because there is some correlation between oxygen content and absorbency, and if anything, oxygen content is not as good a predictor of risk as absorbency. The weight of the tampon is as good an indicator as absorbency, but again, they are too closely correlated to be separable. I do not know what it is that measured absorbency is telling us or what it indicates.

DR. KASS. The most convincing data came from the Tri-State Study, which reported that if there was any kind of cross-over between length of time a tampon is worn and risk, it was at ~13 hours, and the effect was negligible. From that fairly large study, it did not appear that length of time was a great variable in rate of disease. Whether that has changed since then, I do not know. We have all seen cases of the kind that Dr. Todd mentioned, but I think that the length of time a tampon is kept in place has not been statistically significant in relation to risk.

Second, with respect to the point about oxygen, as you know, we published a paper on the effect of oxygen on toxin production, and, except at conditions of zero oxygen, there is toxin production, particularly when magnesium levels are low. I agree that it is unlikely that variation in oxygen is going to be a major significant variable if some oxygen is present.

Third, I hope people will keep in mind that most cotton-containing tampons, whether all cotton or partially cotton, have adherent magnesium that is not covalently linked. Cotton itself has no free carboxyl groups. Therefore, any salts that are in the cotton tampon are simply there as contaminants during the manufacturing process. The salts leach out easily, and the salt content varies immensely from batch to batch. Cotton-containing tampons will usually release magnesium and therefore counteract any other tendency toward increased toxin production, and this becomes an important variable in looking at the effect of different products. Unless each product is carefully examined to see how much this particular variable changes from product to product—and I can assure you it changes immensely from batch to batch—you will get peculiar and variable results, and this adds to the underlying argument that we are talking of a surrogate and not of absorbency itself.