PHW250B Week 10 Reader

Topic 1: Confounding

Lecture: Confounding in depth. 2

Topic 2: Using DAGs to Identify Confounding

Lecture: Using DAGs to Detect Confounding. 25

Pearl et al. Causal Inference in Statistics: A Primer. Wiley 2016. Sections 2.4-2.5, 3.3. 46

Topic 3: Negative Controls

Lecture: Negative Controls. 134

Journal Club

Arnold et al. (2017). 144

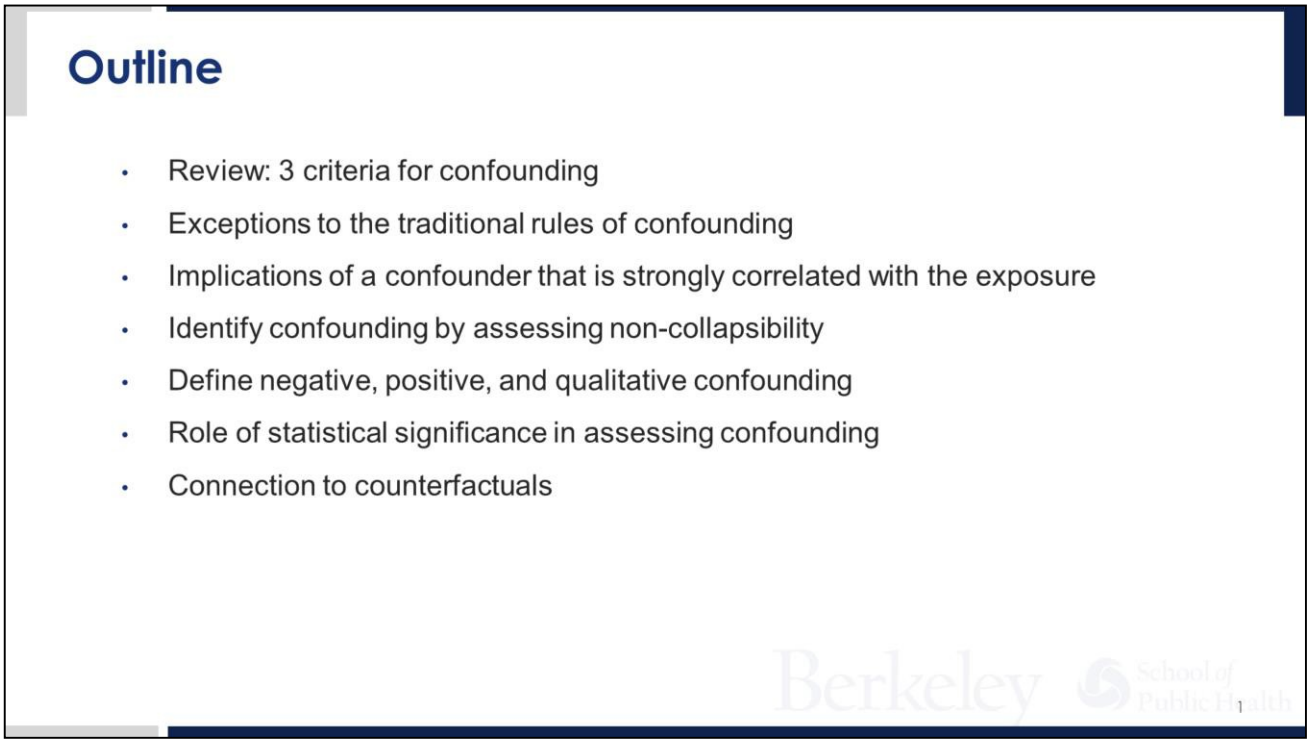
Arnold et al. (2017) - Web Appendix. 154

Arnold et al. (2018). 182

Reingold et al. (1989). 184

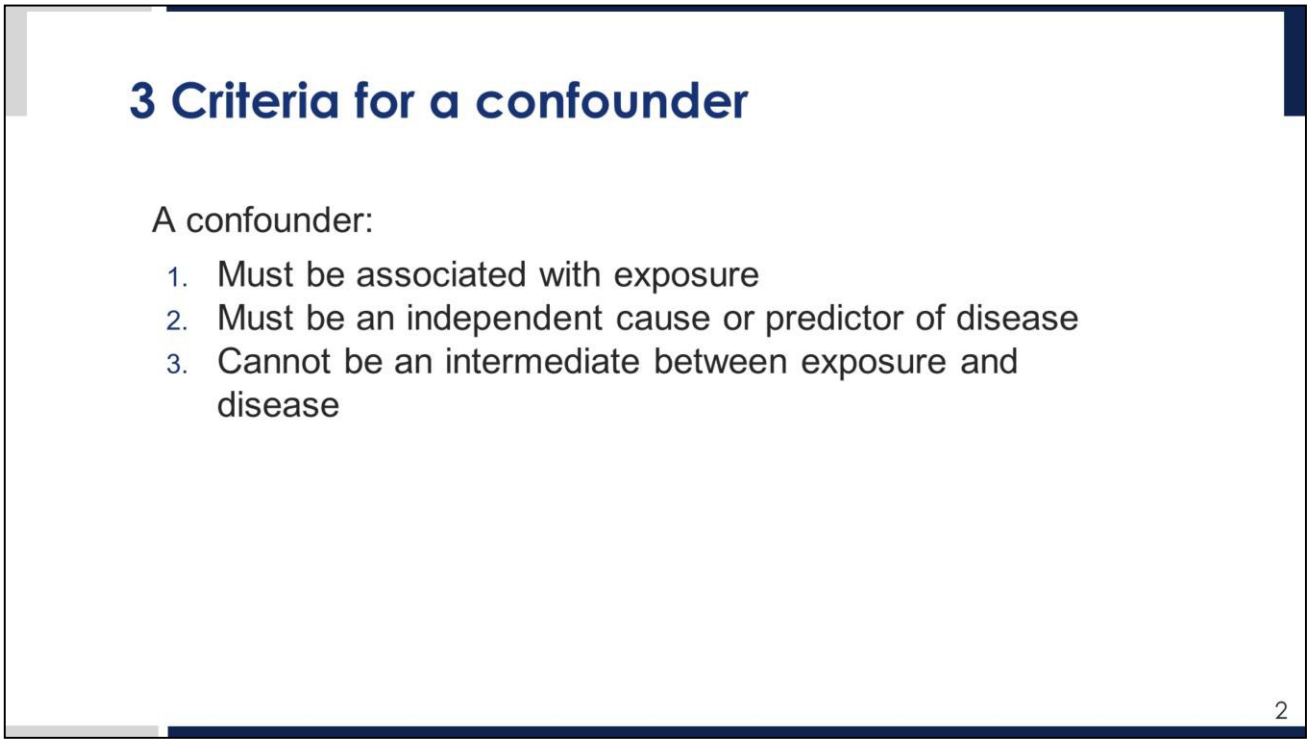
Lecture: Confounding in depth

PRESENTER: Today, we're going to discuss in some depth the concept of confounding in epidemiology.



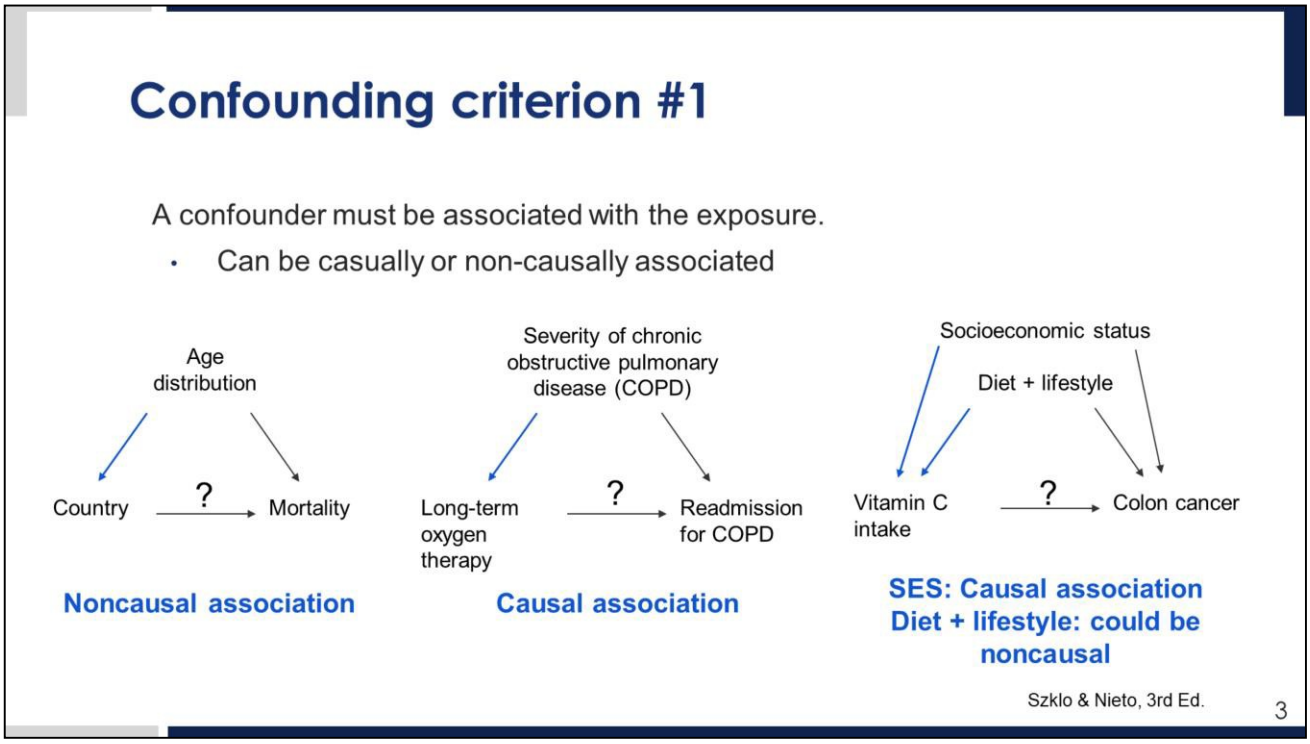
The way this talk will be structured is first, I'll talk about three traditional or classical criteria to define confounding, but then examine some exceptions to those traditional rules. We'll look at the situation as kind of a special situation where the confounder is really strongly correlated with the exposure variable. And then we'll identify confounding by assessing non-collapsibility of strata.

Next, we'll talk about what the difference is between negative and positive confounding, and then sort of a qualitative confounding as kind of a difference in direction. Following that, we'll talk about the role for using statistical testing in assessing confounding or not. It's not always recommended to do that. And then we'll make a link to counterfactuals and the counterfactual framework as we think about confounding.



So first, let's start off with three criteria that are classically given for a confounder. For decades, this is how confounding instruction was taught and sort of limited to this. So a confounder was defined as a variable that had to be associated with an exposure. It had to be an independent clause or predictor of disease. Or you might think of that as outcome.

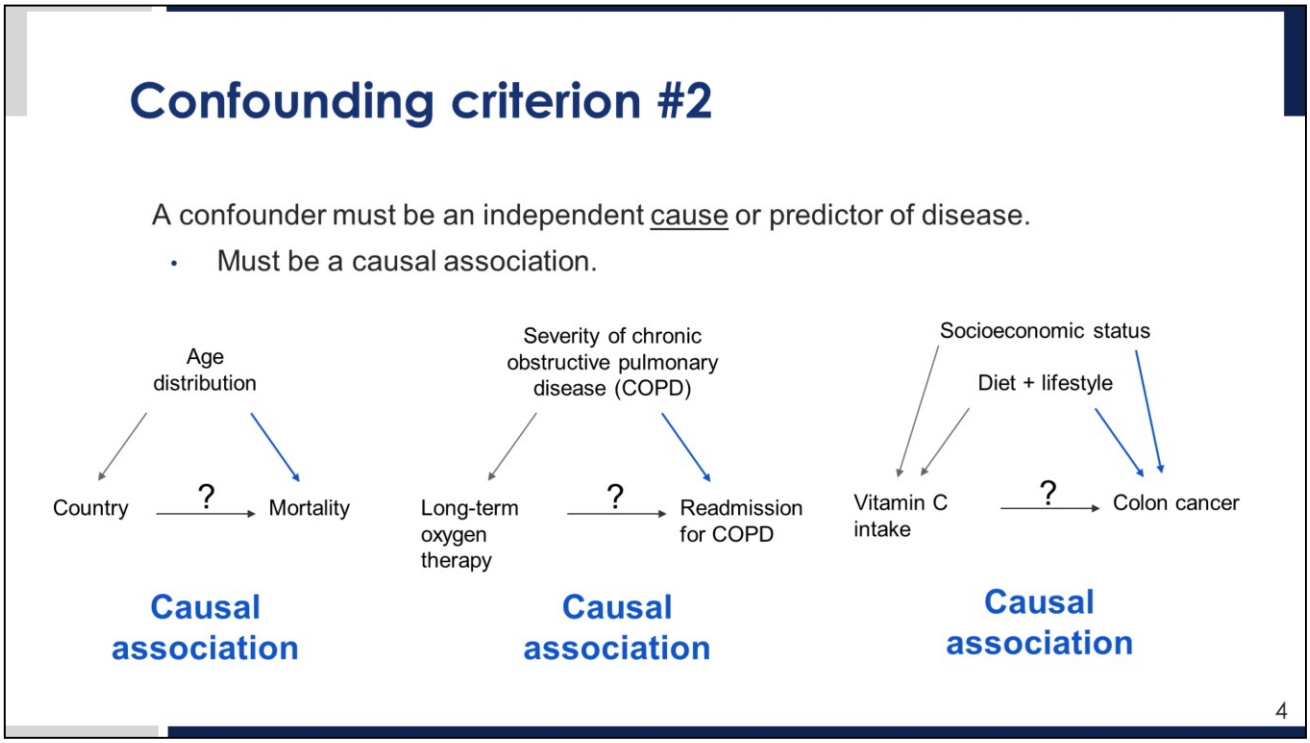
Because remember, we're always looking for the relationship between exposure and outcome. And we are now trying to evaluate whether another variable might be a confounder of that relationship. And finally, it was thought or believed that a confounder cannot be an intermediate between exposure and disease. So we'll go through each of these classical rules and then talk about exceptions to them.



So the first confounding criterion is this concept that the confounder must be associated with the exposure. Let's get oriented. Look at the three structured graphs here where we see the relationship.

And the first one is we're looking at country as the exposure, and mortality is the outcome. And age distribution is being evaluated as a confounder. So for it to be a confounder, it was believed that age distribution had to be causally or non-causally associated with country. Of course, that would be a non-causal association because the age distribution wouldn't really be a causal factor in being a country.

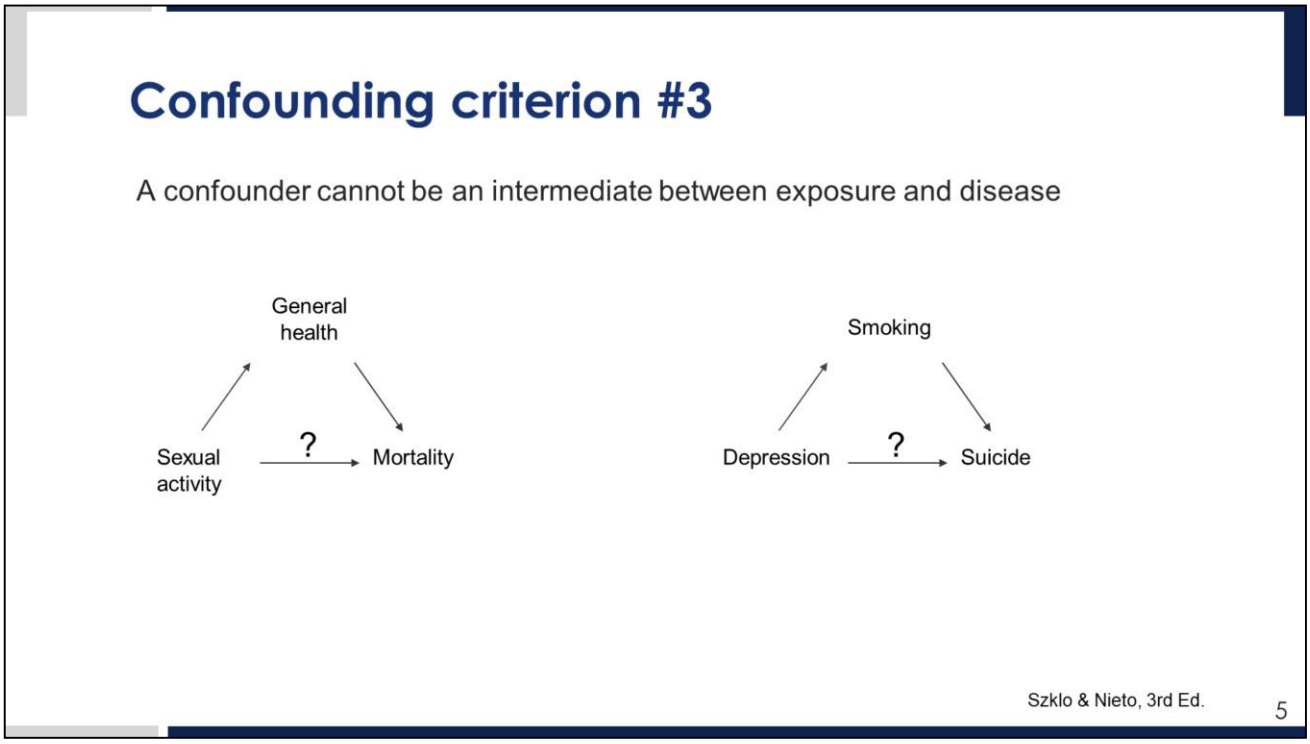
Then let's look at the middle example where there is a causal association between the potential confounder. And in this case, the potential confounder is the severity of a particular condition, called COPD. And when we evaluate it here, the idea is that it has a causal association with long-term oxygen therapy or exposure of interest.



And finally, in this third and more complicated example, we're looking at the relationship between vitamin C intake and colon cancer. And it's believed the SES as a confounder has a causal association with vitamin C intake. And separately, diet and lifestyle might have a non-causal relationship. But both of them could be confounders because they have either a causal or a non-causal association with the exposure of interest.

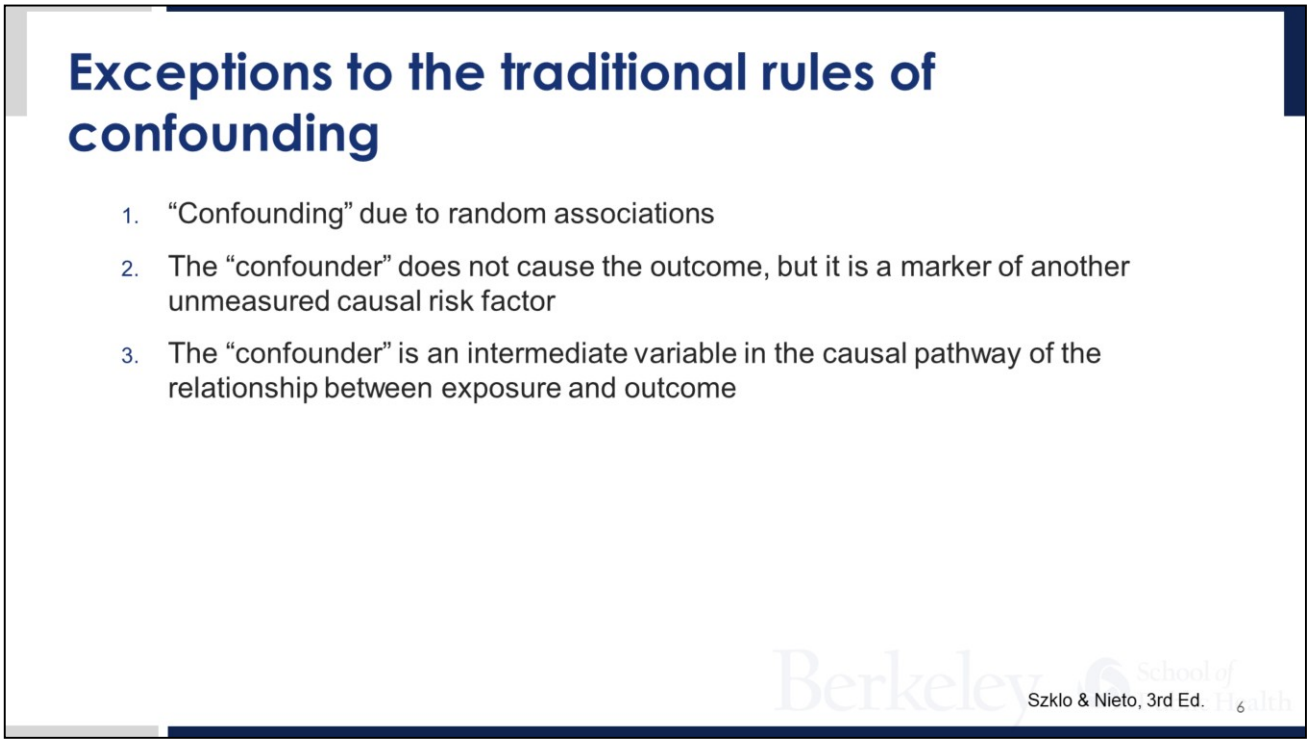
The second confounding criterion was that a confounder had to be an independent cause or predictor of disease, that is, of the outcome. So in the same three figures, the idea is to look at whether the potential confounder, age distribution, COPD, or socioeconomic status, and diet and lifestyle, what's their relationship to the outcome? And the first example, the relationship between age distribution and mortality, it's believed to have a causal relationship. The age does affect, very causally, mortality.

COPD is also felt to have a causal relationship with readmission for COPD. And similarly, it's believed, from other work and other literature, that diet and lifestyle and also socioeconomic status have a causal relationship with colon cancer. But, again, the potential confounder, in this case, age distribution, COPD, socioeconomic status, or diet and lifestyle, the criteria is that they have either a causal relationship or they predict disease and have an association with disease.

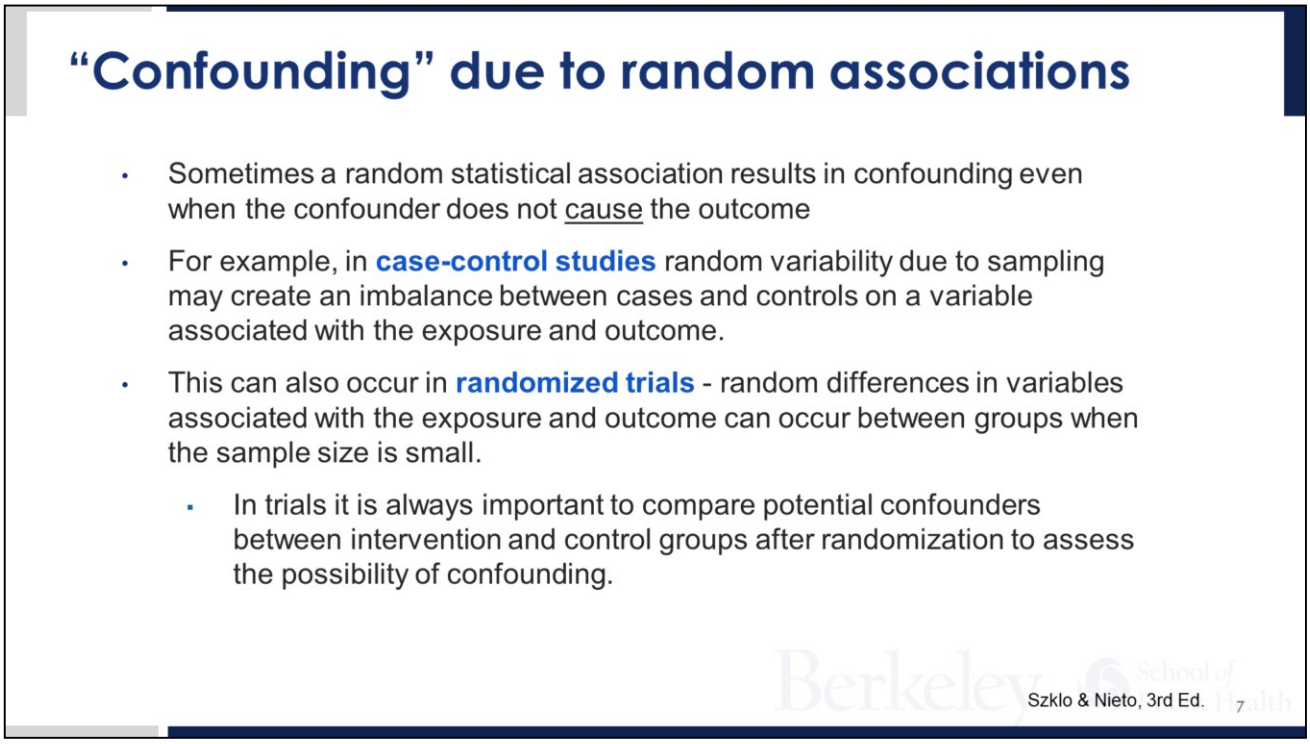


Finally, the third classical criterion for a confounder was that it cannot, should not be an intermediate between exposure and disease. So if you believe that the relationship between sexual activity and mortality flowed through one path with sexual activity influence general health, which influence mortality, that would put general health on a pathway between sexual activity and mortality. So in the classical definition, general health, in this situation, would not be treated as a confounder.

And another example-- it may not be realistic-- but if you believe that depression led to smoking and smoking then led to suicide, this relationship expressed in this diag would suggest that smoking is on the pathway between depression and suicide. So it would not be treated as a confounder in the classical sense.



Now, there are certainly exceptions to the traditional rules of confounding. And three that we'll talk about would be a situation where the confounding that you observe might just be due to random associations and variables. The second is that the potential confounder doesn't cause the outcome, but it's a marker of some other unmeasured causal risk factor that is causal for the outcome. And finally, the third is that this confounder is an intermediate variable in the causal pathway of the relationship between exposure and disease. So we'll talk about situations where even though classically, you don't want a confounder to be a variable that's intermediate in the causal pathway, there are exceptions where this is treated as a confounder. We'll talk about those.

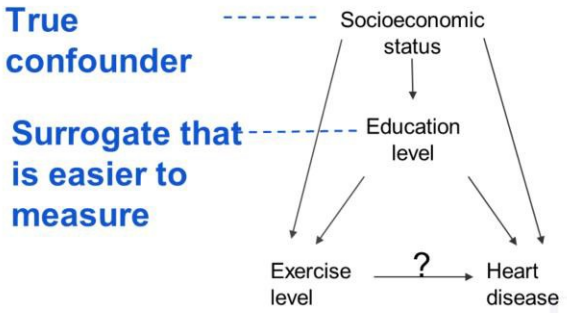


So the first is this situation of, basically, inducing confounding by random associations of variables. So sometimes in your studies, a random statistical association will arise or result in confounding, even when the confounder does not cause the outcome. So, for example, in case control studies, random variability due to sampling may create an imbalance between cases and controls on a variable that's associated with the exposure and outcome. So this is just induced by randomness.

Similarly, in randomized trials, there can be random differences. Even in randomized trials you can have random differences in the variables associated with exposure and outcome that are found in each of the groups, the exposed and the non-exposed group. And this can potentially happen more often when the sample size is small. So in trials, it's always important to compare potential confounders between intervention and control groups after randomization to assess the possibility of confounding.

The "confounder" does not cause the outcome, but it is a marker of another unmeasured causal risk factor

- A "confounder" may be a surrogate for a true confounder.
- Education level and gender are often surrogates of confounders.
- (More on this in the video on DAGs and confounding.)



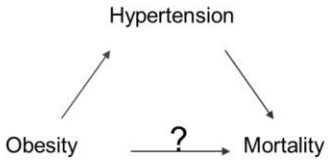
Szklo & Nieto, 3rd Ed. 8

Another situation described as the confounder not causing the outcome, but it is a marker of another unmeasured or causal factor. So a confounder that we're interested in studying might be a surrogate for a true confounder. A possible confounder might be a surrogate for a true confounder.

For instance, education level and gender are often surrogates of confounders. And that's true because it's so much easier to measure education and gender. So in this particular example, we're looking at the relationship between exercise and heart disease, that exposure-outcome relationship. So the surrogate that we measure is education level. But the true confounder might be SES, Socioeconomic Status. And education level can often serve as a very good surrogate, at times, for SES.

The "confounder" is an intermediate variable

- There are some uncommon cases when it is acceptable to define an intermediate as a confounder.
- The Szklo & Nieto textbook discusses how you can control for an intermediate if interested in the effect of an exposure on the outcome other than through that pathway.
  - Example: If you're interested in the effect of obesity on mortality through pathways other than hypertension, you could control for hypertension.
- More recent epidemiologic articles have argued against this. This is related to a more advanced topic of direct vs. indirect effects.
- For this course, it's important to know that controlling for an intermediate is something that should be done with caution and careful thought because it affects the quantity you are estimating.



Szklo & Nieto, 3rd Ed. 9

The third exception is the situation where we said that generally, a confounders should not be an intermediate variable on a pathway between an exposure and outcome. But here's an exception. So there are some uncommon cases when it's acceptable to define an intermediate as a confounder, even though that's classically not what we want to do. And your Szklo and Nieto textbook discusses how you can control for an intermediate variable if you're interested in the effect of an exposure on the outcome other than through that pathway.

So even though there is that pathway from obesity to hypertension to mortality, which would usually mean hypertension shouldn't be a confounder, if we control for hypertension-- that is, we block that path-- then there are situations where we want to do that because we're interested in the direct effect of obesity on mortality. So some recent epi-articles, are you against doing this? And this is a kind of more advanced topic dealing with direct versus indirect effects. Think of direct and indirect as how we get from obesity to mortality, directly or through some other variable. That's indirect.

So for this course, it's important to know that controlling for an intermediate is something that you do with caution and careful thought, because it affects the quantity you are estimating. And you need to know what you're doing and why you're choosing to do that.

Implications of a confounder that is strongly correlated with the exposure

- A confounding variable that is strongly correlated with the exposure of interest may be difficult to adjust for
  - This is called **collinearity.**
- **Example:** the exposure is air pollution and the confounder is area of residence.
  - It would be difficult and potentially impossible to control for area of residence because there is no variation left once controlling for it
- Ideally, there is variation so that when stratifying the confounder and the exposure (and outcome), all cells in a 2x2 table have data

	Confounder present	Confounder absent
Exposure present	A	(no data)
Exposure absent	(no data)	D

	Areas near freeway	Areas far from freeway
High level air pollution	A	(no data)
Low level air pollution	(no data)	D

Szklo & Nieto, 3rd Ed. 10

Now, there are implications when a confounder is strongly correlated with the exposure. So remember, we talked about confounder has to be causally or non-causally associated with an exposure. But this can go too far. And this is a situation called collinearity, when an exposure and a confounder are so tightly linked that there's no data left in the other situation. Let me do this with a specific example.

Let's say we're looking at the relationship between high levels of air pollution and disease. So we might want to look at distance from a freeway as a potential confounding variable. So when we do that, we find that people who have high levels of air pollution all live near freeways. And people who live far from the freeway have no data. So in cell A, those are the people who live near the freeway and have high air pollution. But because of air pollution and the freeways are so tightly linked, areas far from the freeway have no high levels of air pollution.

And similarly, among the people exposed to low levels of air pollution, they all live far from the freeway. So there are no people, no data, in the areas near the freeway with low levels of air pollution. So this collinearity, this tight linking of air pollution and freeway distance, result in the fact that we have no data in a couple of critical cells there. And you can see that in the table, how that's been expressed.

So we can't control for area of residence because we don't have the data with which to do it. So when there's no variability like this, when there's this tight collinearity, we can't adjust for a confounder that so strongly correlate with an exposure. This specific example, distance to the freeway, is so tightly correlated with air pollution level that we can't adjust for distance from the freeway.

Assessing confounding by non-collapsibility of strata

- Does the crude exposure-outcome association have the same direction and similar magnitude as **confounder stratified associations**?

Exhibit 5–4 Stratified Analyses of the Association Between Gender and Malaria (from Exhibit 5–2), According to Whether Individuals Work Mainly Outdoors or Indoors

Mostly outdoor occupation

	Cases	Controls
Males	53	15
Females	10	3
Total	63	18

Odds ratio = 1.06

Mostly indoor occupation

	Cases	Controls
Males	35	53
Females	52	79
Total	87	132

Odds ratio = 1.00

Stratified ORs

The Crude OR = 1.71, which is different from the confounder stratified ORs.

The apparent increase in malaria risk associated with male gender disappeared when accounting for the location of occupation.

Szklo & Nieto, 3rd Ed. 11

Now, a common way to evaluate confounding is to look at the non-collapsibility ability of strata. So I want to talk through this example a bit. What we're asking is whether the crude measure of exposure and outcome has the same direction and similar magnitude as a confounder stratified association. So what do all these words mean?

So this is a study looking at whether gender is associated with malaria. So males and females is the gender variable. Cases and controls is the outcome variable. This is a case-control study.

And what we see in these two separate tables, we've stratified by occupation. One stratum is outdoor occupation. The other stratum is indoor occupation. So our stratification variable is, in this case, the confounder we're trying to evaluate. We're trying to see whether occupation location, outdoor versus indoor, is a confounder of the relationship between gender and malaria.

Assessing confounding by non-collapsibility of strata

- Does the crude exposure-outcome association have the same direction and similar magnitude as the **confounder adjusted associations**?
- This is the most common method
- Can compare crude estimate to Mantel-Haenszel adjusted estimate or to an estimate from an adjusted regression model.

Table 5–1 Association between medical interventions and risk of readmission to a hospital in chronic obstructive pulmonary disease (COPD) patients: estimating the proportion of risk explained by markers of COPD severity (FEV_1, PO_2, and previous admission to a hospital)

	Crude Hazard Ratio	*Adjusted Hazard Ratio**	*Excess Risk Explained by Covariates†*
Long-term oxygen therapy	2.36‡	1.38	72%
Respiratory rehabilitation	1.77‡	1.28	64%
Anticholinergics	3.52‡	2.10‡	56%
Under the care of pulmonologist¶	2.16‡	1.73‡	37%

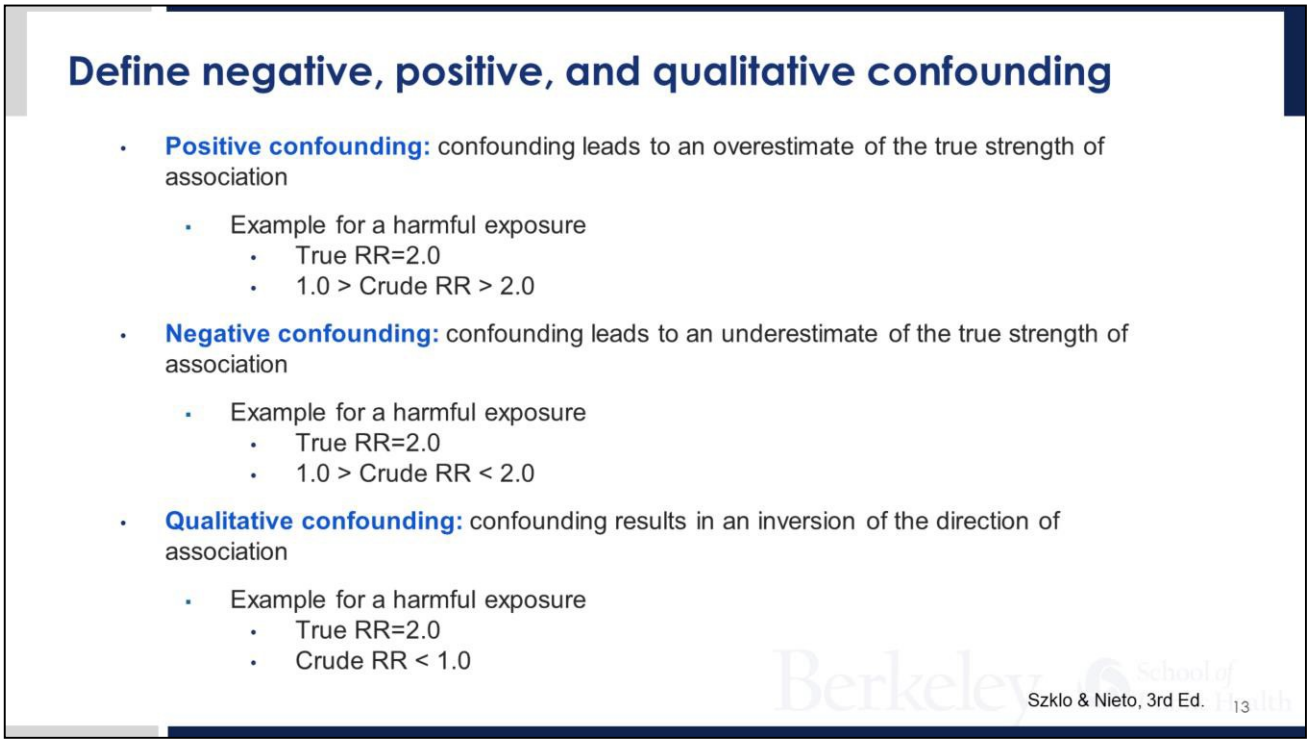
Szklo & Nieto, 3rd Ed. 12

So what do we see? In the stratified odds ratios, that is, in the two different strata, if we calculate an odds ratio-- and you should do that yourself to see that you get these numbers-- in the upper one, we see an odds ratio of 1.06. And in the lower one, we see an odds ratio of 1.00.

Now, these need to be compared to the crude odds ratio. Now, the crude odds ratio you should calculate yourself. But it would result from combining the two tables together. So in other words, we would create a new two-by-two table where we would add 53 and 35 to create cell A, and 15 plus 53 to create cell B, and 10 plus 52 to create cell C, and 3 plus 79 to create cell D. That two-by-two table we call the crude odds ratio, because we haven't stratified by anything. And if you do that, you'll get a crude odds ratio of 1.71.

Now, there's another little step in the reasoning here. We want to look at whether the stratified estimates are similar to each other. And in this case, they're very obviously similar to each other, 1.06 and 1.00. So if you were to average these two, and you'd use a technique like the Mantel-Haenszel method that you learned in 250A, you would come up with a value somewhere between 1.00 and 1.06. That average value is very different than the crude value of 1.71.

So the fact that this collapsed value of 1.71 is so different from this stratified average implies that there is confounding. So in this case, the outdoor, the occupation site, outdoor versus indoor, is a confounder of the relationship between gender and malaria.
We'll continue in our discussion of how to look at this collapsibility issue across the strata. So what we want to ask is, does the crude exposure outcome association have the same direction and similar magnitude as the confounder adjusted association? So in the prior slide we saw that we had our stratified results. And then an adjusted result would be the average of those two. So this table, now, is going to show us several different examples and ask us to evaluate how much confounding is present.

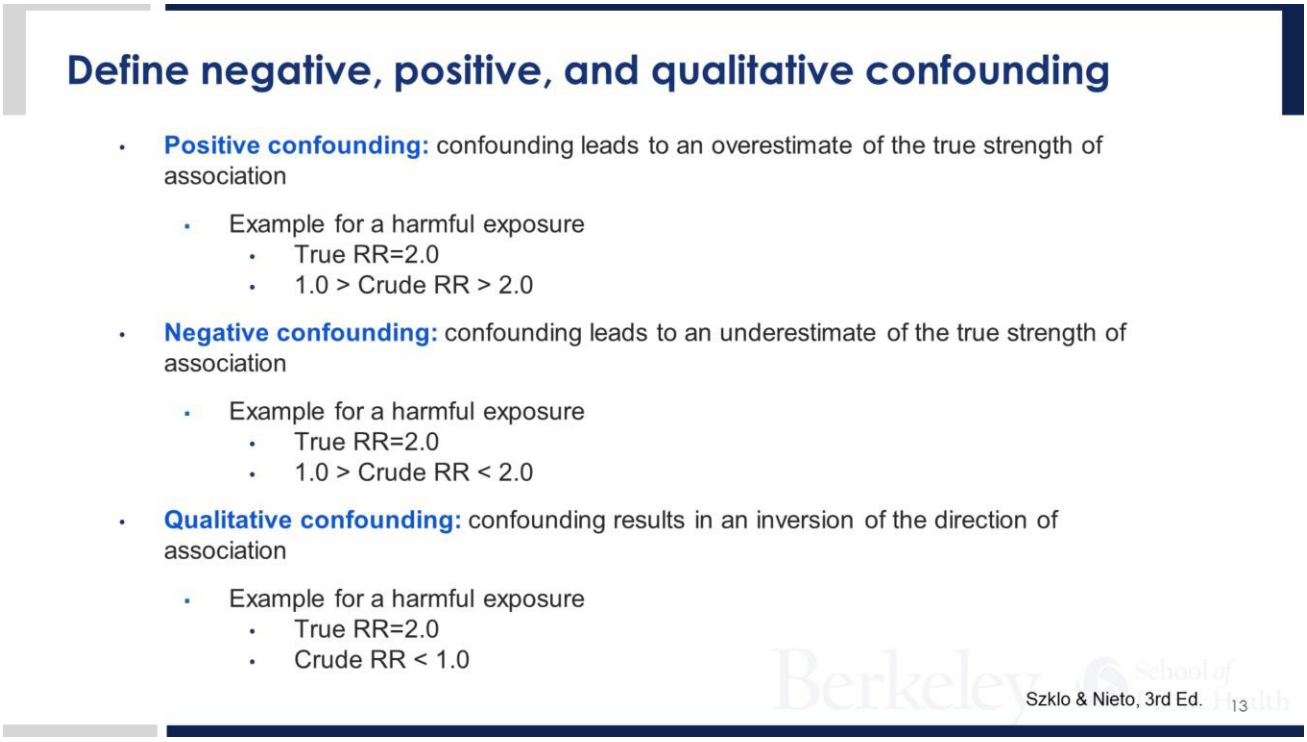


In these examples below, we're looking at the exposure being different medical interventions and the risk of re-admission being the outcome. And we're going to adjust for different variables and see whether adjusting changes our assessment of the risk. And in this case, we're going to use a hazard ratio, which is very similar to a relative risk.

But let's look, for example, at the first row. When we adjust for long-term oxygen therapy, the relationship between medical intervention and risk of readmission to the hospital, which is 2.36 if we don't adjust, becomes 1.38 if we do. So you see there's a big change from the crude to the adjusted value. It's a 72% change. So that implies that's, in a sense, the amount of confounding that's occurring by that particular variable.

Let's do another one, the last one, being under the care of the pulmonologists. With no adjustment for that confounder, the crude ratio was 2.16. But it becomes 1.73 when adjusted. So in all these examples here, we're comparing the crude to the adjusted. If the crude and the adjusted are quite different-- and we'll talk about what makes them different in just a moment-- if they're quite different, then we believe that confounding is present.

Now, in talking about confounding, we often talk about positive, negative, and qualitative confounding. For positive confounding, this is a situation that leads to an overestimate of the true strength of an association. So if my crude value is farther



from the null than my adjusted value, then I have positive confounding. And that can occur on either side of the null.

So if my crude value is 3 and my adjusted value is 2, that's positive founding. But similarly, if my crude value is 0.7 and my adjusted value is 0.9, that's also positive confounding. What's negative confounding? Well, negative confounding leads to an underestimate of the strength of the association.

So if my crude value is 4.0 and my adjusted value were 6.0, the crude would be closer to the null than the adjusted. So that would be negative confounding. And similarly, on the other side of the null, think about that a little bit. It's distance from the null. It's not the absolute value of the number. It's distance from the null.

And finally, what is qualitative confounding? Well, this just means a complete reversal in the direction of the association. So if my crude estimate of an effect were 2.0 and my adjusted estimate of effect were 0.7, that would be qualitative confounding, because the direction has completely changed from an increased risk, 2.0, to a decreased risk, 0.7.

Define negative, positive, and qualitative confounding

Table 5–8 Directions of the Associations of the Confounder with the Exposure and the Outcome, and Expectation of Change of Estimate with Adjustment (Assume a Direct Relationship Between Exposure and Outcome, ie, for Exposed/Unexposed, RR, or Odds Ratio > 1.0)

Association of Confounder with Exposure Is	*Association of Confounder with Outcome Is*	*Type of Confounding*	*Expectation of Change from Unadjusted to Adjusted Estimate*
Direct*	Direct*	Positive‡	Unadjusted > Adjusted
Direct*	Inverse†	Negative§	Unadjusted < Adjusted
Inverse†	Inverse†	Positive‡	Unadjusted > Adjusted
Inverse†	Direct*	Negative§	Unadjusted < Adjusted

Szklo & Nieto, 3rd Ed. 14

And this table just creates several different combinations of all these different possibilities. And this is dealing with the relationship, now, between the association of the confounder with the exposure. And here, we see in this first row, we see the association of the confounder with exposure. If that's direct and the association of the confounder with the outcome is also direct and the type of confounding is positive, then our adjusted value is going to be less than our unadjusted.

And you can go through each of these and see what happens. But basically, this is meant to remind us how complicated this can get, because we have to be thinking both about the relationship with the confounder with the exposure, the relationship of the confounder with the outcome, and then the type of confounding that's occurring.

Examples of negative, positive, and qualitative confounding

Table 5–7 Hypothetical Examples of Unadjusted and Adjusted Relative Risks According to Type of Confounding (Positive or Negative)

Example No.	*Type of Confounding*	*Unadjusted Relative Risk*	*Adjusted Relative Risk*
1	Positive	3.5	1.0
2	Positive	3.5	2.1
3	Positive	0.3	0.7
4	Negative	1.0	3.2
5	Negative	1.5	3.2
6	Negative	0.8	0.2
7	Qualitative	2.0	0.7
8	Qualitative	0.6	1.8

Szklo & Nieto, 3rd Ed. 15

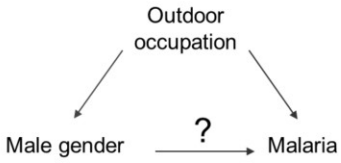
Here's some examples of negative, positive, and qualitative confounding. Let's work through a couple of these. And, again, these are hypothetical. But imagine we have positive confounding present where, in this situation, the unadjusted risk is 3.5 and the adjusted risk is 1.0. Well, this is positive confounding, because the unadjusted, or crude risk, is farther from the null than the adjusted risk.

Let's do an example of a negative confounding. Negative confounding, example number 4, here, our unadjusted relative, or crude, risk is 1.0. But our adjusted is 3.2. So the crude value, the confounded value, has moved toward the null.

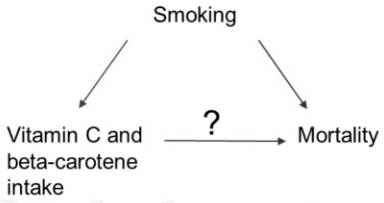
And finally, an example of qualitative confounding, example number 7, the unadjusted relative risk is 2, which means an increase in risk. And the adjusted is 0.7, which is a decrease in risk. So this is a reversal in direction. So this is called qualitative confounding.

Confounding is not an "all or none" phenomenon

- Sometimes a confounding variable is responsible for the **entire relationship** between the exposure and outcome
- Assuming no other confounders, adjusting for this confounder will have a **large effect** on the estimated measure of association — it may cause it to be null.



- Sometimes it is only responsible for **part of the relationship**.
- Assuming no other confounders, adjusting for this confounder will have a **small effect** on the estimated measure of association.



Szklo & Nieto, 3rd Ed. 16

Now, confounding is not an all or none phenomenon, although you can have situations where the confounding variable is responsible for the entire relationship between the exposure and outcome. So, for example, let's say there are no other confounders and that when we adjust for a confounder and the relationship between male gender is the exposure and malaria is the outcome, if there is a large effect on the estimated measure of association, it may turn it into a null association. So without adjustment you might see a strong association between male gender and malaria. And that's what we saw in the earlier slides.

But when we adjust it for occupation, that relationship went away. There was no relationship between male gender and malaria. So the confounding, there, induced a relationship, in a sense. It caused the entire relationship we thought we were seeing in the crude relationship.

But sometimes, a confounder is only responsible for part of a relationship. So assuming no other confounders, adjusting for a confounder in this situation will have a small effect on the estimated measure of association. So, for example, let's say that smoking is a small confounder of vitamin C and beta carotene intake in the relationship of vitamin C and beta carotene intake with mortality. In this situation, the difference between the confounded result and the adjusted result, adjusted for smoking, might be small.

Confounding is not an "all or none" phenomenon
Example: strong confounder

Crude OR = 1.71

Exhibit 5–4 Stratified Analyses of the Association Between Gender and Malaria (from Exhibit 5–2), According to Whether Individuals Work Mainly Outdoors or Indoors

Mostly outdoor occupation

	Cases	Controls	
Males	53	15	**Odds ratio = 1.06**
Females	10	3	
Total	63	18	

Mostly indoor occupation

	Cases	Controls	
Males	35	53	**Odds ratio = 1.00**
Females	52	79	
Total	87	132	

The association is does not differ greatly when stratifying by the confounder.

The stratum-specific estimates are very different from the crude estimate.

The pooled OR would be close to 1.00 based on these stratified estimates.

Here's that example we worked with gender and malaria where occupational location was a very strong confounder, because the relationship between gender and malaria was 1.71 until we adjusted for occupation location when it went down to almost 1.

Confounding is not an "all or none" phenomenon
Example: weak confounder

Table 5–6 Unadjusted and Smoking-Adjusted All-Cause Mortality Ratios Rate in the Western Electric Company Study

	Vitamin C/Beta Carotene Intake Index Rate Ratios		
Rate Ratios	*Low*	*Moderate*	*High*
Unadjusted	1.00	0.82	0.79
Adjusted*	1.00	0.85	0.81

There are very small differences between unadjusted and adjusted rate ratios (and no difference for the low intake category). This suggests that smoking was a weak confounder.

Here's an example with a weak confounder. This is looking at the relationship between vitamin C and beta carotene intake and all cause mortality. And we've now split the confounder into three levels, low, moderate, and high. Smoking is the confounder. So we're looking at low, moderate, and high levels of smoking.

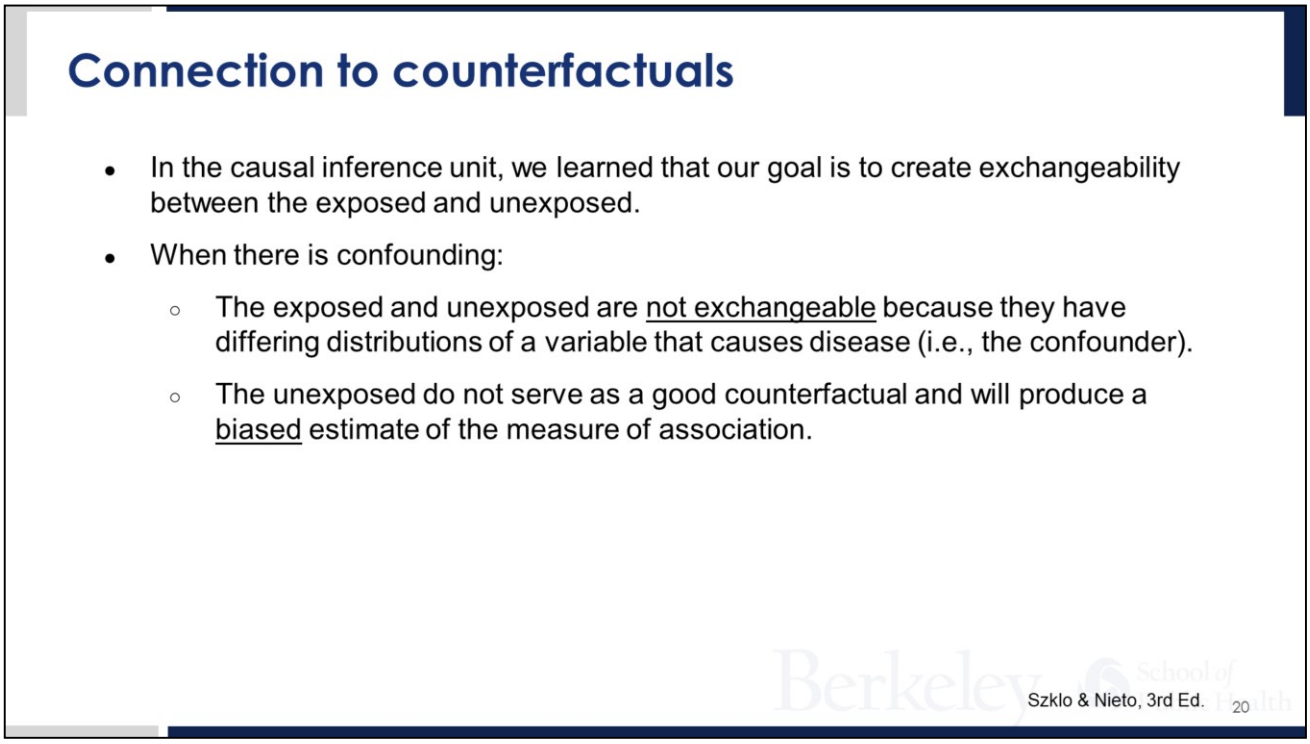
And what we see is in the unadjusted analysis, for example, moderate levels of smoking, going from unadjusted to adjusted goes from 0.82 to 0.85. That's a very small change. So this is a weak confounder. Smoking is only a weak confounder of this relationship. Similarly, in the high smokers, going from unadjusted to adjusted, goes from 0.79 to point. 0.81, again, very small impact. So this smoking is, here, a very weak confounder.

Role of statistical significance in assessing confounding

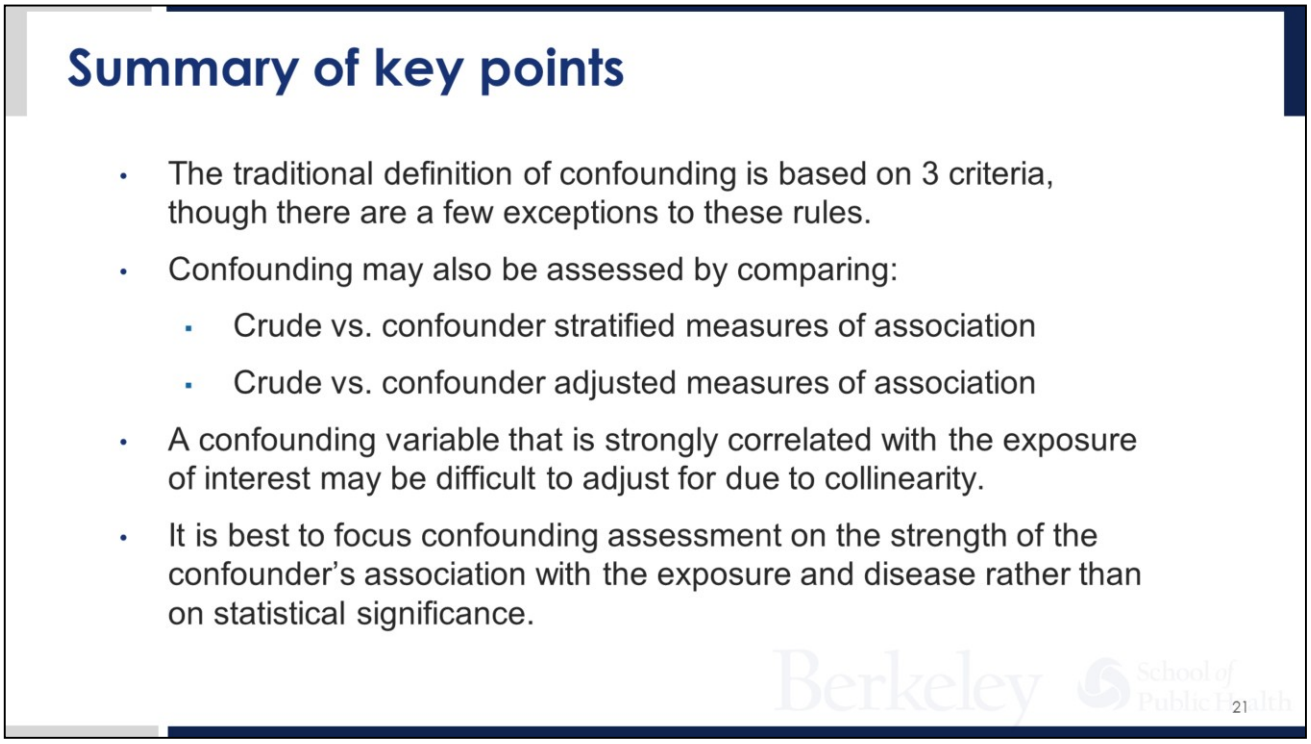
- It is inappropriate to rely solely on statistical significance to identify a confounder, especially when the exposure or outcome is strongly associated with the confounder.
- If using a p-value for the strength of association with a confounder, it is recommended to use a cutoff of 0.2 instead of 0.05 to reduce the chance of a Type II error (and err on the side of detecting a confounder).
- However, generally it is best to focus on the magnitude of the association of the confounder with the exposure and outcome.

Szklo & Nieto, 3rd Ed. 19

Sometimes, people will talk about whether we use statistical testing or statistical significance to assess confounding. Generally, it's felt to be inappropriate to rely solely on statistical significance to identify a confounder, especially when the exposure or outcome is strongly associated with the confounder. If you're using a p-value to test the strength of association with the confounder, it's recommended to use a cut-off 0.2, not 0.05. This will reduce the chance of a type II error. However, generally, it's best to focus on the magnitude of the association of the confounder with the exposure and outcome, not the statistical significance of that association.



There's also a connection between confounding and counterfactuals. Remember, back in the causal inference units, we talked a lot about how our goal was to create exchangeability between exposed and unexposed persons in a study. So when there is confounding present, the exposed and unexposed individuals are not exchangeable, because they have differing distributions of the confounding variable that causes disease. And in that situation, the unexposed do not serve as a good counterfactual. They'll produce a biased estimate of the measure of association.



So let's summarize the key points from this talk. The traditional definition of confounding is based on three criteria, though there are a few exceptions to these rules. And we went through each of the rules and each of the exceptions.

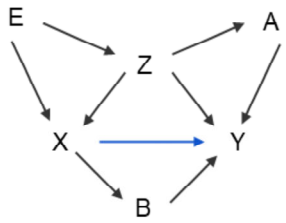
Confounding can be assessed by comparing the crude versus the confounder stratified measure of association or the crude versus confounder adjusted measure of association. A confounding variable that is strongly correlated with the exposure of interest may be difficult to adjust for due to collinearity. Think back to that example with the air pollution being so tightly linked to distance from the freeway. We had no ability to study the effect of the confounder because of that strong collinearity. And finally, it's best to focus confounding assessment on the strength of the confounder's association with the exposure and the disease, rather than on statistical significance.

Lecture: Using DAGs to Detect Confounding

In this video, I'll talk about how to use directed acyclic graphs, or DAGs, to detect confounding.

Intuition behind using DAGs to assess confounding

- We are only interested in directed paths from X to Y.
- Any other unblocked paths between X and Y imply statistical dependence due to reasons other than the causal relationship between X and Y (e.g., confounding).
- DAGs help us:
  - 1) Identify confounders
  - 2) Assess the implications of restriction, stratification, and multivariate adjustment.



So take a look at this DAG on the right side of this slide. We're interested in the relationship between X and exposure and Y and outcome.

And that means that we're only interested in directed paths from X to Y. So in our figure here, that's the wider blue arrow. That's the causal effect of X on Y.

Any other association or statistical relationship between X and Y that's not due to the direct effect of X on Y could be considered to be a source of confounding or other bias. Another way of saying this is that any other unblocked paths between X and Y imply statistical dependence due to reasons other than the causal relationship between X and Y.

So our goal in this video is to talk about how we can remove confounding of the X/Y relationship so that in our data, in our estimation, we're actually estimating the causal effect of X on Y. Directed acyclic graphs can help us identify confounders and decide which confounders to adjust for, and then also assess the implications of restrictions, stratification, and multi-variate adjustment.

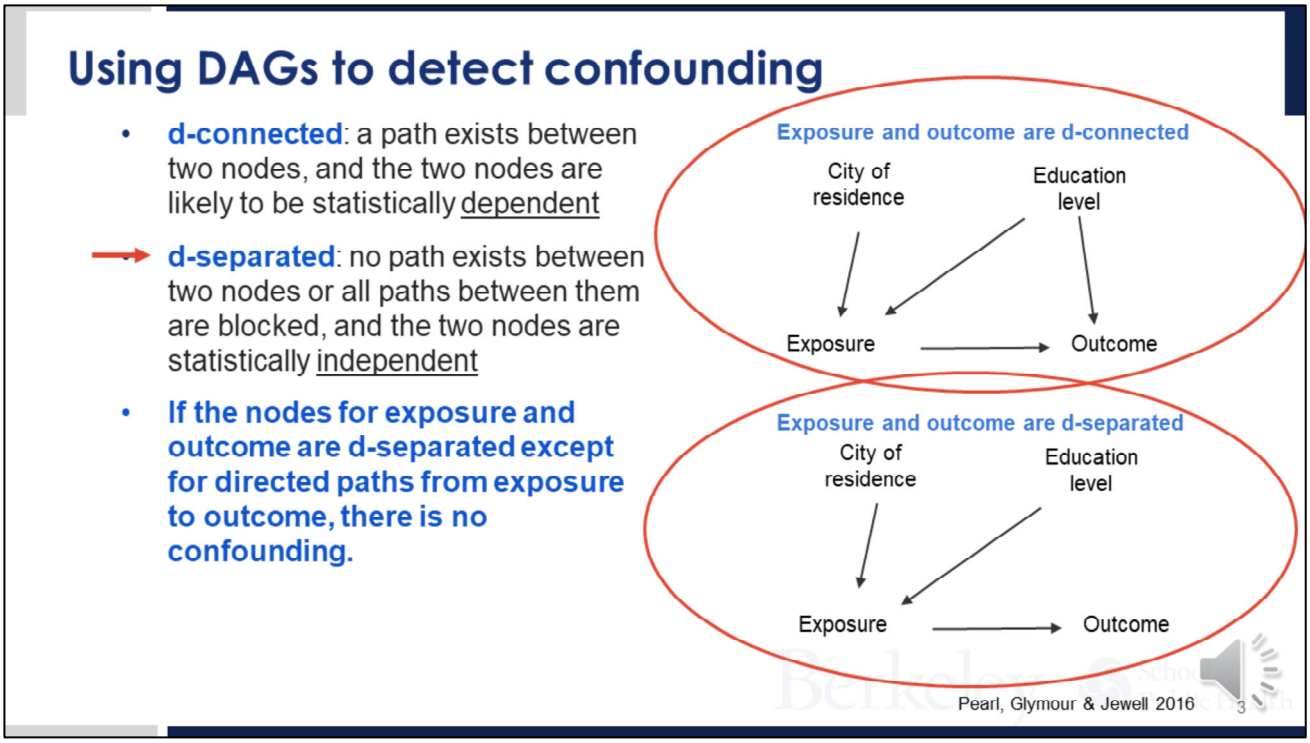
Using DAGs to detect confounding

- Most real DAGs are complex and include many different potential pathways between the exposure and outcome.
- We can use a process called **d-separation** to predict dependencies between nodes in a DAG.
- The “d” stands for “directional”

Pearl, Glymour & Jewell 2016 2

So we've looked at a variety of different DAGs in the course so far. And in most real studies, they're very complex. And they have many different potential pathways between the exposure and outcome beside the direct arrow from exposure to outcome.

We're going to talk about a process called "d-separation." So this helps us predict dependencies between different nodes in a DAG, and the "d" stands for "directional." So we use the term "d-connected" to say that a path exists between two nodes and that the two nodes are likely to be statistically dependent as a result of that path.



*Let's look at the DAG on the upper right-hand corner here. So we're interested in the relationship between the exposure and the outcome. We also have city of residence and education level.

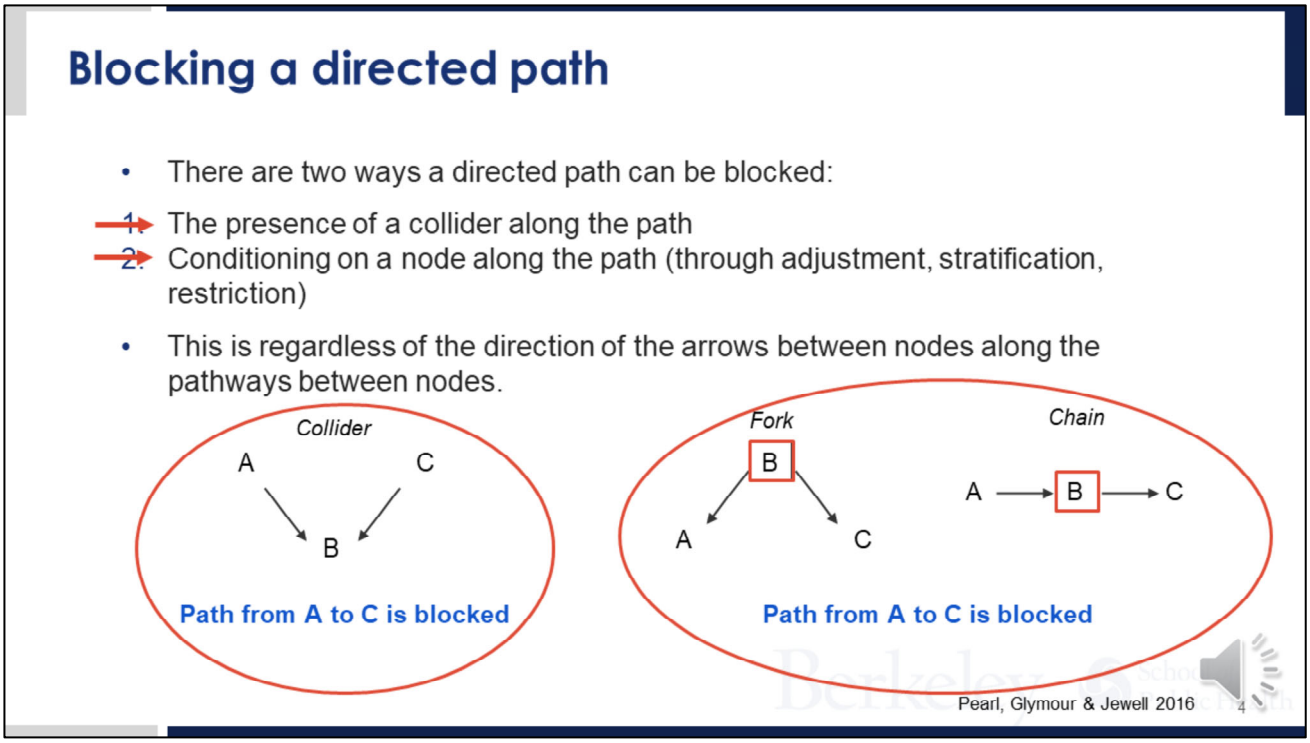
And in this DAG, we can say that the exposure and the outcome are d-connected. So if we removed temporarily the arrow from exposure to outcome, that's the causal effect of interest, so all paths between exposure and outcome besides that path are the ones we're looking at.

We can see that the exposure and outcome are still d-connected because there is a path from exposure to education level to outcome. This path is irregardless of the direction of the arrows. So even though the flow is going from education level into exposure, we still call this path d- connected. And that's because these variables, exposure and outcome, are statistically dependent because of this set of arrows between education level and exposure and education level and outcome.

*We call two nodes d-separated if no path exists between two nodes or all paths between them are blocked. We'll come back to what that means in a moment. And as a result, the two nodes are considered statistically independent if they are d-separated.

*So in the DAG on the bottom right, the exposure and outcome are d-separated, except for the path from exposure directly to outcome because in this DAG, there's

no longer an arrow from education level to outcome. So the only path from exposure to outcome is the one that we're interested in.
If the nodes for exposure and outcome are d-separated except for the directed path from exposure to outcome, then there's no confounding of the exposure outcome relationship. So this is our goal.



Now, on the last slide I briefly mentioned that a path from exposure to outcome could be blocked. There is two different ways a directed path can be blocked that we're going to discuss. The first is the presence of a collider along the path.

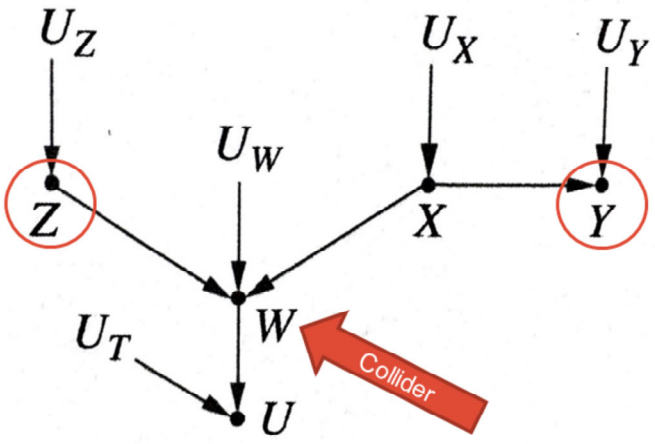
*So if you look at the first DAG on the bottom left, there's a collider between A and C. So node B is the collider. When a collider exists, statistically it blocks the relationship or the path from A to C.

The second type of blocking is conditioning on the node along the path. This can be done through adjustment, stratification, or restriction. *So in the examples on the bottom right, we have a fork structure where B leads to A and B leads to C. *If we condition on B by stratifying on B or restricting to one value of B, for example, then we are blocking the path from A to C.

And then the final DAG has a chain structure from A to B and B to C. And if we *control for B, or condition on B, then we're also blocking the path from A to C. So again, this is regardless of the direction of the arrows between the nodes.

What you want to look for are these particular structures-- colliders, forks, and chains. And then for collider-- if there is a collider along the path between exposure and outcome, then that path is blocked. For forks and chains, we want to see if the central node in the fork and in the chain is conditioned on. And if so, then the path is blocked.

Example #1: Are Z and Y d-separated?

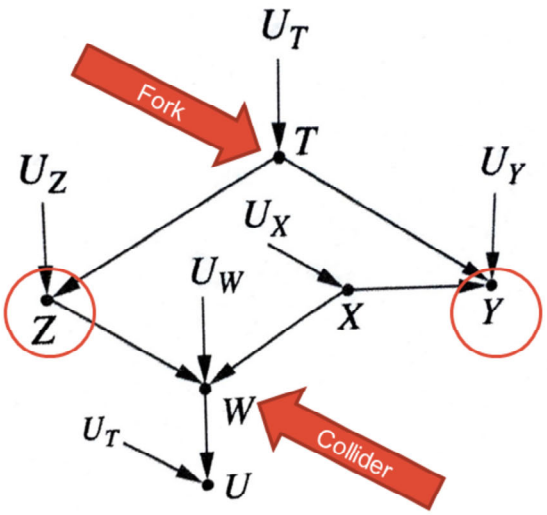


Pearl, Glymour & Jewell 2016

Let's go for an example. So in this example, we want to assess whether node Z-- *right here-- and node Y-- *right here-- are d-separated. So let's look for any blockages on the path from Z to Y.

Well, the path that exists goes from Z to W, W to X, and X to Y. And we can see that W is a collider on the path from Z to X. Because W is a collider, it means the path from Z to Y is blocked, and so Z and Y are d-separated.

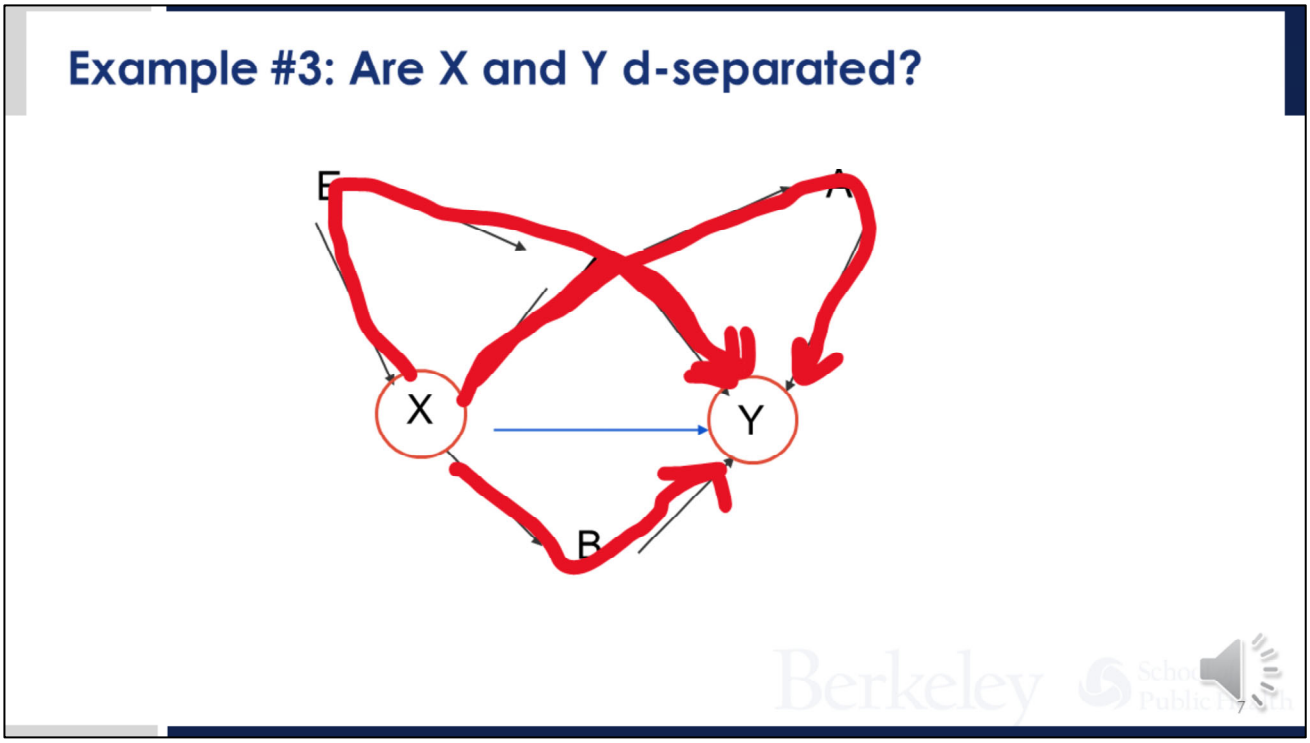
Example #2: Are Z and Y d-separated?



Pearl, Glymour & Jewell 2016

Here is a second example. Again, we're going to look at whether Z and Y are d-separated in this example. So we have a similar structure as in the last example, where there is a collider, W, on the path from Z to Y. But in this example, there is yet another path. It's a fork. So we have T into Z and T into Y.

And because T is not conditioned on-- right, there's no box around T-- it means there is an open path from Z to Y. And so Z and Y are not to d-separated. They are d- connected through the path through T.



Here's a third example. In this example, we want to see whether X and Y are d-separated. And the answer is no. There are actually a number of different paths between X and Y.

For example, there is a fork from Z to X and Z to Y. There's a path from X to Z to A to Y, from X to E to Z to Y, and then another path from X to B to Y. So there are many different paths between X and Y besides the causal effect of X on Y.

The Backdoor Criterion

- This criterion helps us assess under what conditions is the association between exposure and outcome unconfounded. (i.e., under what conditions is there a causal effect between the exposure and outcome)
- **Formal definition:** For a pair of nodes X and Y, a set of variables Z satisfies the backdoor criterion if no node in Z is:
  a. A descendant of X, and
  b. Z blocks every path between X and Y that contains an arrow into X

Pearl, Glymour & Jewell 2016 8

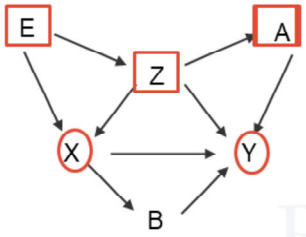
So now I'd like to talk about the backdoor criterion. And this criterion helps us assess under what conditions the association between exposure and outcome are unconfounded. In other words, under what conditions is there a causal effect of the exposure on the outcome?

So here is the formal definition. It's a little jargony, but let's just see how this goes. So, "For a pair of nodes, X and Y, a set of variables, Z, satisfies the backdoor criterion if no node in Z is a descendant of X and Z blocks every path between X and Y that contains an arrow into X."

The Backdoor Criterion

- This criterion helps us assess under what conditions is the association between exposure and outcome unconfounded. (i.e., under what conditions is there a causal effect between the exposure and outcome)
- **Formal definition:** For a pair of nodes X and Y, a set of variables Z satisfies the backdoor criterion if no node in Z is:
  a. A descendant of X, and
  b. Z blocks every path between X and Y that contains an arrow into X

In this DAG, conditioning on the following sets could meet the criterion:
- Z,
- Z, A
- Z, E
- Z, A, E



Pearl, Glymour & Jewell 2016

So this is the DAG from the last slide. Let's take a look at this. We're interested in the effect of X on Y. And we want to identify a set of variables that we're going to call Z.

Now, I realize this is confusing because we have a node, Z, in this DAG. But just keep in mind that when we say "a set of variables, Z," that that's just a general term. A set may or may not include that node Z in the DAG.

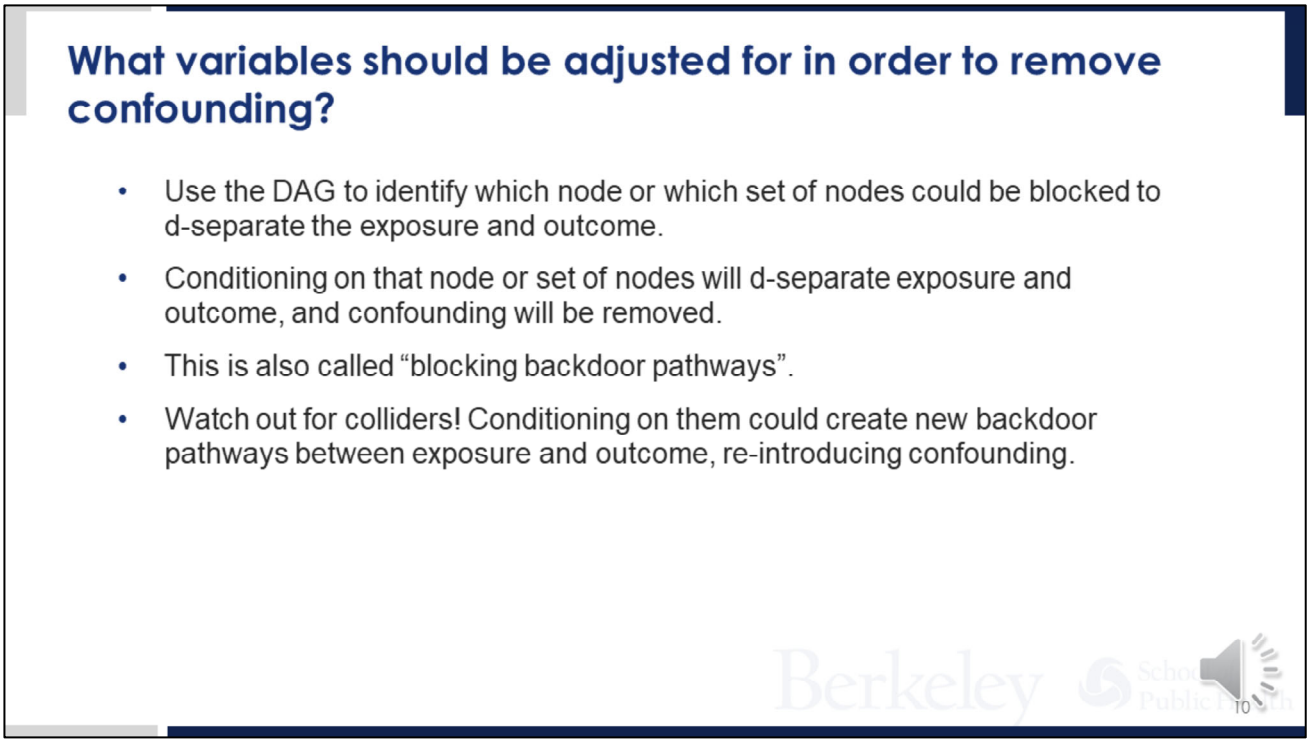
So we want to look for a set of variables that satisfies these two criteria. So the first is that the set cannot include a descendant of X. OK, what's a descendant of X in this DAG? Well, B, because X leads to B. So B cannot be included in the set that meets this definition. All the other nodes can meet the first criterion here, that they are not descendants of X.

The second is that Z blocks every path between X and Y that contains an arrow into X. So currently, none of the paths between X and Y are blocked. But if we condition on certain sets of nodes, we're able to meet the second criterion. So just *conditioning on Z would actually block all the remaining pathways between X and Y, except for the one through B.

And then we could also use these other sets here. So we *could condition on Z and A. *We can condition on Z and E. *We could condition on all three. None of them are colliders, so none of them are going to introduce additional backdoor pathways.

*Now we'll go over some more subject matter examples in a moment, but when it comes to choosing between these different sets, all of them are sufficient for confounder adjustment. So they're all sufficient to remove confounding.

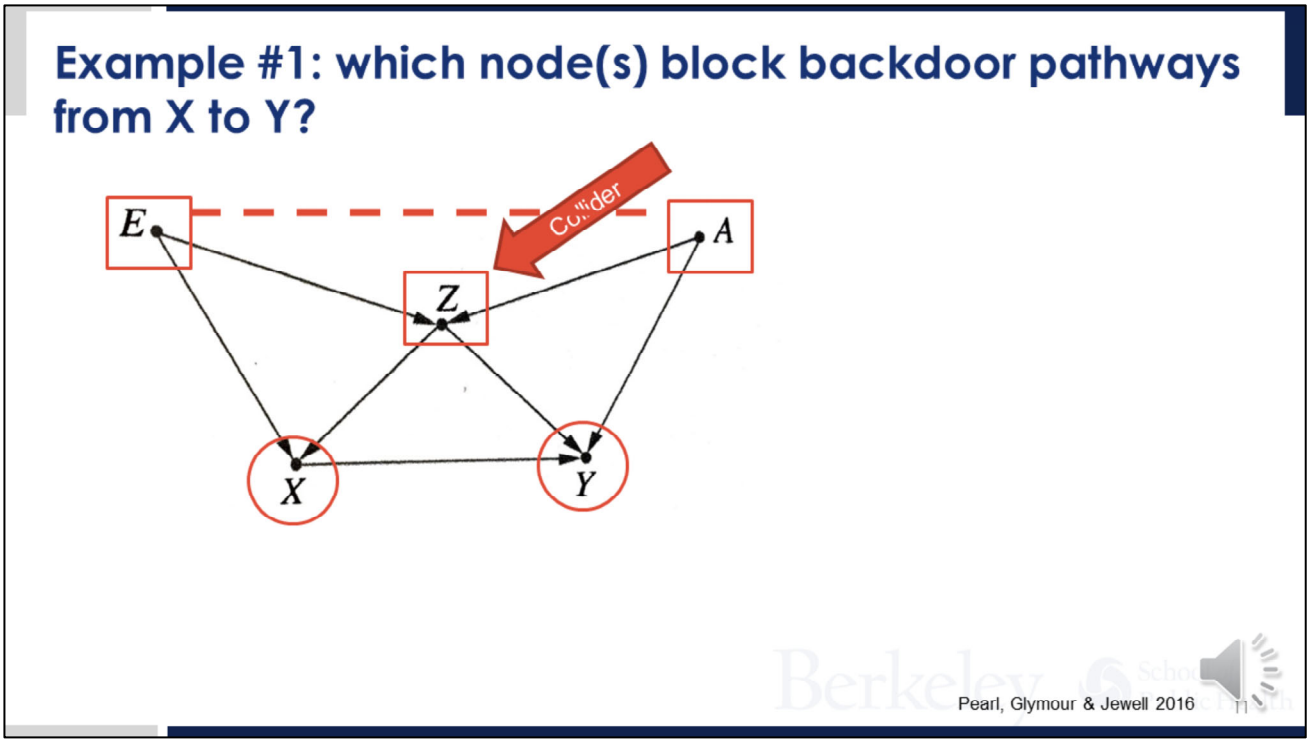
The choice between just conditioning on Z or conditioning on Z and A or Z and A and E will come down to other factors, such as how common these particular variables are, and how many variables you may be adjusting for in your model for other reasons. So those factors are really beyond the scope of this particular video. But we'll return to them later in the course.



So how do we know what variables should be adjusted for in order to remove confounding? Well, we can use that DAG to identify which node or set of nodes can be blocked, to d-separate the exposure and the outcome, except for that direct arrow from exposure to outcome.

And as I just discussed, conditioning on that node or set of nodes will separate the exposure and outcome, removing confounding. And another way of referring to this is that we're trying to block all the backdoor pathways from exposure to outcome.

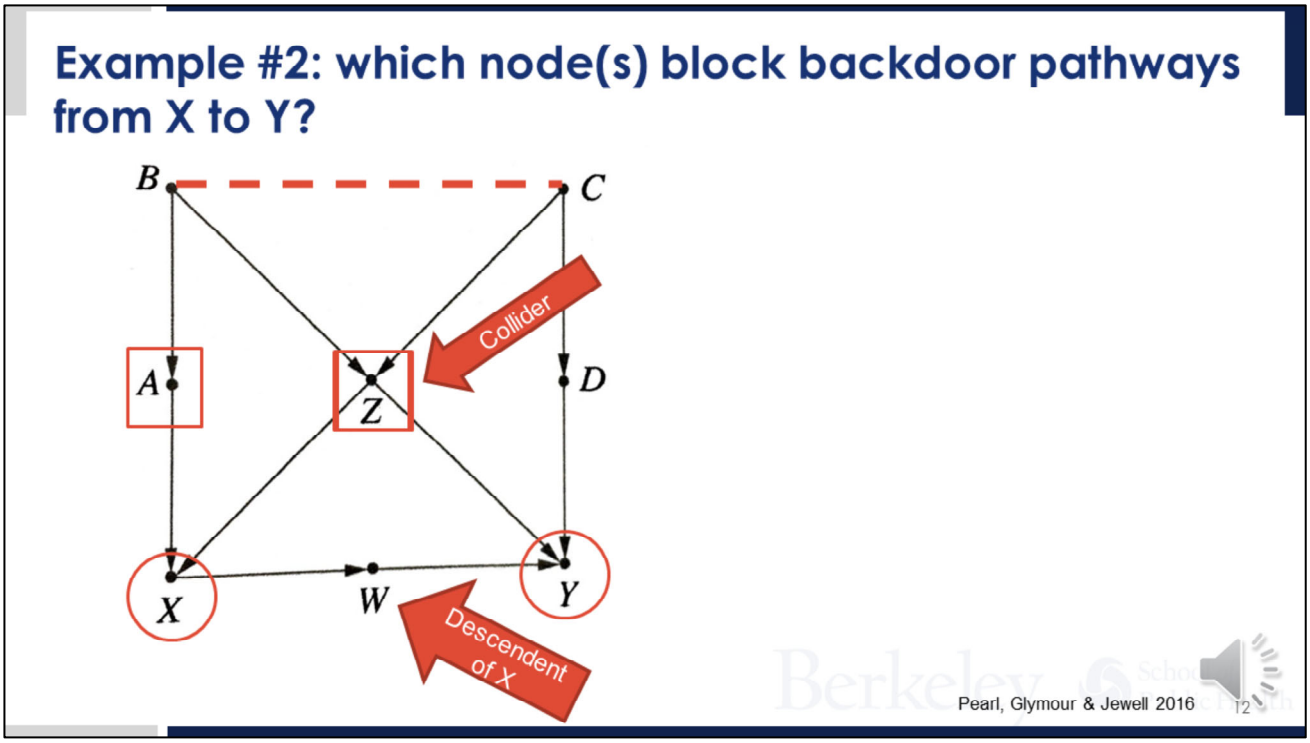
And as I've mentioned, we need to watch out for colliders because conditioning on them could create new backdoor pathways between the exposure and outcome that weren't originally present in the DAG. And that could reintroduce confounding after you've done some initial adjustment.



Let's go over an example. So in this example, we're interested in the effect of X on Y. And we want to identify which node or set of nodes we need to control for or adjust for or stratify on to block backdoor pathway from X to Y. So pause the video for a moment and come up with the nodes or set of nodes that you would want to adjust for.

OK, so we have a path from X to E to Z to Y, from X to Z to A to Y, and from X to E to Z to A to Y. And the first thing I want to note is that we have a collider on Z. So if we condition on Z, it's going to open up a backdoor pathway between E and A, its two parents.

So we wouldn't want a condition only on Z. But if we condition on Z, as long as we also condition on A and/or E, we can block that new pathway. And then all of the backdoor pathways will be blocked from X to Y.



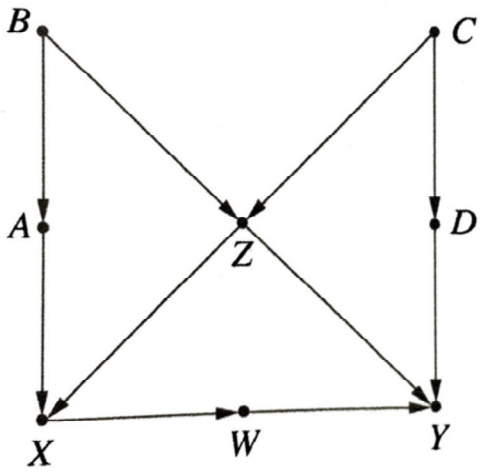
Here's a second example. So again, we're interested in the effect of X on Y. Take a moment to think about which nodes you would condition on to block backdoor pathways between X and Y.

OK, so again, I immediately see a collider with the node Z. So it's a collider of the relationship between B and C. *So we want to be careful about conditioning on Z.

And again, there are also a number of different ways we can get from X to Y besides the path from X to W to Y. Now, we don't want a *condition on W because it's a descendant of X. So let's focus on the other nodes-- A, B, C, D, and Z.

Conditioning on A would be sufficient to block the path from X to A to B to Z to Y. But it would still leave that path through Z. So that means that all of the different sets that block backdoor pathways from X to Y require conditioning on Z. And that will open another path through B to C, which requires us to condition on additional nodes that close that backdoor pathway.

Example #2: which node(s) block backdoor pathways from X to Y?



1. *Sets of 2 nodes:* $\{Z, A\}, \{Z, B\}, \{Z, C\}, \{Z, D\}$
2. *Sets of 3 nodes:* $\{Z, A, B\}, \{Z, A, C\}, \{Z, A, D\}, \{Z, B, C\}, \{Z, B, D\}, \{Z, C, D\}$
3. *Sets of 4 nodes:* $\{Z, A, B, C\}, \{Z, A, B, D\}, \{Z, A, C, D\}, \{Z, B, C, D\}$
4. *Sets of 5 nodes:* $\{Z, A, B, C, D\}$

Pearl, Glymour & Jewell 2016 13

There's lots of different ways to do this. And they're all listed here. And just note quickly that all of them involve conditioning on Z. So take a minute to pause the video and check this out for yourself, because there's lots of different ways that we can do this in this particular DAG.

Comparing DAG vs. traditional approach to assessing confounding

Traditional criteria:

1. Must be associated with exposure

   DAG: arrow from W to X (or X and W d-connected)

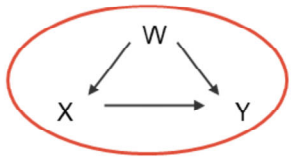
1. Must be an independent cause or predictor of disease

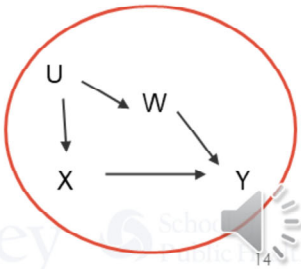
   DAG: arrow from W to Y

1. Cannot be an intermediate between exposure and disease

   DAG: W not on path from X to Y

→ In #1, for the association to exist when one variable does not cause the other, they have to share a common cause – the common cause may be unmeasured





14

Now I'd like to compare the DAG-based approach to the traditional approach using the three criteria to assess confounding. So the first of the traditional criteria is that the confounder must be associated with the exposure. And in a DAG, if our confounder is W, this would mean there's an arrow from W to X. Or if the DAG is more complex, we could say that X and W need to be d-connected.

The second criterion is that the confounder must be an independent cause or predictor of disease. So that means that there needs to be an arrow from W to Y. Another way of saying this is that Y and W must be d-connected.

And then finally, the confounder can't be an intermediate between the exposure and disease. So in a DAG, that would mean that W is not on the path from X to Y. *So in the top right DAG, W meets the criterion under both the traditional and DAG-based approaches.

*And one other thing to note is that in the second criterion that the confounder must be associated with the outcome, for the association to exist when the confounder doesn't directly cause the exposure X. That means they have to share a common cause. And that can be an unmeasured cause of W and of X.

*And so that's indicated in the bottom right DAG. So there is a backdoor pathway from X to U to W to Y. And this can lead to confounding under both criteria.

What do DAGs offer that the traditional criteria do not?

- Clear identification of colliders
- Sufficiency of confounder adjustment
- Most often, these two approaches agree.
  - When they do not, it is the three criteria that fail to detect confounding.

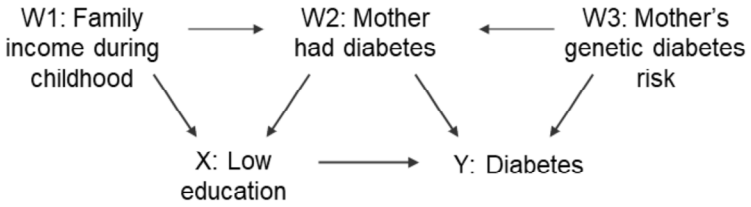
At first glance, it may appear that these two approaches will always yield the same result. But that's actually not the case. The differences are that the DAG approach can clearly identify colliders and that it also can be used to assess this sufficiency of a potential set of confounders that you're interested in adjusting for, instead of adjusting for all the confounders that meet the traditional criteria.

The traditional approach allows us to think about confounders one at a time, whereas the DAG-based approach allows us to think about confounders adjusted for it together, or at one time. When the two approaches don't agree, it's always the three criteria that fail to detect confounding. And we're going to look at that in an example in the next slide.

Example of when the traditional criteria fail to capture confounding

Using traditional criteria, should we control for W2? Yes.

1. Must be associated with exposure ✔
2. Must be an independent cause or predictor of disease ✔
3. Cannot be an intermediate between exposure and disease ✔



16

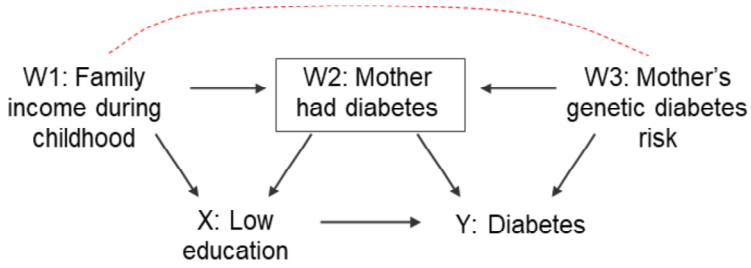
So here in this DAG, we're interested in the effect of low education, X, on diabetes, Y. And we have three different variables noted as W1, W2, and W3 that also appear in the DAG. So using the traditional criteria, should we control for W2?

Well, let's look at the first criterion. So is W2 associated with the exposure? Yes, there is an arrow from W2 to X. Is W2 an independent clause or predictor of disease? Yes, there is an arrow from W2 to Y. And is W2 an intermediate on the path from exposure to disease? No. So under the three criteria, we would say yes, we should control for W2.

Example of when the traditional criteria fail to capture confounding

Using the backdoor criterion, should we control for W2? No

W2 is a collider, and controlling for it opens a backdoor pathway through W1 and W3. As a result, if we control for W2, we also have to control for W1 or W3.



17

Now if we use the backdoor criterion under the DAG-based approach, should we control for W2? And the answer is no. W2 is a collider. And when we control for W2, that means we're going to open a new pathway between W1 and W3. And this pathway would be a backdoor pathway from X to Y.

And so if we want to control for W2 alone, there would still be confounding of X and Y. We can also control for W1 or also control for W3 along with W2. And that would be sufficient to close that backdoor pathway.

But it's important to note here that the traditional criteria wouldn't have told us that. And that's because, again, it's only considering one variable at a time. So hopefully this example has helped convince you that it's always a good practice to use DAGs. They just allow for a more nuanced assessment of potential confounding and analysis.

Summary of key points

- To assess whether an exposure-outcome relationship is confounded, we can assess whether the two nodes are d-separated in a DAG.
- To determine which node or set of nodes needs to be adjusted for to remove confounding, we can identify the node(s) that meet the backdoor criterion.
  - Conditioning on these nodes blocks any backdoor pathways from X to Y and ensures d-separation of X and Y
- Most often, the traditional and DAG-based approaches agree, but when they do not, it is the three criteria that fail to detect confounding.

To summarize, to assess whether an exposure outcome relationship is confounded, we can assess whether two nodes are d-separated in a DAG. And we can also use the backdoor criterion. So this helps us assess which node or set of nodes needs to be adjusted for to remove confounding.

Conditioning on these nodes will block any backdoor pathways from X to Y and insure a d-separation of X and Y. And most often, the traditional and DAG-based approaches will agree, but when they do not, it is the three criteria that fail to detect confounding. And that's why we recommend a DAG-based approach.

1

Preliminaries: Statistical and Causal Models

1.1 Why Study Causation

The answer to the question "why study causation?" is almost as immediate as the answer to "why study statistics." We study causation because we need to make sense of data, to guide actions and policies, and to learn from our success and failures. We need to estimate the effect of smoking on lung cancer, of education on salaries, of carbon emissions on the climate. Most ambitiously, we also need to understand *how* and *why* causes influence their effects, which is not less valuable. For example, knowing whether malaria is transmitted by mosquitoes or "mal-air," as many believed in the past, tells us whether we should pack mosquito nets or breathing masks on our next trip to the swamps.

Less obvious is the answer to the question, "why study causation as a separate topic, distinct from the traditional statistical curriculum?" What can the concept of "causation," considered on its own, tell us about the world that tried-and-true statistical methods can't?

Quite a lot, as it turns out. When approached rigorously, causation is not merely an aspect of statistics; it is an addition to statistics, an enrichment that allows statistics to uncover workings of the world that traditional methods alone cannot. For example, and this might come as a surprise to many, none of the problems mentioned above can be articulated in the standard language of statistics.

To understand the special role of causation in statistics, let's examine one of the most intriguing puzzles in the statistical literature, one that illustrates vividly why the traditional language of statistics must be enriched with new ingredients in order to cope with cause–effect relationships, such as the ones we mentioned above.

1.2 Simpson's Paradox

Named after Edward Simpson (born 1922), the statistician who first popularized it, the paradox refers to the existence of data in which a statistical association that holds for an entire population is reversed in every subpopulation. For instance, we might discover that students who

Causal Inference in Statistics: A Primer, First Edition. Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell.

Companion Website: www.wiley.com/go/Pearl/Causality

smoke get higher grades, on average, than nonsmokers get. But when we take into account the students' age, we might find that, in every age group, smokers get lower grades than nonsmokers get. Then, if we take into account both age and income, we might discover that smokers once again get *higher* grades than nonsmokers of the same age and income. The reversals may continue indefinitely, switching back and forth as we consider more and more attributes. In this context, we want to decide whether smoking causes grade increases and in which direction and by how much, yet it seems hopeless to obtain the answers from the data.

In the classical example used by Simpson (1951), a group of sick patients are given the option to try a new drug. Among those who took the drug, a lower percentage recovered than among those who did not. However, when we partition by gender, we see that *more* men taking the drug recover than do men are not taking the drug, and more women taking the drug recover than do women are not taking the drug! In other words, the drug appears to help men and women, but hurt the general population. It seems nonsensical, or even impossible—which is why, of course, it is considered a paradox. Some people find it hard to believe that numbers could even be combined in such a way. To make it believable, then, consider the following example:

Example 1.2.1 *We record the recovery rates of 700 patients who were given access to the drug. A total of 350 patients chose to take the drug and 350 patients did not. The results of the study are shown in Table 1.1.*

The first row shows the outcome for male patients; the second row shows the outcome for female patients; and the third row shows the outcome for all patients, regardless of gender. In male patients, drug takers had a better recovery rate than those who went without the drug (93% vs 87%). In female patients, again, those who took the drug had a better recovery rate than nontakers (73% vs 69%). However, in the combined population, those who did not take the drug had a better recovery rate than those who did (83% vs 78%).

The data seem to say that if we know the patient's gender—male or female—we can prescribe the drug, but if the gender is unknown we should not! Obviously, that conclusion is ridiculous. If the drug helps men and women, it must help *anyone*; our lack of knowledge of the patient's gender cannot make the drug harmful.

Given the results of this study, then, should a doctor prescribe the drug for a woman? A man? A patient of unknown gender? Or consider a policy maker who is evaluating the drug's overall effectiveness on the population. Should he/she use the recovery rate for the general population? Or should he/she use the recovery rates for the gendered subpopulations?

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

The answer is nowhere to be found in simple statistics. In order to decide whether the drug will harm or help a patient, we first have to understand the story behind the data—the causal mechanism that led to, or *generated*, the results we see. For instance, suppose we knew an additional fact: Estrogen has a negative effect on recovery, so women are less likely to recover than men, regardless of the drug. In addition, as we can see from the data, women are significantly *more* likely to take the drug than men are. So, the reason the drug appears to be harmful overall is that, if we select a drug user at random, that person is more likely to be a woman and hence less likely to recover than a random person who does not take the drug. Put differently, being a woman is a common cause of both drug taking and failure to recover. Therefore, to assess the effectiveness, we need to compare subjects of the same gender, thereby ensuring that any difference in recovery rates between those who take the drug and those who do not is not ascribable to estrogen. This means we should consult the segregated data, which shows us unequivocally that the drug is helpful. This matches our intuition, which tells us that the segregated data is "more specific," hence more informative, than the unsegregated data.

With a few tweaks, we can see how the same reversal can occur in a continuous example. Consider a study that measures weekly exercise and cholesterol in various age groups. When we plot exercise on the X-axis and cholesterol on the Y-axis and segregate by age, as in Figure 1.1, we see that there is a general trend downward in each group; the more young people exercise, the lower their cholesterol is, and the same applies for middle-aged people and the elderly. If, however, we use the same scatter plot, but we don't segregate by gender (as in Figure 1.2), we see a general trend upward; the more a person exercises, the higher their cholesterol is. To resolve this problem, we once again turn to the story behind the data. If we know that older people, who are more likely to exercise (Figure 1.1), are also more likely to have high cholesterol regardless of exercise, then the reversal is easily explained, and easily resolved. Age is a common cause of both treatment (exercise) and outcome (cholesterol). So we should look at the age-segregated data in order to compare same-age people and thereby eliminate the possibility that the high exercisers in each group we examine are more likely to have high cholesterol due to their age, and not due to exercising.

However, and this might come as a surprise to some readers, segregated data does not always give the correct answer. Suppose we looked at the same numbers from our first example of drug taking and recovery, instead of recording participants' gender, patients' blood pressure were

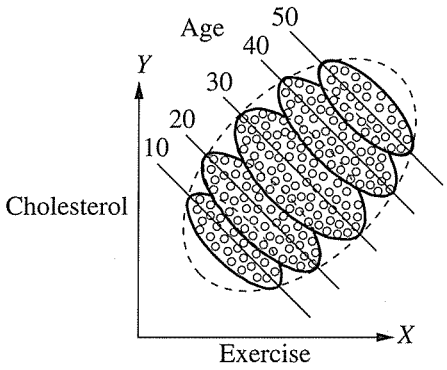


Figure 1.1 Results of the exercise–cholesterol study, segregated by age

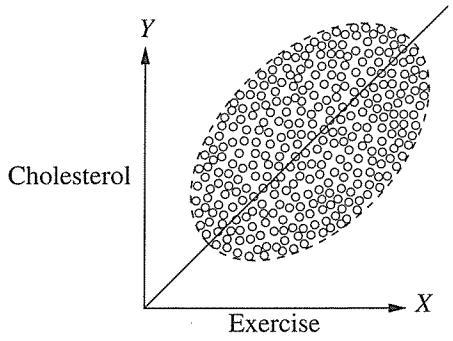


Figure 1.2 Results of the exercise–cholesterol study, unsegregated. The data points are identical to those of Figure 1.1, except the boundaries between the various age groups are not shown

recorded at the end of the experiment. In this case, we know that the drug affects recovery by lowering the blood pressure of those who take it—but unfortunately, it also has a toxic effect. At the end of our experiment, we receive the results shown in Table 1.2. (Table 1.2 is numerically identical to Table 1.1, with the exception of the column labels, which have been switched.)

Now, would you recommend the drug to a patient?

Once again, the answer follows from the way the data were generated. In the general population, the drug might improve recovery rates because of its effect on blood pressure. But in the subpopulations—the group of people whose posttreatment BP is high and the group whose posttreatment BP is low—we, of course, would not see that effect; we would only see the drug's toxic effect.

As in the gender example, the purpose of the experiment was to gauge the overall effect of treatment on rates of recovery. But in this example, since lowering blood pressure is one of the mechanisms by which treatment affects recovery, it makes no sense to separate the results based on blood pressure. (If we had recorded the patients' blood pressure *before* treatment, and if it were BP that had an effect on treatment, rather than the other way around, it would be a different story.) So we consult the results for the general population, we find that treatment increases the probability of recovery, and we decide that we *should* recommend treatment. Remarkably, though the numbers are the same in the gender and blood pressure examples, the correct result lies in the segregated data for the former and the aggregate data for the latter.

None of the information that allowed us to make a treatment decision—not the timing of the measurements, not the fact that treatment affects blood pressure, and not the fact that blood

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

pressure affects recovery—was found in the data. In fact, as statistics textbooks have traditionally (and correctly) warned students, correlation is not causation, so there is no statistical method that can determine the causal story from the data alone. Consequently, there is no statistical method that can aid in our decision.

Yet statisticians interpret data based on causal assumptions of this kind all the time. In fact, the very paradoxical nature of our initial, qualitative, gender example of Simpson's problem is derived from our strongly held conviction that treatment cannot affect sex. If it could, there would be no paradox, since the causal story behind the data could then easily assume the same structure as in our blood pressure example. Trivial though the assumption "treatment does not cause sex" may seem, there is no way to test it in the data, nor is there any way to represent it in the mathematics of standard statistics. There is, in fact, no way to represent *any* causal information in contingency tables (such as Tables 1.1 and 1.2), on which statistical inference is often based.

There are, however, *extra*-statistical methods that can be used to express and interpret causal assumptions. These methods and their implications are the focus of this book. With the help of these methods, readers will be able to mathematically describe causal scenarios of any complexity, and answer decision problems similar to those posed by Simpson's paradox as swiftly and comfortably as they can solve for X in an algebra problem. These methods will allow us to easily distinguish each of the above three examples and move toward the appropriate statistical analysis and interpretation. A calculus of causation composed of simple logical operations will clarify the intuitions we already have about the nonexistence of a drug that cures men and women but hurts the whole population and about the futility of comparing patients with equal blood pressure. This calculus will allow us to move beyond the toy problems of Simpson's paradox into intricate problems, where intuition can no longer guide the analysis. Simple mathematical tools will be able to answer practical questions of policy evaluation as well as scientific questions of how and why events occur.

But we're not quite ready to pull off such feats of derring-do just yet. In order to rigorously approach our understanding of the causal story behind data, we need four things:

1. A working definition of "causation."
2. A method by which to formally articulate causal assumptions—that is, to create causal models.
3. A method by which to link the structure of a causal model to features of data.
4. A method by which to draw conclusions from the combination of causal assumptions embedded in a model and data.

The first two parts of this book are devoted to providing methods for modeling causal assumptions and linking them to data sets, so that in the third part, we can use those assumptions and data to answer causal questions. But before we can go on, we must define causation. It may seem intuitive or simple, but a commonly agreed-upon, completely encompassing definition of causation has eluded statisticians and philosophers for centuries. For our purposes, the definition of causation is simple, if a little metaphorical: A variable X is a *cause* of a variable Y if Y in any way relies on X for its value. We will expand slightly upon this definition later, but for now, think of causation as a form of listening; X is a cause of Y if Y listens to X and decides its value in response to what it hears.

Readers must also know some elementary concepts from probability, statistics, and graph theory in order to understand the aforementioned causal methods. The next two sections

will therefore provide the necessary definitions and examples. Readers with a basic understanding of probability, statistics, and graph theory may skip to Section 1.5 with no loss of understanding.

Study questions

Study question 1.2.1

What is wrong with the following claims?

(a) *"Data show that income and marriage have a high positive correlation. Therefore, your earnings will increase if you get married."*
(b) *"Data show that as the number of fires increase, so does the number of fire fighters. Therefore, to cut down on fires, you should reduce the number of fire fighters."*
(c) *"Data show that people who hurry tend to be late to their meetings. Don't hurry, or you'll be late."*

Study question 1.2.2

A baseball batter Tim has a better batting average than his teammate Frank. However, someone notices that Frank has a better batting average than Tim against both right-handed and left-handed pitchers. How can this happen? (Present your answer in a table.)

Study question 1.2.3

Determine, for each of the following causal stories, whether you should use the aggregate or the segregated data to determine the true effect.

(a) *There are two treatments used on kidney stones: Treatment A and Treatment B. Doctors are more likely to use Treatment A on large (and therefore, more severe) stones and more likely to use Treatment B on small stones. Should a patient who doesn't know the size of his or her stone examine the general population data, or the stone size-specific data when determining which treatment will be more effective?*
(b) *There are two doctors in a small town. Each has performed 100 surgeries in his career, which are of two types: one very difficult surgery and one very easy surgery. The first doctor performs the easy surgery much more often than the difficult surgery and the second doctor performs the difficult surgery more often than the easy surgery. You need surgery, but you do not know whether your case is easy or difficult. Should you consult the success rate of each doctor over all cases, or should you consult their success rates for the easy and difficult cases separately, to maximize the chance of a successful surgery?*

Study question 1.2.4

In an attempt to estimate the effectiveness of a new drug, a randomized experiment is conducted. In all, 50% of the patients are assigned to receive the new drug and 50% to receive a placebo. A day before the actual experiment, a nurse hands out lollipops to some patients who

show signs of depression, mostly among those who have been assigned to treatment the next day (i.e., the nurse's round happened to take her through the treatment-bound ward). Strangely, the experimental data revealed a Simpson's reversal: Although the drug proved beneficial to the population as a whole, drug takers were less likely to recover than nontakers, among both lollipop receivers and lollipop nonreceivers. Assuming that lollipop sucking in itself has no effect whatsoever on recovery, answer the following questions:

(a) Is the drug beneficial to the population as a whole or harmful?
(b) Does your answer contradict our gender example, where sex-specific data was deemed more appropriate?
(c) Draw a graph (informally) that more or less captures the story. (Look ahead to Section 1.4 if you wish.)
(d) How would you explain the emergence of Simpson's reversal in this story?
(e) Would your answer change if the lollipops were handed out (by the same criterion) a day after the study?

[Hint: Use the fact that receiving a lollipop indicates a greater likelihood of being assigned to drug treatment, as well as depression, which is a symptom of risk factors that lower the likelihood of recovery.]

1.3 Probability and Statistics

Since statistics generally concerns itself not with absolutes but with likelihoods, the language of probability is extremely important to it. Probability is similarly important to the study of causation because most causal statements are uncertain (e.g., "careless driving causes accidents," which is true, but does not mean that a careless driver is certain to get into an accident), and probability is the way we express uncertainty. In this book, we will use the language and laws of probability to express our beliefs and uncertainty about the world. To aid readers without a strong background in probability, we provide here a glossary of the most important terms and concepts they will need to know in order to understand the rest of the book.

1.3.1 Variables

A *variable* is any property or descriptor that can take multiple values. In a study that compares the health of smokers and nonsmokers, for instance, some variables might be the age of the participant, the gender of the participant, whether or not the participant has a family history of cancer, and how many years the participant has been smoking. A variable can be thought of as a question, to which the value is the answer. For instance, "How old is this participant?" "38 years old." Here, "age" is the variable, and "38" is its value. The probability that variable X takes value x is written $P(X = x)$. This is often shortened, when context allows, to $P(x)$. We can also discuss the probability of multiple values at once; for instance, the probability that $X = x$ and $Y = y$ is written $P(X = x, Y = y)$, or $P(x, y)$. Note that $P(X = 38)$ is specifically interpreted as the probability that an individual randomly selected from the population is aged 38.

A variable can be either *discrete* or *continuous*. Discrete variables (sometimes called *categorical* variables) can take one of a finite or countably infinite set of values in any range. A variable describing the state of a standard light switch is discrete, because it has two values: "on"

and "off." Continuous variables can take any one of an infinite set of values on a continuous scale (i.e., for any two values, there is some third value that lies between them). For instance, a variable describing in detail a person's weight is continuous, because weight is measured by a real number.

1.3.2 Events

An *event* is any assignment of a value or set of values to a variable or set of variables. "$X = 1$" is an event, as is "$X = 1$ or $X = 2$," as is "$X = 1$ and $Y = 3$," as is "$X = 1$ or $Y = 3$." "The coin flip lands on heads," "the subject is older than 40," and "the patient recovers" are all events. In the first, "outcome of the coin flip" is the variable, and "heads" is the value it takes. In the second, "age of the subject" is the variable and "older than 40" describes a set of values it may take. In the third, "the patient's status" is the variable and "recovery" is the value. This definition of "event" runs counter to our everyday notion, which requires that some change occur. (For instance, we would not, in everyday conversation, refer to a person being a certain age as an event, but we would refer to that person *turning* a year older as such.) Another way of thinking of an event in probability is this: Any declarative statement (a statement that can be true or false) is an event.

Study questions

Study question 1.3.1

Identify the variables and events invoked in the lollipop story of Study question 1.2.4

1.3.3 Conditional Probability

The probability that some event A occurs, given that we know some other event B has occurred, is the *conditional probability of A given B*. The conditional probability that $X = x$, given that $Y = y$, is written $P(X = x|Y = y)$. As with unconditional probabilities, this is often shortened to $P(x|y)$. Often, the probability that we assign to the event "$X = x$" changes drastically, depending on the knowledge "$Y = y$" that we condition on. For instance, the probability that you have the flu right now is fairly low. But, that probability would become much higher if you were to take your temperature and discover that it is 102 °F.

When dealing with probabilities represented by frequencies in a data set, one way to think of conditioning is *filtering* a data set based on the value of one or more variables. For instance, suppose we looked at the ages of U.S. voters in the last presidential election. According to the Census Bureau, we might get the data set shown in Table 1.3.

In Table 1.3, there were 132,948,000 votes cast in total, so we would estimate that the probability that a given voter was younger than the age of 45 is

$$P(\textit{Voter's Age} < 45) = \frac{20{,}539{,}000 + 30{,}756{,}000}{132{,}448{,}000} = \frac{51{,}295{,}000}{132{,}948{,}000} = 0.38$$

Suppose, however, we want to estimate the probability that a voter was younger than the age of 45, given that we *know* he was elder than the age of 29. To find this out, we simply filter

Table 1.3 Age breakdown of voters in 2012 election (all numbers in thousands)

Age group	# of voters
18–29	20,539
30–44	30,756
45–64	52,013
65+	29,641
	132,948

Table 1.4 Age breakdown of voters over the age of 29 in 2012 election (all numbers in thousands)

Age group	# of voters
30–44	30,756
45–64	52,013
65+	29,641
	112,409

the data to form a new set (shown in Table 1.4), using only the cases where voters were elder than 29.

In this new data set, there are 112,409,000 total votes, so we would estimate that

$$P(Voter\ Age < 45 | Voter\ Age > 29) = \frac{30{,}756{,}000}{112{,}409{,}000} = 0.27$$

Conditional probabilities such as these play an important role in investigating causal questions, as we often want to compare how the probability (or, equivalently, risk) of an outcome changes under different filtering, or exposure, conditions. For example, how does the probability of developing lung cancer for smokers compare to the analogous probability for nonsmokers?

Study questions

Study question 1.3.2

Consider Table 1.5 showing the relationship between gender and education level in the U.S. adult population.

(a) Estimate P(High School).
(b) Estimate P(High School OR Female).
(c) Estimate P(High School | Female).
(d) Estimate P(Female | High School).

Table 1.5 The proportion of males and females achieving a given education level

Gender	Highest education achieved	Occurrence (in hundreds of thousands)
Male	Never finished high school	112
Male	High school	231
Male	College	595
Male	Graduate school	242
Female	Never finished high school	136
Female	High school	189
Female	College	763
Female	Graduate school	172

1.3.4 Independence

It might happen that the probability of one event remains unaltered with the observation of another. For example, while observing your high temperature increases the probability that you have the flu, observing that your friend Joe is 38 years old does not change the probability at all. In cases such as this, we say that the two events are *independent*. Formally, events A and B are said to be independent if

$$P(A|B) = P(A) \tag{1.1}$$

that is, the knowledge that B has occurred gives us no additional information about the probability of A occurring. If this equality does not hold, then A and B are said to be *dependent*. Dependence and independence are symmetric relations—if A is dependent on B, then B is dependent on A, and if A is independent of B, then B is independent of A. (Formally, if $P(A|B) = P(A)$, then it *must* be the case that $P(B|A) = P(B)$.) This makes intuitive sense; if "smoke" tells us something about "fire," then "fire" must tell us something about "smoke."

Two events A and B are *conditionally independent* given a third event C if

$$P(A|B, C) = P(A|C) \tag{1.2}$$

and $P(B|A, C) = P(B|C)$. For example, the event "smoke detector is on" is dependent on the event "there is a fire nearby." But these two events may become independent conditional on the third event "there is smoke nearby"; smoke detectors respond to the presence of smoke only, not to its cause. When dealing with data sets, or probability tables, A and B are conditionally independent given C if A and B are independent in the new data set created by filtering on C. If A and B are independent in the original unfiltered data set, they are called *marginally independent*.

Variables, like events, can be dependent or independent of each other. Two variables X and Y are considered independent if for every value x and y that X and Y can take, we have

$$P(X = x|Y = y) = P(X = x) \tag{1.3}$$

(As with independence of events, independence of variables is a symmetrical relation, so it follows that Eq. (1.3) implies $P(Y = y|X = x) = P(Y = y)$.) If for any pair of values of X and Y,

this equality does not hold, then X and Y are said to be *dependent*. In this sense, independence of variables can be understood as a set of independencies of events. For instance, "height" and "musical talent" are independent variables; for every height h and level of musical talent m, the probability that a person is h feet high would not change upon discovering that he/she has m amount of talent.

1.3.5 Probability Distributions

A *probability distribution* for a variable X is the set of probabilities assigned to each possible value of X. For instance, if X can take three values—1, 2, and 3—a possible probability distribution for X would be "$P(X = 1) = 0.5, P(X = 2) = 0.25, P(X = 3) = 0.25$." The probabilities in a probability distribution must lie between 0 and 1, and must sum to 1. An event with probability 0 is impossible; an event with probability 1 is certain.

Continuous variables also have probability distributions. The probability distribution of a continuous variable X is represented by a function f, called the *density function*. When f is plotted on a coordinate plane, the probability that the value of variable X lies between values a and b is the area under the curve between a and b—or, as those who have taken calculus will know, $\int_a^b f(x)dx$. The area under the entire curve—that is, $\int_{-\infty}^{\infty} f(x)dx$—must of course be equal to 1.

Sets of variables can also have probability distributions, called *joint distributions*. The joint distribution of a set of variables V is the set of probabilities of each possible combination of variable values in V. For instance, if V is a set of two variables—X and Y—each of which can take two values—1 and 2—then one possible joint distribution for V is "$P(X = 1, Y = 1) = 0.2, P(X = 1, Y = 2) = 0.1, P(X = 2, Y = 1) = 0.5, P(X = 2, Y = 2) = 0.2$." Just as with single-variable distributions, probabilities in a joint distribution must sum to 1.

1.3.6 The Law of Total Probability

There are several universal probabilistic truths that are useful to know. First, for any two mutually exclusive events A and B (i.e., A and B cannot co-occur), we have

$$P(A \text{ or } B) = P(A) + P(B) \tag{1.4}$$

It follows that, for any two events A and B, we have

$$P(A) = P(A, B) + P(A, \text{"not } B\text{"}) \tag{1.5}$$

because the events "A and B" and "A and 'not B'" are mutually exclusive—and because if A is true, then either "A and B" or "A and 'not B'" must be true. For example, "Dana is a tall man" and "Dana is a tall woman" are mutually exclusive, and if Dana is tall, then he or she must be either a tall man or a tall woman; therefore, $P(\textit{Dana is tall}) = P(\textit{"Dana is a tall man"}) + P(\textit{"Dana is a tall woman"})$.

More generally, for any set of events $B_1, B_2, \ldots, B_n$ such that exactly one of the events must be true (an exhaustive, mutually exclusive set, called a *partition*), we have

$$P(A) = P(A, B_1) + P(A, B_2) + \cdots + P(A, B_n) \tag{1.6}$$

This rule, known as the *law of total probability*, becomes somewhat obvious as soon as we put it in real-world terms: If we pull a random card from a standard deck, the probability that the card is a Jack will be equal to the probability that it's a Jack *and* a spade, plus the probability that it's a Jack *and* a heart, plus the probability that it's a Jack *and* a club, plus the probability that it's a Jack *and* a diamond. Calculating the probability of an event A by summing up its probabilities over all B_i is called *marginalizing over* B, and the resulting probability $P(A)$ is called the *marginal probability* of A.

If we know the probability of B and the probability of A conditional on B, we can deduce the probability of A and B by simple multiplication:

$$P(A, B) = P(A|B)P(B) \tag{1.7}$$

For instance, the probability that Joe is funny and smart is equal to the probability that a smart person is funny, multiplied by the probability that Joe is smart. The division rule

$$P(A|B) = P(A, B)/P(B)$$

which is formally regarded as a definition of conditional probabilities, is justified by viewing conditioning as a filtering operation, as we have done in Tables 1.3 and 1.4. When we condition on B, we remove from the table all events that conflict with B. The resulting subtable, like the original, represents a probability distribution, and like all probability distributions, it must sum to one. Since the probabilities of the subtables rows in the original distribution summed to $P(B)$ (by definition), we can determine their probabilities in the new distribution by multiplying each by $1/P(B)$.

Equation (1.7) implies that the notion of independence, which until now we have used informally to mean "giving no additional information," has a numerical representation in the probability distribution. In particular, for events A and B to be independent, we require that

$$P(A, B) = P(A)P(B)$$

For example, to check if the outcomes of two coins are truly independent, we should count the frequency at which both show up tails, and make sure that it equals the product of the frequencies at which each of the coins shows up tails.

Using (1.7) together with the symmetry $P(A, B) = P(B, A)$, we can immediately obtain one of the most important laws of probability, *Bayes' rule*:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.8}$$

With the help of the multiplication rule in (1.7), we can express the law of total probability as a weighted sum of conditional probabilities:

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_k)P(B_k) \tag{1.9}$$

This is very useful, because often we will find ourselves in a situation where we cannot assess $P(A)$ directly, but we can through this decomposition. It is generally easier to assess conditional probabilities such as $P(A|B_k)$, which are tied to specific contexts, rather than $P(A)$, which is not attached to a context. For instance, suppose we have a stock of gadgets from two sources: 30% of them are manufactured by factory A, in which one out of 5000 is defective, whereas 70%

are manufactured by factory B, in which one out of 10,000 is defective. To find the probability that a randomly chosen gadget will be defective is not a trivial mental task, but when broken down according to Eq. (1.9) it becomes easy:

$$\begin{aligned} P(\textit{defective}) &= P(\textit{defective}|A)P(A) + P(\textit{defective}|B)P(B) \\ &= \frac{0.30}{5{,}000} + \frac{0.70}{10{,}000} \\ &= \frac{1.30}{10{,}000} = 0.00013 \end{aligned}$$

Or, to take a somewhat harder example, suppose we roll two dice, and we want to know the probability that the second roll is higher than the first, $P(A) = P(\textit{Roll}\ 2 > \textit{Roll}\ 1)$. There is no obvious way to calculate this probability all at once. But if we break it down into contexts $B_1, \ldots, B_6$ by conditioning on the value of the first die, it becomes easy to solve:

$$\begin{aligned} P(\textit{Roll}\ 2 > \textit{Roll}\ 1) &= P(\textit{Roll}\ 2 > \textit{Roll}\ 1 | \textit{Roll}\ 1 = 1)P(\textit{Roll}\ 1 = 1) \\ &\quad + P(\textit{Roll}\ 2 > \textit{Roll}\ 1 | \textit{Roll}\ 1 = 2)P(\textit{Roll}\ 1 = 2) \\ &\quad + \cdots + P(\textit{Roll}\ 2 > \textit{Roll}\ 1 | \textit{Roll}\ 1 = 6) \times P(\textit{Roll}\ 1 = 6) \\ &= \left(\frac{5}{6} \times \frac{1}{6}\right) + \left(\frac{4}{6} \times \frac{1}{6}\right) + \left(\frac{3}{6} \times \frac{1}{6}\right) + \left(\frac{2}{6} \times \frac{1}{6}\right) + \left(\frac{1}{6} \times \frac{1}{6}\right) + \left(\frac{0}{6} \times \frac{1}{6}\right) \\ &= \frac{5}{12} \end{aligned}$$

The decomposition described in Eq. (1.9) is sometimes called "the law of alternatives" or "extending the conversation"; in this book, we will refer to it as *conditionalizing on B*.

1.3.7 *Using Bayes' Rule*

When using Bayes' rule, we sometimes loosely refer to event A as the "hypothesis" and event B as the "evidence." This naming reflects the reason that Bayes' theorem is so important: In many cases, we know or can easily determine $P(B|A)$ (the probability that a piece of evidence will occur, given that our hypothesis is correct), but it's much harder to figure out $P(A|B)$ (the probability of the hypothesis being correct, given that we obtain a piece of evidence). Yet the latter is the question that we most often want to answer in the real world; generally, we want to update our belief in some hypothesis, $P(A)$, after some evidence B has occurred, to $P(A|B)$. To precisely use Bayes' rule in this manner, we must treat each hypothesis as an event and assign to all hypotheses for a given situation a probability distribution, called a *prior*.

For example, suppose you are in a casino, and you hear a dealer shout "11!" You happen to know that the only two games played at the casino that would occasion that event are craps and roulette and that there are exactly as many craps games as roulette games going on at any moment. What is the probability that the dealer is working at a game of craps, given that he shouted "11?"

In this case, "craps" is our hypothesis, and "11" is our evidence. It's difficult to figure out this probability off-hand. But the reverse—the probability that an 11 will result in a given round of craps—is easy to calculate; it is specified by the game. Craps is a game in which gamblers bet

on the sum of a roll of two dice. So 11 will be the sum in $\frac{2}{36} = \frac{1}{18}$ of cases: $P(\text{“11”}|\textit{“craps”}) = \frac{1}{18}$. In roulette, there are 38 equally probable outcomes, so $P(\text{“11”}|\textit{“roulette”}) = \frac{1}{38}$. In this situation, there are two possible hypotheses; “craps” and “roulette.” Since there are an equal number of craps and roulette games, $P\textit{“craps”}(\textit{“craps”}) = \frac{1}{2}$, our prior belief before we hear the “11” shout. Using, the law of total probability,

$$\begin{aligned} P(\text{“11”}) &= P(\text{“11”}|\textit{“craps”})P(\textit{“craps”}) + P(\text{“11”}|\textit{“roulette”})P(\textit{“roulette”}) \\ &= \frac{1}{2} \times \frac{1}{18} + \frac{1}{2} \times \frac{1}{38} = \frac{7}{171} \end{aligned}$$

We have now fairly easily obtained all the information we need to determine $P(\textit{“craps”}|\text{“11”})$:

$$P(\textit{“craps”}|\text{“11”}) = \frac{P(\text{“11”}|\textit{“craps”}) \times P(\textit{“craps”})}{P(\text{“11”})} = \frac{1/18 \times 1/2}{7/171} = 0.679$$

Another informative example of Bayes' rule in action is the Monty Hall problem, a classic brain teaser in statistics. In the problem, you are a contestant on a game show, hosted by Monty Hall. Monty shows you three doors—A, B, and C—behind one and only one of which is a new car. (The other two doors have goats.) If you guess correctly, the car is yours; otherwise, you get a goat. You guess A at random. Monty, who is forbidden from revealing where the car is, then opens Door C, which, of course, has a goat behind it. He tells you that you can now switch to Door B, or stick with Door A. Whichever you pick, you'll get what's behind it.

Are you better off opening Door A, or switching to Door B?

Many people, when they first encounter the problem, reason that, since the location of the car is independent of the door you first choose, switching doors neither gains nor loses you anything; the probability that the car is behind Door A is equal to the probability that it is behind Door B.

But the correct answer, as decades of statistics students have found to their consternation, is that you are twice as likely to win the car if you switch to Door B as you are if you stay with Door A. The reasoning often given for this counterintuitive solution is that, when you originally chose a door, you had a $\frac{1}{3}$ probability of picking the door with the car. Since Monty *always* opens a door with a goat, no matter whether you initially chose the car or not, you have received no new information since then. Therefore, there is still a $\frac{1}{3}$ probability that the door you picked hides the car, and the remaining $\frac{2}{3}$ probability must lie with the only other closed door left.

We can prove this surprising fact using Bayes' rule. Here we have three variables: X, the door chosen by the player; Y, the door behind which the car is hidden; and Z, the door which the host opens. X, Y, and Z can all take the values A, B, or C. We want to prove that $P(Y = B|X = A, Z = C) > P(Y = A|X = A, Z = C)$. Our hypothesis is that the car lies behind Door A; our evidence is that Monty opened Door C. We will leave the proof to the reader—see Study question 1.3.5. To further develop your intuition, you might generalize the game to having 100 doors (which contain 1 hidden car and 99 hidden goats). The contestant still chooses one door, but now Monty opens 98 doors—all revealing goats deliberately—before offering the contestant the chance to switch before the final doors are opened. Now, the choice to switch should be obvious.

Why does Monty opening Door C constitute evidence about the location of the car? It didn't, after all, provide any evidence for whether your initial choice of door was correct. And, surely,

when he was about to open a door, be it B or C, you knew in advance that you won't find a car behind it. The answer is that there was no way for Monty to open Door A after you chose it—but he *could* have opened Door B. The fact that he didn't make it more likely that he opened Door C because he was forced to; it provides evidence that the car lies behind Door B. This is a general theme of Bayesian analysis: Any hypothesis that has withstood some test of refutation becomes more likely. Door B was vulnerable to refutation (i.e., Monty could have opened it), but Door A was not. Therefore, Door B becomes a more likely location, whereas Door A does not.

The reader may find it instructive to note that the explanation above is laden with counterfactual terminology; for example, "He could have opened," "because he was forced," "He was about to open." Indeed, what makes the Monty Hall example unique among probability puzzles is its critical dependence on the process that generated the data. It shows that our beliefs should depend not merely on the facts observed but also on the process that led to those facts. In particular, the information that the car is not behind Door C, in itself, is not sufficient to describe the problem; to figure out the probabilities involved, we must also know what options were available to the host before opening Door C. In Chapter 4 of this book we will formulate a theory of counterfactuals that will enable us to describe such processes and alternative options, so as to form the correct beliefs about choices.

There is some controversy attached to Bayes' rule. Often, when we are trying to ascertain the probability of a hypothesis given some evidence, we have no way to calculate the prior probability of the hypothesis, $P(A)$, in terms of fractions or frequencies of cases. Consider: If we did not know the proportion of roulette tables to craps tables in the casino, how on Earth could we determine the prior probability $P(\textit{"craps"})$? We might be tempted to postulate $P(A) = \frac{1}{2}$ as a way of expressing our ignorance. But what if we have a hunch that roulette tables are less common in this casino, or the tone of the voice of the caller reminds us of a craps dealer we heard yesterday? In cases such as this, in order to use Bayes' rule, we substitute, in place of $P(A)$, our *subjective belief* in the relative truth of the hypothesis compared to other possibilities. The controversy stems from the subjective nature of that belief—how are we to know whether the assigned $P(A)$ accurately summarizes the information we have about the hypothesis? Should we insist on distilling all of our pro and con arguments down to a single number? And even if we do, why should we update our subjective beliefs about hypotheses the same way that we update objective frequencies? Some behavioral experiments suggest that people do not update their beliefs in accordance with Bayes' rule—but many believe that they *should*, and that deviations from the rule represent compromises, if not deficiencies in reasoning, and lead to suboptimal decisions. Debate over the proper use of Bayes' theorem continues to this day. Despite these controversies, however, Bayes' rule is a powerful tool for statistics, and we will use it to great effect throughout this book.

Study questions

Study question 1.3.3

Consider the casino problem described in Section 1.3.6

(a) Compute $P(\text{"craps"} \mid \textit{"11"})$ *assuming that there are twice as many roulette tables as craps games at the casino.*

(b) *Compute P("roulette"|"10") assuming that there are twice as many craps games as roulette tables at the casino.*

Study question 1.3.4

Suppose we have three cards. Card 1 has two black faces, one on each side; Card 2 has two white faces; and Card 3 has one white face and one back face. You select a card at random and place it on the table. You find that it is black on the face-up side. What is the probability that the face-down side of the card is also black?

(a) *Use your intuition to argue that the probability that the face-down side of the card is also black is $\frac{1}{2}$. Why might it be greater than $\frac{1}{2}$?*
(b) *Express the probabilities and conditional probabilities that you find easy to estimate (for example, $P(C_D = \text{Black})$), in terms of the following variables:*

I = Identity of the card selected (Card 1, Card 2, or Card 3)

C_D = Color of the face-down side (Black, White)

C_U = Color of the face-up side (Black, White)

Find the probability that the face-down side of the selected card is black, using your estimates above.
(c) *Use Bayes' theorem to find the correct probability of a randomly selected card's back being black if you observe that its front is black?*

Study question 1.3.5 (Monty Hall)

Prove, using Bayes' theorem, that switching doors improves your chances of winning the car in the Monty Hall problem.

1.3.8 Expected Values

In statistics, one often deals with data sets and probability distributions that are too large to effectively examine each possible combination of values. Instead, we use statistical measures to represent, with some loss of information, meaningful features of the distribution. One such measure is the *expected value*, also called the *mean*, which can be used when variables take on numerical values. The expected value of a variable X, denoted $E(X)$, is found by multiplying each possible value of the variable by the probability that the variable will take that value, then summing the products:

$$E(X) = \sum_x x\,P(X = x) \tag{1.10}$$

For instance, a variable X representing the outcome of one roll of a fair six-sided die has the following probability distribution: $P(1) = \frac{1}{6}, P(2) = \frac{1}{6}, P(3) = \frac{1}{6}, P(4) = \frac{1}{6}, P(5) = \frac{1}{6}, P(6) = \frac{1}{6}$. The expected value of X is given by:

$$E(X) = \left(1 \times \frac{1}{6}\right) + \left(2 \times \frac{1}{6}\right) + \left(3 \times \frac{1}{6}\right) + \left(4 \times \frac{1}{6}\right) + \left(5 \times \frac{1}{6}\right) + \left(6 \times \frac{1}{6}\right) = 3.5$$

Similarly, the expected value of any function of X—say, $g(X)$—is obtained by summing $g(x)P(X = x)$ over all values of X.

$$E[g(X)] = \sum_x g(x)P(x) \tag{1.11}$$

For example, if after rolling a die, I receive a cash prize equal to the square of the result, we have $g(X) = X^2$, and the expected prize is

$$E[g(X)] = \left(1^2 \times \frac{1}{6}\right) + \left(2^2 \times \frac{1}{6}\right) + \left(3^2 \times \frac{1}{6}\right) + \left(4^2 \times \frac{1}{6}\right) + \left(5^2 \times \frac{1}{6}\right) + \left(6^2 \times \frac{1}{6}\right) = 15.17 \tag{1.12}$$

We can also calculate the expected value of Y conditional on X, $E(Y|X = x)$, by multiplying each possible value y of Y by $P(Y = y|X = x)$, and summing the products.

$$E(Y|X = x) = \sum_y y\, P(Y = y|X = x) \tag{1.13}$$

$E(X)$ is one way to make a "best guess" of X's value. Specifically, out of all the guesses g that we can make, the choice "$g = E(X)$" minimizes the expected square error $E(g - X)^2$. Similarly, $E(Y|X = x)$ represents a best guess of Y, given that we observe $X = x$. If $g = E(Y|X = x)$, then g minimizes the expected square error $E[(g - Y)^2|X = x)]$.

For example, the expected age of a 2012 voter, as demonstrated by Table 1.3, is

$$E(\textit{Voter's Age}) = 23.5 \times 0.16 + 37 \times 0.23 + 54.5 \times 0.39 + 70 \times 0.22 = 48.9$$

(For this calculation, we have assumed that every age within each category is equally likely, e.g., a voter is as likely to be 18 as 25, and as likely to be 30 as 44. We have also assumed that the oldest age of any voter is 75.) This means that if we were asked to guess the age of a randomly chosen voter, with the understanding that if we were off by e years, we would lose e^2 dollars, we would lose the least money, on average, if we guessed 48.9. Similarly, if we were asked to guess the age of a random voter younger than the age of 45, our best bet would be

$$E[\textit{Voter's Age} \mid \textit{Voter's Age} < 45] = 23.5 \times 0.40 + 37 \times 0.60 = 31.6 \tag{1.14}$$

The use of expectations as a basis for predictions or "best guesses" hinges to a great extent or an implicit assumption regarding the distribution of X or $Y|X = x$, namely that such distributions are approximately *symmetric*. If, however, the distribution of interest is highly *skewed*, other methods of prediction may be better. In such cases, for example, we might use the median of the distribution of X as our "best guess"; this estimate minimizes the expected absolute error $E(|g - X|)$. We will not pursue such alternative measures further here.

1.3.9 Variance and Covariance

The *variance* of a variable X, denoted $Var(X)$ or σ_X^2, is a measure of roughly how "spread out" the values of X in a data set or population are from their mean. If the values of X all hover close

to one value, the variance will be relatively small; if they cover a large range, the variance will be comparatively large. Mathematically, we define the variance of a variable as the average square difference of that variable from its mean. It can be computed by first finding its mean, μ, and then calculating

$$Var(X) = E((X - \mu)^2) \tag{1.15}$$

The *standard deviation* σ_X of a random variable X is the square root of its variance. Unlike the variance, σ_X is expressed in the same units as X. For example, the variance of under-45 voters' age distribution, according to Table 1.3, can easily be calculated to be (Eq. (1.14)):

$$\begin{aligned} Var(X) &= ((23.5 - 31.5)^2 \times 0.41) + ((37 - 31.5)^2 \times 0.59) \\ &= (64 \times 0.41) + (30.25 \times .59) \\ &= 26.24 + 17.85 = 43.09 \text{ years}^2 \end{aligned}$$

while the standard deviation is

$$\sigma_X = \sqrt{(43.09)} = 6.56 \text{ years}$$

This means that, choosing a voter at random, chances are high that his/her age will fall less than 6.56 years away from the average 31.5. This kind of interpretation can be quantified. For example, for a normally distributed random variable X, approximately two-thirds of the population values of X fall within *one* standard deviation of the expectation, or mean. Further, about 95% fall within *two* standard deviations from the mean.

Of special importance is the expectation of the product $(X - E(X))(Y - E(Y))$, which is known as the *covariance* of X and Y,

$$\sigma_{XY} \triangleq E[(X - E(X))(Y - E(Y))] \tag{1.16}$$

It measures the degree to which X and Y *covary*, that is, the degree to which the two variables vary together, or are "associated." This measure of association actually reflects a specific way in which X and Y covary; it measures the extent to which X and Y *linearly* covary. You can think of this as plotting Y versus X and considering the extent to which a straight *line* captures the way in which Y varies as X changes.

The covariance σ_{XY} is often normalized to yield the *correlation coefficient*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{1.17}$$

which is a dimensionless number ranging from -1 to 1, which represents the slope of the best-fit line after we normalize both X and Y by their respective standard deviations. ρ_{XY} is one if and only if one variable can predict the other in a *linear* fashion, and it is zero whenever such a linear prediction is no better than a random guess. The significance of σ_{XY} and ρ_{XY} will be discussed in the next section. At this point, it is sufficient to note that these degrees of covariation can be readily computed from the joint distribution $P(x, y)$, using Eqs. (1.16) and (1.17). Moreover, both σ_{XY} and ρ_{XY} vanish when X and Y are independent. Note that nonlinear relationships between Y and X cannot naturally be captured by a simple numerical summary; they require a full specification of the conditional probability $P(Y = y|X = x)$.

Study questions

Study question 1.3.6

(a) Prove that, in general, both σ_{XY} and ρ_{XY} vanish when X and Y are independent. [Hint: Use Eqs. (1.16) and (1.17).]
(b) Give an example of two variables that are highly dependent and, yet, their correlation coefficient vanishes.

Study question 1.3.7

Two fair coins are flipped simultaneously to determine the payoffs of two players in the town's casino. Player 1 wins a dollar if and only if at least one coin lands on head. Player 2 receives a dollar if and only if the two coins land on the same face. Let X stand for the payoff of Player 1 and Y for the payoff of Player 2.

(a) Find and describe the probability distributions

$$P(x), P(y), P(x, y), P(y|x) \text{ and } P(x|y)$$

(b) Using the descriptions in (a), compute the following measures:

$$E[X], E[Y], E[Y|X = x], E[X|Y = y]$$

$$Var(X), Var(Y), Cov(X, Y), \rho_{XY}$$

(c) Given that Player 2 won a dollar, what is your best guess of Player 1's payoff?
(d) Given that Player 1 won a dollar, what is your best guess of Player 2's payoff?
(e) Are there two events, $X = x$ and $Y = y$, that are mutually independent?

Study question 1.3.8

Compute the following theoretical measures of the outcome of a single game of craps (one roll of two independent dice), where X stands for the outcome of Die 1, Z for the outcome of Die 2, and Y for their sum.

(a)

$$E[X], E[Y], E[Y|X = x], E[X|Y = y], \text{ for each value of } x \text{ and } y, \text{ and}$$

$$Var(X), Var(Y), Cov(X, Y), \rho_{XY}, Cov(X, Z)$$

Table 1.6 describes the outcomes of 12 craps games.
(b) Find the sample estimates of the measures computed in (a), based on the data from Table 1.6. [Hint: Many software packages are available for doing this computation for you.]
(c) Use the results in (a) to determine the best estimate of the sum, Y, given that we measured $X = 3$.

Table 1.6 Results of 12 rolls of two fair dice

	X Die 1	Z Die 2	Y Sum
Roll 1	6	3	9
Roll 2	3	4	7
Roll 3	4	6	10
Roll 4	6	2	8
Roll 5	6	4	10
Roll 6	5	3	8
Roll 7	1	5	6
Roll 8	3	5	8
Roll 9	6	5	11
Roll 10	3	5	8
Roll 11	5	3	8
Roll 12	4	5	9

(d) What is the best estimate of X, given that we measured $Y = 4$?
(e) What is the best estimate of X, given that we measured $Y = 4$ and $Z = 1$? Explain why it is not the same as in (d).

1.3.10 Regression

Often, in statistics, we wish to predict the value of one variable, Y, based on the value of another variable, X. For example, we may want to predict a student's height based on his age. We noted earlier that the best prediction of Y based on X is given by the conditional expectation $E[Y|X = x]$, at least in terms of mean-squared error. But this assumes that we know the conditional expectation, or can compute it, from the joint distribution $P(y, x)$. With *regression*, we make our prediction directly from the data. We try to find a formula, usually a linear function, that takes observed values of X as input and gives values of Y as output, such that the square error between the predicted and actual values of Y is minimized, on average.

We start with a scatter plot that takes every case in our data set and charts them on a coordinate plane, as shown in Figure 1.2. Our predictor, or input, variable goes on the x-axis, and the variable whose value we are predicting goes on the y-axis.

The least squares regression line is the line for which the sum of the squared vertical distances of the points on the scatter plot from the line is minimized. That is, if there are n data points (x, y) on our scatter plot, and for any data point (x_i, y_i), the value y'_i represents the value of the line $y = \alpha + \beta x$ at x_i, then the least squares regression line is the one that minimizes the value

$$\sum_i (y_i - y'_i)^2 = \sum_i (y_i - \alpha - \beta x_i)^2 \tag{1.18}$$

To see how the slope β relates to the probability distribution $P(x, y)$, suppose we play 12 successive rounds of craps, and get the results shown in Table 1.6. If we wanted to predict the sum Y of the die rolls based on the value of $X = Die\ 1$ alone, using the data in Table 1.6, we would use the scatter plot shown in Figure 1.3. For our craps example, the least squares

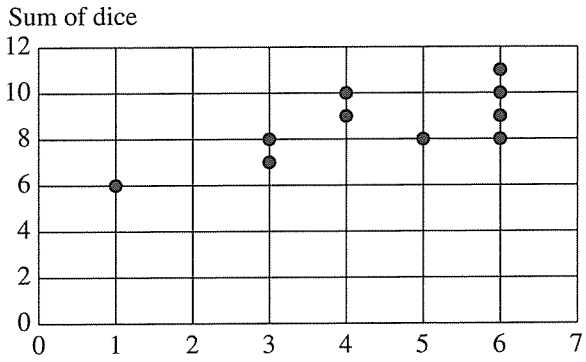


Figure 1.3 Scatter plot of the results in Table 1.6, with the value of Die 1 on the x-axis and the sum of the two dice rolls on the y-axis

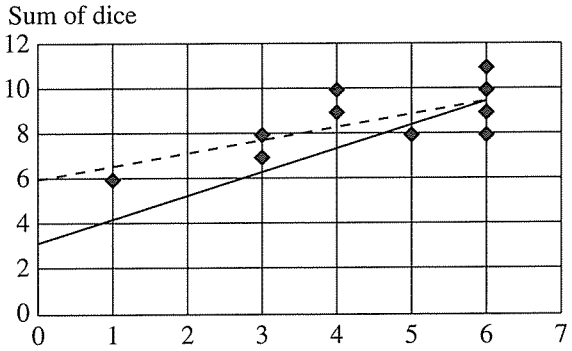


Figure 1.4 Scatter plot of the results in Table 1.6, with the value of Die 1 on the x-axis and the sum of the two dice rolls on the y-axis. The dotted line represents the line of best fit based on the data. The solid line represents the line of best fit we would expect in the population

regression line is shown in Figure 1.4. Note that the regression line for the *sample* that we used is not necessarily the same as the regression line for the *population*. The population is what we get when we allow our sample size to increase to infinity. The solid line in Figure 1.4 represents the theoretical least-square line, which is given by

$$y = 3.5 + 1.0x \tag{1.19}$$

The dashed line represents the sample least-square line, which, due to sampling variations, differs from the theoretical both in slope and in intercept.

In Figure 1.4, we know the equation of the regression line for the population because we know the expected value of the sum of two dice rolls, given that the first die lands on x. The computation is simple:

$$E[Y|X = x] = E[Die\ 2 + X|X = x] = E[Die\ 2] + x = 3.5 + 1.0x$$

This result is not surprising, since Y (the sum of the two dice) can be written as

$$Y = X + Z$$

where Z is the outcome of Die 2, and it stands to reason that if X increases by one unit, say from $X = 3$ to $X = 4$, then $E[Y]$ will, likewise, increase by one unit. The reader might be a bit surprised, however, to find out that the reverse is not the case; the regression of X on Y does not have a slope of 1.0. To see why, we write

$$E[X|Y = y] = E[Y - Z|Y = y] = 1.0y - E[Z|Y = y] \tag{1.20}$$

and realize that the added term, $E[Z|Y = y]$, since it depends (linearly) on y, makes the slope less than unity. We can in fact compute the exact value of $E[X|Y = y]$ by appealing to symmetry and write

$$E[X|Y = y] = E[Z|Y = y]$$

which gives, after substituting in Eq. (1.20),

$$E[X|Y = y] = 0.5y$$

The reason for this reduction is that, when we increase Y by one unit, each of X and Z contributes equally to this increase or average. This matches intuition; observing that the sum of the two dice is $Y = 10$, our best estimate of each is $X = 5$ and $Z = 5$.

In general, if we write the regression equation for Y on X as

$$y = a + bx \tag{1.21}$$

the slope b is denoted by R_{YX}, and it can be written in terms of the covariate σ_{XY} as follows:

$$b = R_{YX} = \frac{\sigma_{XY}}{\sigma_X^2} \tag{1.22}$$

From this equation, we see clearly that the slope of Y on X may differ from the slope of X on Y—that is, in most cases, $R_{YX} \neq R_{XY}$. ($R_{YX} = R_{XY}$ only when the variance of X is equal to the variance of Y.) The slope of the regression line can be positive, negative, or zero. If it is positive, X and Y are said to have a *positive correlation*, meaning that as the value of X gets higher, the value of Y gets higher; if it is negative, X and Y are said to have a *negative correlation*, meaning that as the value of X gets higher, the value of Y gets lower; if it is zero (a horizontal line), X and Y have no linear correlation, and knowing the value of X does not assist us in predicting the value of Y, at least linearly. If two variables are correlated, whether positively or negatively (or in some other way), they are dependent.

1.3.11 Multiple Regression

It is also possible to regress a variable on several variables, using *multiple linear regression*. For instance, if we wanted to predict the value of a variable Y using the values of the variables X and Z, we could perform multiple linear regression of Y on $\{X, Z\}$, and estimate a regression relationship

$$y = r_0 + r_1 x + r_2 z \tag{1.23}$$

which represents an inclined plane through the three-dimensional coordinate system.

We can create a three-dimensional scatter plot, with values of Y on the y-axis, X on the x-axis, and Z on the z-axis. Then, we can cut the scatter plot into slices along the Z-axis. Each slice will constitute a two-dimensional scatter plot of the kind shown in Figure 1.4. Each of those 2-D scatter plots will have a regression line with a slope r_1. Slicing along the X-axis will give the slope r_2.

The slope of Y on X when we hold Z constant is called the *partial regression coefficient* and is denoted by $R_{YX \cdot Z}$. Note that it is possible for R_{YX} to be positive, whereas $R_{YX \cdot Z}$ is negative as shown in Figure 1.1. This is a manifestation of Simpson's Paradox: positive association between Y and X overall, that becomes negative when we condition on the third variable Z.

The computation of partial regression coefficients (e.g., r_1 and r_2 in (1.23)) is greatly facilitated by a theorem that is one of the most fundamental results in regression analysis. It states that if we write Y as a linear combination of variables $X_1, X_2, \ldots, X_k$ plus a noise term ϵ,

$$Y = r_0 + r_1 X_1 + r_2 X_2 + \cdots + r_k X_k + \epsilon \tag{1.24}$$

then, regardless of the underlying distribution of $Y, X_1, X_2, \ldots, X_k$, the best least-square coefficients are obtained when ϵ is uncorrelated with each of the regressors $X_1, X_2, \ldots, X_k$. That is,

$$Cov(\epsilon, X_i) = 0 \quad \text{for} \quad i = 1,2, \ldots, k$$

To see how this *orthogonality principle* is used to our advantage, assume we wish to compute the best estimate of $X = Die\ 1$ given the sum

$$Y = Die\ 1 + Die\ 2$$

Writing

$$X = \alpha + \beta Y + \epsilon$$

our goal is to find α and β in terms of estimable statistical measures. Assuming without loss of generality $E[\epsilon] = 0$, and taking expectation on both sides of the equation, we obtain

$$E[X] = \alpha + \beta E[Y] \tag{1.25}$$

Further multiplying both sides of the equation by X and taking the expectation gives

$$E[X^2] = \alpha E[X] + \beta E[YX] + E[X\epsilon] \tag{1.26}$$

The orthogonality principle dictates $E[X\epsilon] = 0$, and (1.25) and (1.26) yield two equations with two unknowns, α and β. Solving for α and β, we obtain

$$\alpha = E(X) - E(Y)\frac{\sigma_{XY}}{\sigma_Y^2}$$

$$\beta = \frac{\sigma_{XY}}{\sigma_Y^2}$$

which completes the derivation. The slope β could have been obtained from Eq. (1.22), by simply reversing X and Y, but the derivation above demonstrates a general method of computing slopes, in two or more dimensions.

Consider for example the problem of finding the best estimate of Z given two observations, $X = x$ and $Y = y$. As before, we write the regression equation

$$Z = \alpha + \beta_Y Y + \beta_X X + \epsilon$$

But now, to obtain three equations for α, β_Y, and β_X, we also multiply both sides by Y and X and take expectations. Imposing the orthogonality conditions $E[\epsilon Y] = E[\epsilon X] = 0$ and solving the resulting equations gives

$$\beta_Y = R_{ZY \cdot X} = \frac{\sigma_X^2 \sigma_{ZY} - \sigma_{ZX}\sigma_{XY}}{\sigma_Y^2 \sigma_X^2 - \sigma_{YX}^2} \tag{1.27}$$

$$\beta_X = R_{ZX \cdot Y} = \frac{\sigma_Y^2 \sigma_{ZX} - \sigma_{ZY}\sigma_{YX}}{\sigma_Y^2 \sigma_X^2 - \sigma_{YX}^2} \tag{1.28}$$

Equations (1.27) and (1.28) are generic; they give the linear regression coefficients $R_{ZY \cdot X}$ and $R_{ZX \cdot Y}$ for any three variables in terms of their variances and covariances, and as such, they allow us to see how sensitive these slopes are to other model parameters. In practice, however, regression slopes are estimated from sampled data by efficient "least-square" algorithms, and rarely require memorization of mathematical equations. An exception is the task of predicting whether any of these slopes is zero, prior to obtaining any data. Such predictions are important when we contemplate choosing a set of regressors for one purpose or another, and as we shall see in Section 3.8, this task will be handled quite efficiently through the use of causal graphs.

Study question 1.3.9

(a) Prove Eq. (1.22) using the orthogonality principle. [Hint: Follow the treatment of Eq. (1.26).]

(b) Find all partial regression coefficients

$$R_{YX \cdot Z}, R_{XY \cdot Z}, R_{YZ \cdot X}, R_{ZY \cdot X}, R_{XZ \cdot Y}, \text{ and } R_{ZX \cdot Y}$$

for the craps game described in Study question 1.3.7. [Hint: Apply Eq. (1.27) and use the variances and covariances computed for part (a) of this question.]

1.4 Graphs

We learned from Simpson's Paradox that certain decisions cannot be made on the basis of data alone, but instead depend on the story behind the data. In this section, we layout a mathematical language, *graph theory*, in which these stories can be conveyed. Graph theory is not generally taught in high school mathematics, but it provides a useful mathematical language that allows us to address problems of causality with simple operations similar to those used to solve arithmetic problems.

Although the word *graph* is used colloquially to refer to a whole range of visual aids—more or less interchangeably with the word *chart*—in mathematics, a graph is a formally defined

object. A mathematical graph is a collection of *vertices* (or, as we will call them, *nodes*) and edges. The nodes in a graph are connected (or not) by the edges. Figure 1.5 illustrates a simple graph. X, Y, and Z (the dots) are nodes, and A and B (the lines) are edges.

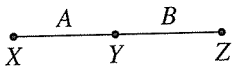


Figure 1.5 An undirected graph in which nodes X and Y are adjacent and nodes Y and Z are adjacent but not X and Z

Two nodes are *adjacent* if there is an edge between them. In Figure 1.5, X and Y are adjacent, and Y and Z are adjacent. A graph is said to be a *complete graph* if there is an edge between every pair of nodes in the graph.

A *path* between two nodes X and Y is a sequence of nodes beginning with X and ending with Y, in which each node is connected to the next by an edge. For instance, in Figure 1.5, there is a path from X to Z, because X is connected to Y, and Y is connected to Z.

Edges in a graph can be *directed* or *undirected*. Both of the edges in Figure 1.5 are undirected, because they have no designated "in" and "out" ends. A directed edge, on the other hand, goes out of one node and into another, with the direction indicated by an arrow head. A graph in which all of the edges are directed is a *directed graph*. Figure 1.6 illustrates a directed graph. In Figure 1.6, A is a directed edge from X to Y and B is a directed edge from Y to Z.

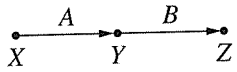


Figure 1.6 A directed graph in which node A is a parent of B and B is a parent of C

The node that a directed edge starts from is called the *parent* of the node that the edge goes into; conversely, the node that the edge goes into is the *child* of the node it comes from. In Figure 1.6, X is the parent of Y, and Y is the parent of Z; accordingly, Y is the child of X, and Z is the child of Y. A path between two nodes is a *directed* path if it can be traced along the arrows, that is, if no node on the path has two edges on the path directed into it, or two edges directed out of it. If two nodes are connected by a directed path, then the first node is the *ancestor* of every node on the path, and every node on the path is the *descendant* of the first node. (Think of this as an analogy to parent nodes and child nodes: parents are the ancestors of their children, and of their children's children, and of their children's children's children, etc.) For instance, in Figure 1.6, X is the ancestor of both Y and Z, and both Y and Z are descendants of X.

When a directed path exists from a node to itself, the path (and graph) is called *cyclic*. A directed graph with no cycles is *acyclic*. For example, in Figure 1.7(a) the graph is acyclic; however, the graph in Figure 1.7(b) is cyclic. Note that in (1) there is no directed path from any node to itself, whereas in (2) there are directed paths from X back to X, for example.

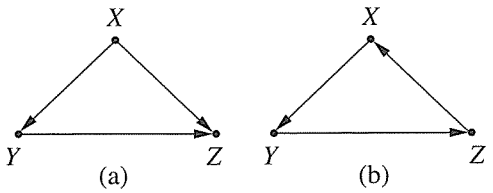


Figure 1.7 (a) Showing acyclic graph and (b) cyclic graph

Study questions

Study question 1.4.1

Consider the graph shown in Figure 1.8:

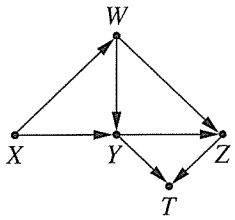


Figure 1.8 A directed graph used in Study question 1.4.1

(a) Name all of the parents of Z.
(b) Name all the ancestors of Z.
(c) Name all the children of W.
(d) Name all the descendants of W.
(e) Draw all (simple) paths between X and T (i.e., no node should appear more than once).
(f) Draw all the directed paths between X and T.

1.5 Structural Causal Models

1.5.1 Modeling Causal Assumptions

In order to deal rigorously with questions of causality, we must have a way of formally setting down our assumptions about the causal story behind a data set. To do so, we introduce the concept of the *structural causal model*, or SCM, which is a way of describing the relevant features of the world and how they interact with each other. Specifically, a structural causal model describes how nature assigns values to variables of interest.

Formally, a structural causal model consists of two sets of variables U and V, and a set of functions f that assigns each variable in V a value based on the values of the other variables in the model. Here, as promised, we expand on our definition of causation: A variable X is a *direct cause* of a variable Y if X appears in the function that assigns Y's value. X is a *cause* of Y if it is a direct cause of Y, or of any cause of Y.

The variables in U are called *exogenous variables*, meaning, roughly, that they are external to the model; we choose, for whatever reason, not to explain how they are caused. The variables in V are *endogenous*. Every endogenous variable in a model is a descendant of at least one exogenous variable. Exogenous variables cannot be descendants of any other variables, and in particular, cannot be a descendant of an endogenous variable; they have no ancestors and are represented as *root* nodes in graphs. If we know the value of every exogenous variable, then using the functions in f, we can determine with perfect certainty the value of every endogenous variable.

For example, suppose we are interested in studying the causal relationships between a treatment X and lung function Y for individuals who suffer from asthma. We might assume that Y also depends on, or is "caused by," air pollution levels as captured by a variable Z. In this case, we would refer to X and Y as endogenous and Z as exogenous. This is because we assume that air pollution is an external factor, that is, it cannot be caused by an individual's selected treatment or their lung function.

Every SCM is associated with a *graphical causal model*, referred to informally as a "graphical model" or simply "graph." Graphical models consist of a set of nodes representing the variables in U and V, and a set of edges between the nodes representing the functions in f. The graphical model G for an SCM M contains one node for each variable in M. If, in M, the function f_X for a variable X contains within it the variable Y (i.e., if X depends on Y for its value), then, in G, there will be a directed edge from Y to X. We will deal primarily with SCMs for which the graphical models are *directed acyclic graphs* (DAGs). Because of the relationship between SCMs and graphical models, we can give a graphical definition of causation: If, in a graphical model, a variable X is the child of another variable Y, then Y is a direct cause of X; if X is a descendant of Y, then Y is a potential cause of X (there are rare *intransitive cases* in which Y will not be a cause of X, which we will discuss in Part Two).

In this way, causal models and graphs encode causal assumptions. For instance, consider the following simple SCM:

SCM 1.5.1 (Salary Based on Education and Experience)

$$U = \{X, Y\}, \quad V = \{Z\}, \quad F = \{f_Z\}$$
$$f_Z : Z = 2X + 3Y$$

This model represents the salary (Z) that an employer pays an individual with X years of schooling and Y years in the profession. X and Y both appear in f_Z, so X and Y are both direct causes of Z. If X and Y had any ancestors, those ancestors would be potential causes of Z.

The graphical model associated with SCM 1.5.1 is illustrated in Figure 1.9.

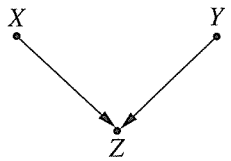


Figure 1.9 The graphical model of SCM 1.5.1, with X indicating years of schooling, Y indicating years of employment, and Z indicating salary

Because there are edges connecting Z to X and Y, we can conclude just by looking at the graphical model that there is some function f_Z in the model that assigns Z a value based on X and Y, and therefore that X and Y are causes of Z. However, without the fuller specification of an SCM, we can't tell from the graph what the function is that defines Z—or, in other words, *how* X and Y cause Z.

If graphical models contain less information than SCMs, why do we use them at all? There are several reasons. First, usually the knowledge that we have about causal relationships is not quantitative, as demanded by an SCM, but qualitative, as represented in a graphical model. We know off-hand that sex is a cause of height and that height is a cause of performance in basketball, but we would hesitate to give numerical values to these relationships. We could, instead of drawing a graph, simply create a partially specified version of the SCM:

SCM 1.5.2 (Basketball Performance Based on Height and Sex)

$$V = \{\text{Height, Sex, Performance}\}, \quad U = \{U_1, U_2, U_3\}, \quad F = \{f1, f2\}$$
$$\text{Sex} = U_1$$
$$\text{Height} = f_1(\text{Sex}, U_2)$$
$$\text{Performance} = f_2(\text{Height, Sex}, U_3)$$

Here, $U = \{U_1, U_2, U_3\}$ represents unmeasured factors that we do not care to name, but that affect the variables in V that we can measure. The U factors are sometimes called "error terms" or "omitted factors." These represent additional unknown and/or random exogenous causes of what we observe.

But graphical models provide a more intuitive understanding of causality than do such partially specified SCMs. Consider the SCM and its associated graphical model introduced above; while the SCM and its graphical model contain the same information, that is, that X causes Z and Y causes Z, that information is more quickly and easily ascertained by looking at the graphical model.

Study questions

Study question 1.5.1

Suppose we have the following SCM. Assume all exogenous variables are independent and that the expected value of each is 0.

SCM 1.5.3

$$V = \{X, Y, Z\}, \quad U = \{U_X, U_Y, U_Z\}, \quad F = \{f_X, f_Y, f_Z\}$$
$$f_X : X = u_X$$
$$f_Y : Y = \frac{X}{3} + U_Y$$
$$f_Z : Z = \frac{Y}{16} + U_Z$$

(a) Draw the graph that complies with the model.
(b) Determine the best guess of the value (expected value) of Z, given that we observe $Y = 3$.
(c) Determine the best guess of the value of Z, given that we observe $X = 3$.
(d) Determine the best guess of the value of Z, given that we observe $X = 1$ and $Y = 3$.
(e) Assume that all exogenous variables are normally distributed with zero means and unit variance, that is, $\sigma = 1$.
  (i) Determine the best guess of X, given that we observed $Y = 2$.
  (ii) (Advanced) Determine the best guess of Y, given that we observed $X = 1$ and $Z = 3$. [Hint: You may wish to use the technique of multiple regression, together with the fact that, for every three normally distributed variables, say X, Y, and Z, we have $E[Y|X = x, Z = z] = R_{YX \cdot Z}x + R_{YZ \cdot X}z$.]
(f) Determine the best guess of the value of Z, given that we know $X = 3$.

1.5.2 Product Decomposition

Another advantage of graphical models is that they allow us to express joint distributions very efficiently. So far, we have presented joint distributions in two ways. First, we have used tables, in which we assigned a probability to every possible combination of values. This is intuitively easy to parse, but in models with many variables, it can take up a prohibitive amount of space; 10 binary variables would require a table with 1024 rows!

Second, in a fully specified SCM, we can represent the joint distributions of n variables with greater efficiency: We need only to specify the n functions that govern the relationships between the variables, and then from the probabilities of the error terms, we can discover all the probabilities that govern the joint distribution. But we are not always in a position to fully specify a model; we may know that one variable is a cause of another but not the form of the equation relating them, or we may not know the distributions of the error terms. Even if we know these objects, writing them down may be easier said than done, especially, when the variables are discrete and the functions do not have familiar algebraic expressions.

Fortunately, we can use graphical models to help overcome both of these barriers through the following rule.

Rule of product decomposition

For any model whose graph is acyclic, the joint distribution of the variables in the model is given by the product of the conditional distributions $P(child|parents)$ over all the "families" in the graph. Formally, we write this rule as

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i|pa_i) \tag{1.29}$$

where pa_i stands for the values of the parents of variable X_i, and the product $\prod_i$ runs over all i, from 1 to n. The relationship (1.29) follows from certain universally true independencies among the variables, which will be discussed in the next chapter in more detail.

For example, in a simple chain graph $X \rightarrow Y \rightarrow Z$, we can write directly:

$$P(X = x, Y = y, Z = z) = P(X = x)P(Y = y|X = x)P(Z = z|Y = y)$$

This knowledge allows us to save an enormous amount of space when laying out a joint distribution. We need not create a probability table that lists a value for every possible triple (x, y, z). It will suffice to create three much smaller tables for X, $(Y|X)$, and $(Z|Y)$, and multiply the values as necessary.

To estimate the joint distribution from a data set generated by the above model, we need not count the frequency of every triple; we can instead count the frequencies of each x, $(y|x)$, and $(z|y)$ and multiply. This saves us a great deal of processing time in large models. It also increases substantially the accuracy of frequency counting. Thus, the assumptions underlying the graph allow us to exchange a "high-dimensional" estimation problem for a few "low-dimensional" probability distribution challenges. The graph therefore simplifies an estimation problem and, simultaneously, provides more precise estimators. If we do not know the graphical structure of an SCM, estimation becomes impossible with large number of variables and small, or moderately sized, data sets—the so-called "curse of dimensionality."

Graphical models let us do all of this without always needing to know the functions relating the variables, their parameters, or the distributions of their error terms.

Here's an evocative, if unrigorous, demonstration of the time and space saved by this strategy: Consider the chain $X \rightarrow Y \rightarrow Z \rightarrow W$, where X stands for clouds/no clouds, Y stands for rain/no rain, Z stands for wet pavement/dry pavement, and W stands for slippery pavement/unslippery pavement.

Using your own judgment, based on your experience of the world, how plausible is it that $P(\textit{clouds, no-rain, dry pavement, slippery pavement}) = 0.23$?

This is quite a difficult question to answer straight out. But using the product rule, we can break it into pieces:

$$P(\textit{clouds})P(\textit{no rain}|\textit{clouds})P(\textit{dry pavement}|\textit{no rain})P(\textit{slippery pavement}|\textit{dry pavement})$$

Our general sense of the world tells us that $P(\textit{clouds})$ should be relatively high, perhaps 0.5 (lower, of course, for those of us living in the strange, weatherless city of Los Angeles). Similarly, $P(\textit{no rain}|\textit{clouds})$ is fairly high—say, 0.75. And $P(\textit{dry pavement}|\textit{no rain})$ would be higher still, perhaps 0.9. But the $P(\textit{slippery pavement}|\textit{dry pavement})$ should be quite low, somewhere in the range of 0.05. So putting it all together, we come to a ballpark estimate of $0.5 \times 0.75 \times 0.9 \times 0.05 = 0.0169$.

We will use this product rule often in this book in cases when we need to reason with numerical probabilities, but wish to avoid writing out large probability tables.

The importance of the product decomposition rule can be particularly appreciated when we deal with estimation. In fact, much of the role of statistics focuses on effective sampling designs, and estimation strategies, that allow us to exploit an appropriate data set to estimate probabilities as precisely as we might need. Consider again the problem of estimating the probability $P(X, Y, Z, W)$ for the chain $X \rightarrow Y \rightarrow Z \rightarrow W$. This time, however, we attempt to estimate the probability from data, rather than our own judgment. The number of (x, y, z, w) combinations that need to be assigned probabilities is $16 - 1 = 15$. Assume that we have 45 random observations, each consisting of a vector (x, y, z, w). On the average, each (x, y, z, w) cell would receive about three samples; some will receive one or two samples, and some remain empty. It is very unlikely that we would obtain a sufficient number of samples in each cell to assess the proportion in the population at large (i.e., when the sample size goes to infinity).

If we use our product decomposition rule, however, the 45 samples are separated into much larger categories. In order to determine $P(x)$, every (x, y, z, w) sample falls into one of only two cells: $(X = 1)$ and $(X = 0)$. Clearly, the probability of leaving either of them empty is much lower, and the accuracy of estimating population frequencies is much higher. The same is true of the divisions we need to make to determine $P(y|x)$: $(Y = 1, X = 1), (Y = 0, X = 1), (Y = 1, X = 0)$, and $(Y = 0, X = 0)$. And to determine $P(z|y)$: $(Y = 1, Z = 1), (Y = 0, Z = 1), (Y = 1, Z = 0)$, and $(Y = 0, Z = 0)$. And to determine $P(w|z)$: $(W = 1, Z = 1), (W = 0, Z = 1), (W = 1, Z = 0)$, and $(W = 0, Z = 0)$. Each of these divisions will give us much more accurate frequencies than our original division into 15 cells. Here we explicitly see the simpler estimation problems allowed by assuming the graphical structure of an SCM and the resulting improved accuracy of our frequency estimates.

This is not the only use to which we can put the qualitative knowledge that a graph provides. As we will see in the next section, graphical models reveal much more information than is obvious at first glance; we can learn a lot about, and infer a lot from, a data set using only the graphical model of its causal story.

Study questions

Study question 1.5.2

Assume that a population of patients contains a fraction r of individuals who suffer from a certain fatal syndrome Z, which simultaneously makes it uncomfortable for them to take a life-prolonging drug X (Figure 1.10). Let $Z = z_1$ and $Z = z_0$ represent, respectively, the presence and absence of the syndrome, $Y = y_1$ and $Y = y_0$ represent death and survival, respectively, and $X = x_1$ and $X = x_0$ represent taking and not taking the drug. Assume that patients not carrying the syndrome, $Z = z_0$, die with probability p_2 if they take the drug and with probability p_1 if they don't. Patients carrying the syndrome, $Z = z_1$, on the other hand, die with probability p_3 if they do not take the drug and with probability p_4 if they do take the drug. Further, patients having the syndrome are more likely to avoid the drug, with probabilities $q_1 = P(x_1|z_0)$ and $q_2 = p(x_1|z_1)$.

(a) Based on this model, compute the joint distributions $P(x, y, z), P(x, y), P(x, z)$, and $P(y, z)$ for all values of x, y, and z, in terms of the parameters $(r, p_1, p_2, p_3, p_4, q_1, q_2)$. [Hint: Use the product decomposition of Section 1.5.2.]

(b) Calculate the difference $P(y_1|x_1) - P(y_1|x_o)$ for three populations: (1) those carrying the syndrome, (2) those not carrying the syndrome, and (3) the population as a whole.

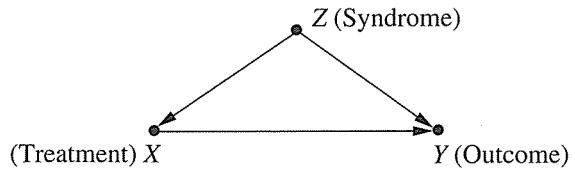


Figure 1.10 Model showing an unobserved syndrome, Z, affecting both treatment (X) and outcome (Y)

(c) *Using your results for (b), find a combination of parameters that exhibits Simpson's reversal.*

Study question 1.5.3

Consider a graph $X_1 \to X_2 \to X_3 \to X_4$ of binary random variables, and assume that the conditional probabilities between any two consecutive variables are given by

$$P(X_i = 1|X_{i-1} = 1) = p$$

$$P(X_i = 1|X_{i-1} = 0) = q$$

$$P(X_1 = 1) = p_0$$

Compute the following probabilities

$$P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0)$$

$$P(X_4 = 1|X_1 = 1)$$

$$P(X_1 = 1|X_4 = 1)$$

$$P(X_3 = 1|X_1 = 0, X_4 = 1)$$

Study question 1.5.4

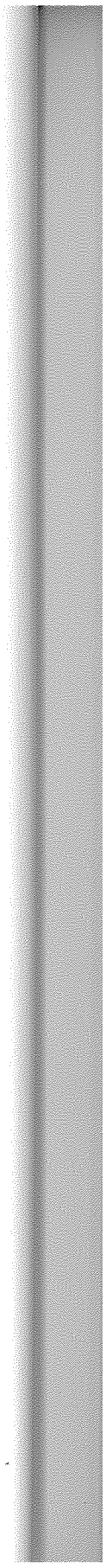
Define the structural model that corresponds to the Monty Hall problem, and use it to describe the joint distribution of all variables.

Bibliographical Notes for Chapter 1

An extensive account of the history of Simpson's paradoxis given in Pearl (2009, pp. 174–182), including many attempts by statisticians to resolve it without invoking causation. A more recent account, geared for statistics instructors is given in (Pearl 2014c). Among the many texts that provide basic introductions to probability theory, Lindley (2014) and Pearl (1988, Chapters 1 and 2) are the closest in spirit to the Bayesian perspective used in Chapter 1. The textbooks by Selvin (2004) and Moore et al. (2014) provide excellent introductions to classical methods ofstatistics, including parameter estimation, hypothesis testing and regression analysis.

The Monty Hall problem, discussed in Section 1.3, appears in many introductory books on probability theory (e.g., Grinstead and Snell 1998, p. 136; Lindley 2014, p. 201) and is mathematically equivalent to the "Three Prisoners Dilemma" discussed in (Pearl 1988, pp. 58–62). Friendly introductions to graphical models are given in Elwert (2013), Glymour and Greenland (2008), and the more advanced texts of Pearl (1988, Chapter 3), Lauritzen (1996) and Koller and Friedman (2009). The product decomposition rule of Section 1.5.2 was used in Howard and Matheson (1981) and Kiiveri et al. (1984) and became the semantic

basis of *Bayesian Networks* (Pearl 1985)—directed acyclic graphs that represent probabilistic knowledge, not necessarily causal. For inference and applications of Bayesian networks, see Darwiche (2009) and Fenton and Neil (2013), and Conrady and Jouffe (2015). The validity of the product decomposition rule for structural causal models was shown in Pearl and Verma (1991).



2

Graphical Models and Their Applications

2.1 Connecting Models to Data

In Chapter 1, we treated probabilities, graphs, and structural equations as isolated mathematical objects with little to connect them. But the three are, in fact, closely linked. In this chapter, we show that the concept of independence, which in the language of probability is defined by algebraic equalities, can be expressed visually using directed acyclic graphs (DAGs). Further, this graphical representation will allow us to capture the probabilistic information that is embedded in a structural equation model.

The net result is that a researcher who has scientific knowledge in the form of a structural equation model is able to predict patterns of independencies in the data, based solely on the structure of the model's graph, without relying on any quantitative information carried by the equations or by the distributions of the errors. Conversely, it means that observing patterns of independencies in the data enables us to say something about whether a hypothesized model is correct. Ultimately, as we will see in Chapter 3, the structure of the graph, when combined with data, will enable us to predict quantitatively the results of interventions without actually performing them.

2.2 Chains and Forks

We have so far referred to causal models as representations of the "causal story" underlying data. Another way to think of this is that causal models represent the *mechanism* by which data were generated. Causal models are a sort of blueprint of the relevant part of the universe, and we can use them to simulate data from this universe. Given a truly complete causal model for, say, math test scores in high school juniors, and given a complete list of values for every exogenous variable in that model, we could theoretically generate a data point (i.e., a test score) for any individual. Of course, this would necessitate specifying all factors that may have an effect on a student's test score, an unrealistic task. In most cases, we will not have such precise knowledge about a model. We might instead have a probability distribution characterizing the exogenous

Causal Inference in Statistics: A Primer, First Edition. Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell.

Companion Website: www.wiley.com/go/Pearl/Causality

variables, which would allow us to generate a distribution of test scores approximating that of the entire student population and relevant subgroups of students.

Suppose, however, that we do not have even a probabilistically specified causal model, but only a graphical structure of the model. We know which variables are caused by which other variables, but we don't know the strength or nature of the relationships. Even with such limited information, we can discern a great deal about the data set generated by the model. From an unspecified graphical causal model—that is, one in which we know which variables are functions of which others, but not the specific nature of the functions that connect them—we can learn which variables in the data set are independent of each other and which are independent of each other conditional on other variables. These independencies will be true of every data set generated by a causal model with that graphical structure, regardless of the specific functions attached to the SCM.

Consider, for instance, the following three hypothetical SCMs, all of which share the same graphical model. The first SCM represents the causal relationships among a high school's funding in dollars (X), its average SAT score (Y), and its college acceptance rate (Z) for a given year. The second SCM represents the causal relationships among the state of a light switch (X), the state of an associated electrical circuit (Y), and the state of a light bulb (Z). The third SCM concerns the participants in a foot race. It represents causal relationships among the hours that participants work at their jobs each week (X), the hours the participants put into training each week (Y), and the completion time, in minutes, the participants achieve in the race (Z). In all three models, the exogenous variables (U_X, U_Y, U_Z) stand in for any unknown or random effects that may alter the relationship between the endogenous variables. Specifically, in SCMs 2.2.1 and 2.2.3, U_Y and U_Z are additive factors that account for variations among individuals. In SCM 2.2.2, U_Y and U_Z take the value 1 if there is some unobserved abnormality, and 0 if there is none.

SCM 2.2.1 (School Funding, SAT Scores, and College Acceptance)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$
$$f_X : X = U_X$$
$$f_Y : Y = \frac{x}{3} + U_Y$$
$$f_Z : Z = \frac{y}{16} + U_Z$$

SCM 2.2.2 (Switch, Circuit, and Light Bulb)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$
$$f_X : X = U_X$$
$$f_Y : Y = \begin{cases} \text{Closed IF } (X = \text{Up AND } U_Y = 0) \text{ OR } (X = \text{Down AND } U_Y = 1) \\ \text{Open otherwise} \end{cases}$$
$$f_Z : Z = \begin{cases} \text{On IF } (Y = \text{Closed AND } U_Z = 0) \text{ OR } (Y = \text{Open AND } U_Z = 1) \\ \text{Off otherwise} \end{cases}$$

SCM 2.2.3 (Work Hours, Training, and Race Time)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$
$$f_X : X = U_X$$
$$f_Y : Y = 84 - x + U_Y$$
$$f_Z : Z = \frac{100}{y} + U_Z$$

SCMs 2.2.1–2.2.3 share the graphical model shown in Figure 2.1.

SCMs 2.2.1 and 2.2.3 deal with continuous variables; SCM 2.2.2 deals with categorical variables. The relationships between the variables in 2.1.1 are all positive (i.e., the higher the value of the parent variable, the higher the value of the child variable); the correlations between the variables in 2.2.3 are all negative (i.e., the higher the value of the parent variable, the lower the value of the child variable); the correlations between the variables in 2.2.2 are not linear at all, but logical. No two of the SCMs share any functions in common. But because they share a common graphical structure, the data sets generated by all three SCMs must share certain independencies—and we can predict those independencies simply by examining the graphical model in Figure 2.1. The independencies shared by data sets generated by these three SCMs, and the dependencies that are likely shared by all such SCMs, are these:

1. ***Z and Y are dependent***
   For some $z, y, P(Z = z|Y = y) \neq P(Z = z)$
2. ***Y and X are dependent***
   For some $y, x, P(Y = y|X = x) \neq P(Y = y)$
3. ***Z and X are likely dependent***
   For some $z, x, P(Z = z|X = x) \neq P(Z = z)$
4. ***Z and X are independent, conditional on Y***
   For all $x, y, z, P(Z = z|X = x, Y = y) = P(Z = z|Y = y)$

To understand why these independencies and dependencies hold, let's examine the graphical model. First, we will verify that any two variables with an edge between them are dependent. Remember that an arrow from one variable to another indicates that the first variable causes the second—and, more importantly, that the value of the first variable is part of the function that determines the value of the second. Therefore, the second variable *depends* on the first for

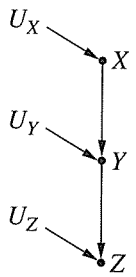


Figure 2.1 The graphical model of SCMs 2.2.1–2.2.3

its value; there is some case in which changing the value of the first variable changes the value of the second. That means that when we examine those variables in the data set, the probability that one variable takes a given value will change, given that we know the value of the other variable. So in any causal model, regardless of the specific functions, two variables connected by an edge are dependent. By this reasoning, we can see that in SCMs 2.2.1–2.2.3, Z and Y are dependent, and Y and X are dependent.

From these two facts, we can conclude that Z and X are *likely* dependent. If Z depends on Y for its value, and Y depends on X for its value, then Z likely depends on X for its value. There are pathological cases in which this is not true. Consider, for example, the following SCM, which also has the graph in Figure 2.1.

SCM 2.2.4 (Pathological Case of Intransitive Dependence)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = \begin{cases} a & \text{IF } X = 1 \text{ AND } U_Y = 1 \\ b & \text{IF } X = 2 \text{ AND } U_Y = 1 \\ c & \text{IF } U_Y = 2 \end{cases}$$

$$f_Y : Z = \begin{cases} i & \text{IF } Y = c \text{ OR } U_Z = 1 \\ j & \text{IF } U_Z = 2 \end{cases}$$

In this case, no matter what value U_Y and U_Z take, X will have no effect on the value that Z takes; changes in X account for variation in Y between a and b, but Y doesn't affect Z unless it takes the value c. Therefore, X and Z vary independently in this model. We will call cases such as these *intransitive cases*.

However, intransitive cases form only a small number of the cases we will encounter. In most cases, the values of X and Z vary together just as X and Y do, and Y and Z. Therefore, they are likely dependent in the data set.

Now, let's consider point 4: Z and X are independent conditional on Y. Remember that when we condition on Y, we filter the data into groups based on the value of Y. So we compare all the cases where $Y = a$, all the cases where $Y = b$, and so on. Let's assume that we're looking at the cases where $Y = a$. We want to know whether, *in these cases only*, the value of Z is independent of the value of X. Previously, we determined that X and Z are likely dependent, because when the value of X changes, the value of Y likely changes, and when the value of Y changes, the value of Z is likely to change. Now, however, examining *only the cases where* $Y = a$, when we select cases with different values of X, the value of U_Y changes so as to keep Y at $Y = a$, but since Z depends only on Y and U_Z, not on U_Y, the value of Z remains unaltered. So selecting a different value of X doesn't change the value of Z. So, in the case where $Y = a$, X is independent of Z. This is of course true no matter which specific value of Y we condition on. So X is independent of Z, conditional on Y.

This configuration of variables—three nodes and two edges, with one edge directed into and one edge directed out of the middle variable—is called a *chain*. Analogous reasoning to the above tells us that in any graphical model, given any two variables X and Y, if the only path between X and Y is composed entirely of chains, then X and Y are independent conditional on any intermediate variable on that path. This independence relation holds regardless of the functions that connect the variables. This gives us a rule:

Rule 1 (Conditional Independence in Chains) *Two variables, X and Y, are conditionally independent given Z, if there is only one unidirectional path between X and Y and Z is any set of variables that intercepts that path.*

An important note: Rule 1 only holds when we assume that the error terms U_X, U_Y, and U_Z are independent of each other. If, for instance, U_X were a cause of U_Y, then conditioning on Y would not necessarily make X and Z independent—because variations in X could still be associated with variations in Y, through their error terms.

Now, consider the graphical model in Figure 2.2. This structure might represent, for example, the causal mechanism that connects a day's temperature in a city in degrees Fahrenheit (X), the number of sales at a local ice cream shop on that day (Y), and the number of violent crimes in the city on that day (Z). Possible functional relationships between these variables are given in SCM 2.2.5. Or the structure might represent, as in SCM 2.2.6, the causal mechanism that connects the state (up or down) of a switch (X), the state (on or off) of one light bulb (Y), and the state (on or off) of a second light bulb (Z). The exogenous variables U_X, U_Y, and U_Z represent other, possibly random, factors that influence the operation of these devices.

SCM 2.2.5 (Temperature, Ice Cream Sales, and Crime)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$
$$f_X : X = U_X$$
$$f_Y : Y = 4x + U_Y$$
$$f_Z : Z = \frac{x}{10} + U_Z$$

SCM 2.2.6 (Switch and Two Light Bulbs)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$
$$f_X : X = U_X$$
$$f_Y : Y = \begin{cases} \text{On IF } (X = \text{Up AND } U_Y = 0) \text{ OR } (X = \text{Down AND } U_Y = 1) \\ \text{Off otherwise} \end{cases}$$
$$f_Z : Z = \begin{cases} \text{On IF } (X = \text{Up AND } U_Z = 0) \text{ OR } (X = \text{Down AND } U_Z = 1) \\ \text{Off otherwise} \end{cases}$$

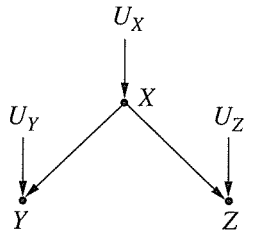


Figure 2.2 The graphical model of SCMs 2.2.5 and 2.2.6

If we assume that the error terms U_X, U_Y, and U_Z are independent, then by examining the graphical model in Figure 2.2, we can determine that SCMs 2.2.5 and 2.2.6 share the following dependencies and independencies:

1. ***X and Y are dependent.***
   For some $x, y, P(X = x|Y = y) \neq P(X = x)$
2. ***X and Z are dependent.***
   For some $x, z, P(X = x|Z = z) \neq P(X = x)$
3. ***Z and Y are likely dependent.***
   For some $z, y, P(Z = z|Y = y) \neq P(Z = z)$
4. ***Y and Z are independent, conditional on X.***
   For all $4x, y, z, P(Y = y|Z = z, X = x) = P(Y = y|X = x)$

Points 1 and 2 follow, once again, from the fact that Y and Z are both directly connected to X by an arrow, so when the value of X changes, the values of both Y and Z change. This tells us something further, however: If Y changes when X changes, and Z changes when X changes, then it is likely (though not certain) that Y changes together with Z, and vice versa. Therefore, since a change in the value of Y gives us information about an associated change in the value of Z, Y, and Z are likely dependent variables.

Why, then, are Y and Z independent conditional on X? Well, what happens when we condition on X? We filter the data based on the value of X. So now, we're only comparing cases where the value of X is constant. Since X does not change, the values of Y and Z do not change in accordance with it—they change only in response to U_Y and U_Z, which we have assumed to be independent. Therefore, any additional changes in the values of Y and Z must be independent of each other.

This configuration of variables—three nodes, with two arrows emanating from the middle variable—is called a *fork*. The middle variable in a fork is the *common cause* of the other two variables, and of any of their descendants. If two variables share a common cause, and if that common cause is part of the only path between them, then analogous reasoning to the above tells us that these dependencies and conditional independencies are true of those variables. Therefore, we come by another rule:

Rule 2 (Conditional Independence in Forks) *If a variable X is a common cause of variables Y and Z, and there is only one path between Y and Z, then Y and Z are independent conditional on X.*

2.3 Colliders

So far we have looked at two simple configurations of edges and nodes that can occur on a path between two variables: chains and forks. There is a third such configuration that we speak of separately, because it carries with it unique considerations and challenges. The third configuration contains a *collider* node, and it occurs when one node receives edges from two other nodes. The simplest graphical causal model containing a collider is illustrated in Figure 2.3, representing a common effect, Z, of two causes X and Y.

As is the case with every graphical causal model, all SCMs that have Figure 2.3 as their graph share a set of dependencies and independencies that we can determine from the graphical

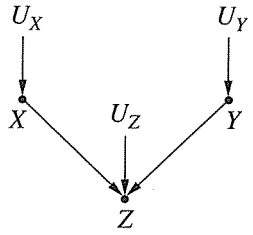


Figure 2.3 A simple collider

model alone. In the case of the model in Figure 2.3, assuming independence of U_X, U_Y, and U_Z, these independencies are as follows:

1. ***X and Z are dependent.***
   For some $x, z, P(X = x|Z = z) \neq P(X = x)$
2. ***Y and Z are dependent.***
   For some $y, z, P(Y = y|Z = z) \neq P(Y = y)$
3. ***X and Y are independent.***
   For all $x, y, P(X = x|Y = y) = P(X = x)$
4. ***X and Y are dependent conditional on Z.***
   For some $x, y, z, P(X = x|Y = y, Z = z) \neq P(X = x|Z = z)$

The truth of the first two points was established in Section 2.2. Point 3 is self-evident; neither X nor Y is a descendant or an ancestor of the other, nor do they depend for their value on the same variable. They respond only to U_X and U_Y, which are assumed independent, so there is no causal mechanism by which variations in the value of X should be associated with variations in the value of Y. This independence also reflects our understanding of how causation operates in time; events that are independent in the present do not become dependent merely because they may have common effects in the future.

Why, then, does point 4 hold? Why would two independent variables suddenly become dependent when we condition on their common effect? To answer this question, we return again to the definition of conditioning as filtering by the value of the conditioning variable. When we condition on Z, we limit our comparisons to cases in which Z takes the same value. But remember that Z depends, for its value, on X and Y. So, when comparing cases where Z takes, for example, the value, any change in value of X must be compensated for by a change in the value of Y—otherwise, the value of Z would change as well.

The reasoning behind this attribute of colliders—that conditioning on a collision node produces a dependence between the node's parents—can be difficult to grasp at first. In the most basic situation where $Z = X + Y$, and X and Y are independent variables, we have the following logic: If I tell you that $X = 3$, you learn nothing about the potential value of Y, because the two numbers are independent. On the other hand, if I start by telling you that $Z = 10$, then telling you that $X = 3$ immediately tells you that Y must be 7. Thus, X and Y are dependent, *given that* $Z = 10$.

This phenomenon can be further clarified through a real-life example. For instance, suppose a certain college gives scholarships to two types of students: those with unusual musical talents and those with extraordinary grade point averages. Ordinarily, musical talent and scholastic achievement are independent traits, so, in the population at large, finding a person with musical

talent tells us nothing about that person's grades. However, discovering that a person is on a scholarship changes things; knowing that the person lacks musical talent then tells us immediately that he is likely to have high grade point average. Thus, two variables that are marginally independent become dependent upon learning the value of a third variable (scholarship) that is a common effect of the first two.

Let's examine a numerical example. Consider a simultaneous (independent) toss of two fair coins and a bell that rings whenever at least one of the coins lands on heads. Let the outcomes of the two coins be denoted X and Y, respectively, and let Z stand for the state of the bell, with $Z = 1$ representing ringing, and $Z = 0$ representing silence. This mechanism can be represented be a collider as in Figure 2.3, in which the outcomes of the two coins are the parent nodes, and the state of the bell is the collision node.

If we know that Coin 1 landed on heads, it tells us nothing about the outcome of Coin 2, due to their independence. But suppose that we hear the bell ring and then we learn that Coin 1 landed on tails. We now know that Coin 2 must have landed on heads. Similarly, if we assume that we've heard the bell ring, the probability that Coin 1 landed on heads changes if we learn that Coin 2 also landed on heads. This particular change in probability is somewhat subtler than the first case.

To see the latter calculation, consider the initial probabilities as shown in Table 2.1.

We see that

$$P(X = \textit{"Heads"} | Y = \textit{"Heads"}) = P(X = \textit{"Tails"} | Y = \textit{"Tails"}) = \frac{1}{2}$$

That is, X and Y are independent. Now, let's condition on $Z = 1$ and $Z = 0$ (the bell ringing and not ringing). The resulting data subsets are shown in Table 2.2.

By calculating the probabilities in these tables, we obtain

$$P(X = \textit{"Heads"} | Z = 1) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

If we further filter the $Z = 1$ subtable to examine only those cases where $Y = \textit{"Heads"}$, we get

$$P(X = \textit{"Heads"} | Y = \textit{"Heads"}, Z = 1) = \frac{1}{2}$$

We see that, given $Z = 1$, the probability of $X = \textit{"Heads"}$ changes from $\frac{2}{3}$ to $\frac{1}{2}$ upon learning that $Y = \textit{"Heads."}$ So, clearly, X and Y are dependent given $Z = 1$. A more pronounced dependence occurs, of course, when the bell does not ring ($Z = 0$), because then we know that both coins must have landed on tails.

Table 2.1 Probability distribution for two flips of a fair coin, with X representing flip one, Y representing flip two, and Z representing a bell that rings if either flip results in heads

X	Y	Z	$P(X, Y, Z)$
Heads	Heads	1	0.25
Heads	Tails	1	0.25
Tails	Heads	1	0.25
Tails	Tails	0	0.25

Table 2.2 Conditional probability distributions for the distribution in Table 2.2. (Top: Distribution conditional on $Z = 1$. Bottom: Distribution conditional on $Z = 0$)

X	Y	$P(X, Y\|Z = 1)$
Heads	Heads	0.333
Heads	Tails	0.333
Tails	Heads	0.333
Tails	Tails	0
X	Y	$Pr(X, Y\|Z = 0)$
Heads	Heads	0
Heads	Tails	0
Tails	Heads	0
Tails	Tails	1

Another example of colliders in action—one that may serve to further illuminate the difficulty that such configurations can present to statisticians—is the Monty Hall Problem, which we first encountered in Section 1.3. At its heart, the Monty Hall Problem reflects the presence of a collider. Your initial choice of door is one parent node; the door behind which the car is placed is the other parent node; and the door Monty opens to reveal a goat is the collision node, causally affected by both the other two variables. The causation here is clear: If you choose Door A, and if Door A has a goat behind it, Monty is forced to open whichever of the remaining doors that has a goat behind it.

Your initial choice and the location of the car are independent; that's why you initially have a $\frac{1}{3}$ chance of choosing the door with the car behind it. However, as with the two independent coins, conditional on Monty's choice of door, your initial choice and the placement of the prizes are dependent. Though the car may only be behind Door B in $\frac{1}{3}$ of cases, it will be behind Door B in $\frac{2}{3}$ of cases in which you choose Door A and Monty opened Door C.

Just as conditioning on a collider makes previously independent variables dependent, so too does conditioning on any descendant of a collider. To see why this is true, let's return to our example of two independent coins and a bell. Suppose we do not hear the bell directly, but instead rely on a witness who is somewhat unreliable; whenever the bell *does not ring*, there is 50% chance that our witness will falsely report that it did. Letting W stand for the witness's report, the causal structure is shown in Figure 2.4, and the probabilities for all combinations of X, Y, and W are shown in Table 2.3.

The reader can easily verify that, based on this table, we have

$$P(X = \text{“Heads”}|Y = \text{“Heads”}) = P(X = \text{“Heads”}) = \frac{1}{2}$$

and

$$P(X = \text{“Heads”}|W = 1) = (0.25 + 0.25) \text{ or } \div (0.25 + 0.25 + 0.25 + 0.125) = \frac{0.5}{0.85}$$

and

$$P(X = \text{“Heads”}|Y = \text{“Heads”}, W = 1) = 0.25 \text{ or } \div (0.25 + 0.25) = 0.5 < \frac{0.5}{0.85}$$

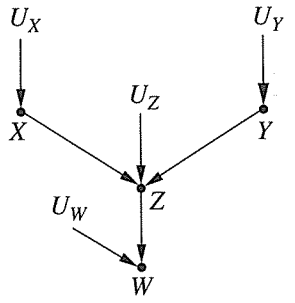


Figure 2.4 A simple collider, Z, with one child, W, representing the scenario from Table 2.3, with X representing one coin flip, Y representing the second coin flip, Z representing a bell that rings if either X or Y is heads, and W representing an unreliable witness who reports on whether or not the bell has rung

Table 2.3 Probability distribution for two flips of a fair coin and a bell that rings if either flip results in heads, with X representing flip one, Y representing flip two, and W representing a witness who, with variable reliability, reports whether or not the bell has rung

X	Y	W	$P(X, Y, W)$
Heads	Heads	1	0.25
Heads	Tails	1	0.25
Tails	Heads	1	0.25
Tails	Tails	1	0.125
Tails	Tails	0	0.125

Thus, X and Y are independent before reading the witness report, but become dependent thereafter.

These considerations lead us to a third rule, in addition to the two we established in Section 2.2.

Rule 3 (Conditional Independence in Colliders) *If a variable Z is the collision node between two variables X and Y, and there is only one path between X and Y, then X and Y are unconditionally independent but are dependent conditional on Z and any descendants of Z.*

Rule 3 is extremely important to the study of causality. In the coming chapters, we will see that it allows us to test whether a causal model could have generated a data set, to discover models from data, and to fully resolve Simpson's Paradox by determining which variables to measure and how to estimate causal effects under confounding.

Remark Inquisitive students may wonder why it is that dependencies associated with conditioning on a collider are so surprising to most people—as in, for example, the Monty Hall example. The reason is that humans tend to associate dependence with causation. Accordingly, they assume (wrongly) that statistical dependence between two variables can only exist if there is a causal mechanism that generate such dependence; that is, either one of the variables causes the other or a third variable causes both. In the case of a collider, they are surprised to find a

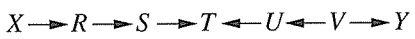


Figure 2.5 A directed graph for demonstrating conditional independence (error terms are not shown explicitly)

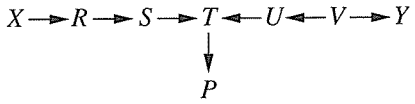


Figure 2.6 A directed graph in which P is a descendant of a collider

dependence that is created in a third way, thus violating the assumption of "no correlation without causation."

Study questions

Study question 2.3.1

(a) *List all pairs of variables in Figure 2.5 that are independent conditional on the set* $Z = \{R, V\}$.
(b) *For each pair of nonadjacent variables in Figure 2.5, give a set of variables that, when conditioned on, renders that pair independent.*
(c) *List all pairs of variables in Figure 2.6 that are independent conditional on the set* $Z = \{R, P\}$.
(d) *For each pair of nonadjacent variables in Figure 2.6, give a set of variables that, when conditioned on, renders that pair independent.*
(e) *Suppose we generate data by the model described in Figure 2.5, and we fit them with the linear equation* $Y = a + bX + cZ$. *Which of the variables in the model may be chosen for* Z *so as to guarantee that the slope* b *would be equal to zero? [Hint: Recall, a non zero slope implies that* Y *and* X *are dependent given* Z.*]*
(f) *Continuing question (e), suppose we fit the data with the equation:*

$$Y = a + bX + cR + dS + eT + fP$$

which of the coefficients would be zero?

2.4 *d*-separation

Causal models are generally not as simple as the cases we have examined so far. Specifically, it is rare for a graphical model to consist of a single path between variables. In most graphical models, pairs of variables will have multiple possible paths connecting them, and each such a path will traverse a variety of chains, forks, and colliders. The question remains whether there is a criterion or process that can be applied to a graphical causal model of *any* complexity in order to predict dependencies that are shared by all data sets generated by that graph.

There is, indeed, such a process: *d-separation*, which is built upon the rules established in the previous section. *d*-separation (the *d* stands for "directional") allows us to determine, for any pair of nodes, whether the nodes are *d-connected*, meaning there exists a connecting path between them, or *d-separated*, meaning there exists no such path. When we say that a pair of nodes are *d*-separated, we mean that the variables they represent are definitely independent; when we say that a pair of nodes are *d-connected*, we mean that they are possibly, or most likely, dependent.[1]

Two nodes X and Y are *d*-separated if every path between them (should any exist) is *blocked*. If even one path between X and Y is unblocked, X and Y are *d*-connected. The paths between variables can be thought of as pipes, and dependence as the water that flows through them; if even one pipe is unblocked, some water can pass from one place to another, and if a single path is clear, the variables at either end will be dependent. However, a pipe need only be blocked in one place to stop the flow of water through it, and similarly, it takes only one node to block the passage of dependence in an entire path.

There are certain kinds of nodes that can block a path, depending on whether we are performing unconditional or conditional *d*-separation. If we are not conditioning on any variable, then only colliders can block a path. The reasoning for this is fairly straightforward: as we saw in Section 2.2, unconditional dependence can't pass through a collider. So if every path between two nodes X and Y has a collider in it, then X and Y cannot be unconditionally dependent; they must be marginally independent.

If, however, we are conditioning on a set of nodes Z, then the following kinds of nodes can block a path:

- A collider that is not conditioned on (i.e., not in Z), and that has no descendants in Z.
- A chain or fork whose middle node is in Z.

The reasoning behind these points goes back to what we learned in Sections 2.2 and 2.3. A collider does not allow dependence to flow between its parents, thus blocking the path. But Rule 3 tells us that when we condition on a collider or its descendants, the parent nodes may become dependent. So a collider whose collision node is not in the conditioning set Z would block dependence from passing through a path, but one whose collision node, or its descendants, *is* in the conditioning set would not. Conversely, dependence can pass through noncolliders—chains and forks—but Rules 1 and 2 tell us that when we condition on them, the variables on either end of those paths become independent (when we consider one path at a time). So any noncollision node in the conditioning set would block dependence, whereas one that is not in the conditioning set would allow dependence through.

We are now prepared to give a general definition of *d*-separation:

Definition 2.4.1 (*d*-separation) *A path p is blocked by a set of nodes Z if and only if*

1. *p contains a chain of nodes $A \to B \to C$ or a fork $A \leftarrow B \to C$ such that the middle node B is in Z (i.e., B is conditioned on), or*
2. *p contains a collider $A \to B \leftarrow C$ such that the collision node B is not in Z, and no descendant of B is in Z.*

[1] The *d*-connected variables will be dependent for almost all sets of functions assigned to arrows in the graph, the exception being the sorts of intransitive cases discussed in Section 2.2.

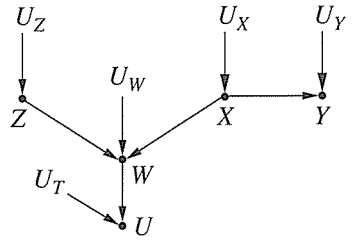


Figure 2.7 A graphical model containing a collider with child and a fork

If Z blocks every path between two nodes X and Y, then X and Y are d-separated, conditional on Z, and thus are independent conditional on Z.

Armed with the tool of d-separation, we can now look at some more complex graphical models and determine which variables in them are independent and dependent, both marginally and conditional on other variables. Let's take, for example, the graphical model in Figure 2.7. This graph might be associated with any number of causal models. The variables might be discrete, continuous, or a mixture of the two; the relationships between them might be linear, exponential, or any of an infinite number of other relations. No matter the model, however, d-separation will always provide the same set of independencies in the data the model generates.

In particular, let's look at the relationship between Z and Y. Using an empty conditioning set, they are d-separated, which tells us that Z and Y are unconditionally independent. Why? Because there is no unblocked path between them. There is only one path between Z and Y, and that path is blocked by a collider ($Z \rightarrow W \leftarrow X$).

But suppose we condition on W. d-separation tells us that Z and Y are d-connected, conditional on W. The reason is that our conditioning set is now $\{W\}$, and since the only path between Z and Y contains a fork (X) that is not in that set, and the only collider (W) on the path is in that set, that path is not blocked. (Remember that conditioning on colliders "unblocks" them.) The same is true if we condition on U, because U is a descendant of a collider along the path between Z and Y.

On the other hand, if we condition on the set $\{W, X\}$, Z and Y remain independent. This time, the path between Z and Y is blocked by the first criterion, rather than the second: There is now a noncollider node (X) on the path that is in the conditioning set. Though W has been unblocked by conditioning, one blocked node is sufficient to block the entire path. Since the only path between Z and Y is blocked by this conditioning set, Z and Y are d-separated conditional on $\{W, X\}$.

Now, consider what happens when we add another path between Z and Y, as in Figure 2.8. Z and Y are now unconditionally dependent. Why? Because there is a path between them ($Z \leftarrow T \rightarrow Y$) that contains no colliders. If we condition on T, however, that path is blocked, and Z and Y become independent again. Conditioning on $\{T, W\}$, on the other hand, makes them d-connected again (conditioning on T blocks the path $Z \leftarrow T \rightarrow Y$, but conditioning on W unblocks the path $Z \rightarrow W \leftarrow X \rightarrow Y$). And if we add X to the conditioning set, making it $\{T, W, X\}, Z$, and Y become independent yet again! In this graph, Z and Y are d-connected (and therefore likely dependent) conditional

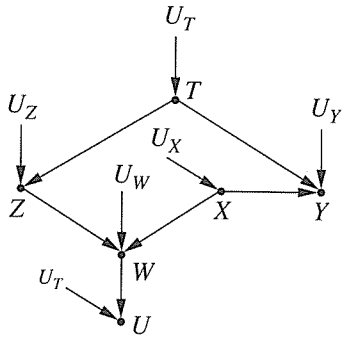


Figure 2.8 The model from Figure 2.7 with an additional forked path between Z and Y

on $W, U, \{W, U\}, \{W, T\}, \{U, T\}, \{W, U, T\}, \{W, X\}, \{U, X\}$, and $\{W, U, X\}$. They are d-separated (and therefore independent) conditional on $T, \{X, T\}, \{W, X, T\}, \{U, X, T\}$, and $\{W, U, X, T\}$. Note that T is in every conditioning set that d-separates Z and Y; that's because T is the only node in a path that unconditionally d-connects Z and Y, so unless it is conditioned on, Z and Y will always be d-connected.

Study questions

Study question 2.4.1

Figure 2.9 below represents a causal graph from which the error terms have been deleted. Assume that all those errors are mutually independent.

(a) For each pair of nonadjacent nodes in this graph, find a set of variables that d-separates that pair. What does this list tell us about independencies in the data?
(b) Repeat question (a) assuming that only variables in the set $\{Z_3, W, X, Z_1\}$ can be measured.
(c) For each pair of nonadjacent nodes in the graph, determine whether they are independent conditional on all other variables.
(d) For every variable V in the graph, find a minimal set of nodes that renders V independent of all other variables in the graph.
(e) Suppose we wish to estimate the value of Y from measurements taken on all other variables in the model. Find the smallest set of variables that would yield as good an estimate of Y as when we measured all variables.
(f) Repeat question (e) assuming that we wish to estimate the value of Z_2.
(g) Suppose we wish to predict the value of Z_2 from measurements of Z_3. Would the quality of our prediction improve if we add measurement of W? Explain.

2.5 Model Testing and Causal Search

The preceding sections demonstrate that causal models have *testable implications* in the data sets they generate. For instance, if we have a graph G that we believe might have generated

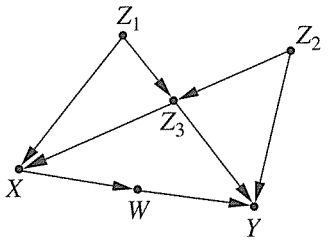


Figure 2.9 A causal graph used in study question 2.4.1, all U terms (not shown) are assumed independent

a data set S, d-separation will tell us which variables in G must be independent conditional on which other variables. Conditional independence is something we can test for using a data set. Suppose we list the d-separation conditions in G, and note that variables A and B must be independent conditional on C. Then, suppose we estimate the probabilities based on S, and discover that the data suggests that A and B are *not* independent conditional on C. We can then reject G as a possible causal model for S.

We can demonstrate it on the causal model of Figure 2.9. Among the many conditional independencies advertised by the model, we find that W and Z_1 are independent given X, because X d-separates W from Z_1. Now suppose we regress W on X and Z_1. Namely, we find the line

$$w = r_X x + r_1 z_1$$

that best fits our data. If it turns out that r_1 is not equal to zero, we know that W depends on Z_1 given X and, consequently, that the model is wrong. [Recall, conditional correlation implies conditional dependence.] Not only we know that the model is wrong, but we also know where it is wrong; the true model must have a path between W and Z_1 that is not d-separated by X. Finally, this is a theoretical result that holds for all acyclic models with independent errors (Verma and Pearl 1990), and we also know that if every d-separation condition in the model matches a conditional independence in the data, then no further test can refute the model. This means that, for any data set whatsoever, one can always find a set of functions F for the model and an assignment of probabilities to the U terms, so as to generate the data precisely.

There are other methods for testing the fitness of a model. The standard way of evaluating fitness involves a statistical hypothesis testing over the entire model, that is, we evaluate how likely it is for the observed samples to have been generated by the hypothesized model, as opposed to sheer chance. However, since the model is not fully specified, we need to first estimate its parameters before evaluating that likelihood. This can be done (approximately) when we assume a linear and Gaussian model (i.e., all functions in the model are linear and all error terms are normally distributed), because, under such assumptions, the joint distribution (also Gaussian) can be expressed succinctly in terms of the model's parameters, and we can then evaluate the likelihood that the observed samples to have been generated by the fully parameterized model (Bollen 1989).

There are, however, a number of issues with this procedure. First, if any parameter cannot be estimated, then the joint distribution cannot be estimated, and the model cannot be tested.

As we shall see in Section 3.8.3, this can occur when some of the error terms are correlated or, equivalently, when some of the variables are unobserved. Second, this procedure tests models globally. If we discover that the model is not a good fit to the data, there is no way for us to determine why that is—which edges should be removed or added to improve the fit. Third, when we test a model globally, the number of variables involved may be large, and if there is measurement noise and/or sampling variation associated with each variable, the test will not be reliable.

d-separation presents several advantages over this global testing method. First, it is nonparametric, meaning that it doesn't rely on the specific functions that connect variables; instead, it uses only the graph of the model in question. Second, it tests models locally, rather than globally. This allows us to identify specific areas, where our hypothesized model is flawed, and to repair them, rather than starting from scratch on a whole new model. It also means that if, for whatever reason, we can't identify the coefficient in one area of the model, we can still get some incomplete information about the rest of the model. (As opposed to the first method, in which if we could not estimate one coefficient, we could not test any part of the model.)

If we had a computer, we could test and reject many possible models in this way, eventually whittling down the set of possible models to only a few whose testable implications do not contradict the dependencies present in the data set. It is a set of models, rather than a single model, because some graphs have indistinguishable implications. A set of graphs with indistinguishable implications is called an *equivalence class*. Two graphs G_1 and G_2 are in the same equivalence class if they share a common skeleton—that is, the same edges, regardless of the direction of those edges—and if they share common *v-structures*, that is, colliders whose parents are not adjacent. Any two graphs that satisfy this criterion have identical sets of *d*-separation conditions and, therefore, identical sets of testable implications (Verma and Pearl 1990).

The importance of this result is that it allows us to search a data set for the causal models that could have generated it. Thus, not only can we start with a causal model and generate a data set—but we can also start with a data set, and reason back to a causal model. This is enormously useful, since the object of most data-driven research is exactly to find a model that explains the data.

There are other methods of causal search—including some that rely on the kind of global model testing with which we began the section—but a full investigation of them is beyond the scope of this book. Those interested in learning more about search should refer to (Pearl 2000; Pearl and Verma 1991; Rebane and Pearl 2003; Spirtes and Glymour 1991; Spirtes et al. 1993).

Study questions

Study question 2.5.1

(a) Which of the arrows in Figure 2.9 can be reversed without being detected by any statistical test? [Hint: Use the criterion for equivalence class.
(b) List all graphs that are observationally equivalent to the one in Figure 2.9.
(c) List the arrows in Figure 2.9 whose directionality can be determined from nonexperimental data.

(d) Write down a regression equation for Y such that, if a certain coefficient in that equation is nonzero, the model of Figure 2.9 is wrong.
(e) Repeat question (d) for variable Z_3.
(f) Repeat question (e) assuming the X is not measured.
(g) How many regression equations of the type described in (d) and (e) are needed to ensure that the model is fully tested, namely, that if it passes all these tests it cannot be refuted additional tests of these kind. [Hint: Ensure that you test every vanishing partial regression coefficient that is implied by the product decomposition (1.29).]

Bibliographical Notes for Chapter 2

The distinction between chains and forks in causal models was made by Simon (1953) and Reichenbach (1956) while the treatment of colliders (or common effect) can be traced back to the English economist Pigou (1911) (see Stigler 1999, pp. 36–41). In epidemiology, colliders came to be associated with "Selection bias" or "Berkson paradox" (Berkson 1946) while in artificial intelligence it came to be known as the "explaining away effect" (Kim and Pearl 1983). The rule of *d*-separation for determining conditional independence by graphs (Definition 2.4.1) was introduced in Pearl (1986) and formally proved in Verma and Pearl (1988) using the theory of graphoids (Pearl and Paz 1987). Gentle introductions to *d*-separation are available in Hayduk et al. (2003), Glymour and Greenland (2008), and Pearl (2000, pp. 335–337). Algorithms and software for detecting *d*-separation, as well as finding minimal separating sets are described in Tian et al. (1998), Kyono (2010), and Textor et al. (2011). The advantages of local over global model testing, are discussed in Pearl (2000, pp. 144–145) and further elaborated in Chen and Pearl (2014). Recent applications of *d*-separation include extrapolation across populations (Pearl and Bareinboim 2014) and handling missing data (Mohan et al. 2013).

3

The Effects of Interventions

3.1 Interventions

The ultimate aim of many statistical studies is to predict the effects of interventions. When we collect data on factors associated with wildfires in the west, we are actually searching for something we can intervene upon in order to decrease wildfire frequency. When we perform a study on a new cancer drug, we are trying to identify how a patient's illness responds when we intervene upon it by medicating the patient. When we research the correlation between violent television and acts of aggression in children, we are trying to determine whether intervening to reduce children's access to violent television will reduce their aggressiveness.

As you have undoubtedly heard many times in statistics classes, "correlation is not causation." A mere association between two variables does not necessarily or even usually mean that one of those variables causes the other. (The famous example of this property is that an increase in ice cream sales is correlated with an increase in violent crime—not because ice cream causes crime, but because both ice cream sales and violent crime are more common in hot weather.) For this reason, the randomized controlled experiment is considered the golden standard of statistics. In a properly randomized controlled experiment, all factors that influence the outcome variable are either static, or vary at random, except for one—so any change in the outcome variable must be due to that one input variable.

Unfortunately, many questions do not lend themselves to randomized controlled experiments. We cannot control the weather, so we can't randomize the variables that affect wildfires. We could conceivably randomize the participants in a study about violent television, but it would be difficult to effectively control how much television each child watches, and nearly impossible to know whether we were controlling them effectively or not. Even randomized drug trials can run into problems when participants drop out, fail to take their medication, or misreport their usage.

In cases where randomized controlled experiments are not practical, researchers instead perform observational studies, in which they merely record data, rather than controlling it. The problem of such studies is that it is difficult to untangle the causal from the merely correlative. Our common sense tells us that intervening on ice cream sales is unlikely to have any effect on crime, but the facts are not always so clear. Consider, for instance, a recent

Causal Inference in Statistics: A Primer, First Edition. Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell.

Companion Website: www.wiley.com/go/Pearl/Causality

University of Winnipeg study that showed that heavy text messaging in teens was correlated with "shallowness." Media outlets jumped on this as proof that texting makes teenagers more shallow. (Or, to use the language of intervention, that intervening to make teens text less would make them less shallow.) The study, however, proved nothing of the sort. It might be the case that shallowness makes teens more drawn to texting. It might be that both shallowness and heavy texting are caused by a common factor—a gene, perhaps—and that intervening on that variable, if possible, would decrease both.

The difference between intervening on a variable and conditioning on that variable should, hopefully, be obvious. When we intervene on a variable in a model, we fix its value. We *change* the system, and the values of other variables often change as a result. When we condition on a variable, we change nothing; we merely narrow our focus to the subset of cases in which the variable takes the value we are interested in. What changes, then, is our perception about the world, not the world itself.

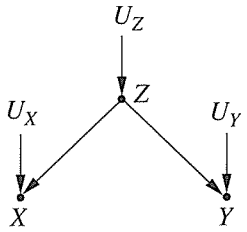


Figure 3.1 A graphical model representing the relationship between temperature (Z), ice cream sales (X), and crime rates (Y)

Consider, for instance, Figure 3.1 that shows a graphical model of our ice cream sales example, with X as ice cream sales, Y as crime rates, and Z as temperature. When we intervene to fix the value of a variable, we curtail the natural tendency of that variable to vary in response to other variables in nature. This amounts to performing a kind of surgery on the graphical model, removing all edges directed into that variable. If we were to intervene to make ice cream sales low (say, by shutting down all ice cream shops), we would have the graphical model shown in Figure 3.2. When we examine correlations in this new graph, we find that crime rates are, of course, totally independent of (i.e., uncorrelated with) ice cream sales since the latter is no longer associated with temperature (Z). In other words, even if we vary the level at which we hold X constant, that variation will not be transmitted to variable Y (crime rates). We see that intervening on a variable results in a totally different pattern of dependencies than conditioning on a variable. Moreover, the latter can be obtained

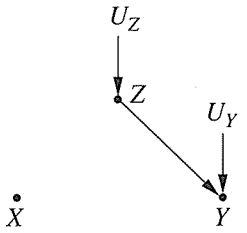


Figure 3.2 A graphical model representing an intervention on the model in Figure 3.1 that lowers ice cream sales

directly from the data set, using the procedures described in Part One, while the former varies depending on the structure of the causal graph. It is the graph that instructs us which arrow should be removed for any given intervention.

In notation, we distinguish between cases where a variable X takes a value x naturally and cases where we fix $X = x$ by denoting the latter $do(X = x)$. So $P(Y = y|X = x)$ is the probability that $Y = y$ conditional on finding $X = x$, while $P(Y = y|do(X = x))$ is the probability that $Y = y$ when we intervene to make $X = x$. In the distributional terminology, $P(Y = y|X = x)$ reflects the population distribution of Y among individuals whose X value is x. On the other hand, $P(Y = y|do(X = x))$ represents the population distribution of Y if *everyone in the population* had their X value fixed at x. We similarly write $P(Y = y|do(X = x), Z = z)$ to denote the conditional probability of $Y = y$, given $Z = z$, in the distribution created by the intervention $do(X = x)$.

Using *do*-expressions and graph surgery, we can begin to untangle the causal relationships from the correlative. In the rest of this chapter, we learn methods that can, astoundingly, tease out causal information from purely observational data, assuming of course that the graph constitutes a valid representation of reality. It is worth noting here that we are making a tacit assumption here that the intervention has no "side effects," that is, that *assigning* the value x for the valuable X for an individual does not alter subsequent variables in a direct way. For example, being "assigned" a drug might have a different effect on recovery than being forced to take the drug against one's religious objections. When side effects are present, they need to be specified explicitly in the model.

3.2 The Adjustment Formula

The ice cream example represents an extreme case in which the correlation between X and Y was totally spurious from a causal perspective, because there was no causal path from X to Y. Most real-life situations are not so clear-cut. To explore a more realistic situation, let us examine Figure 3.3, in which Y responds to both Z and X. Such a model could represent, for example, the first story we encountered for Simpson's paradox, where X stands for drug usage, Y stands for recovery, and Z stands for gender. To find out how effective the drug is in the population, we imagine a hypothetical intervention by which we administer the drug uniformly to the entire population and compare the recovery rate to what would obtain under the complementary intervention, where we prevent everyone from using the drug. Denoting the first intervention by $do(X = 1)$ and the second by $do(X = 0)$, our task is to estimate the difference

$$P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0)) \tag{3.1}$$

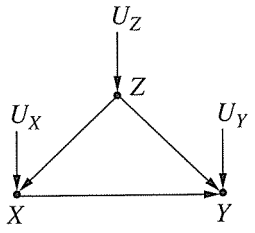


Figure 3.3 A graphical model representing the effects of a new drug, with Z representing gender, X standing for drug usage, and Y standing for recovery

which is known as the "causal effect difference," or "average causal effect" (ACE). In general, however, if X and Y can each take on more than one value, we would wish to predict the general causal effect $P(Y = y|do(X = x))$, where x and y are any two values that X and Y can take on. For example, x may be the dosage of the drug and y the patient's blood pressure.

We know from first principles that causal effects cannot be estimated from the data set itself without a causal story. That was the lesson of Simpson's paradox: The data itself was not sufficient even for determining whether the effect of the drug was positive or negative. But with the aid of the graph in Figure 3.3, we can compute the magnitude of the causal effect from the data. To do so, we simulate the intervention in the form of a graph surgery (Figure 3.4) just as we did in the ice cream example. The causal effect $P(Y = y|do(X = x))$ is equal to the conditional probability $P_m(Y = y|X = x)$ that prevails in the *manipulated* model of Figure 3.4. (This, of course, also resolves the question of whether the correct answer lies in the aggregated or the Z-specific table—when we determine the answer through an intervention, there's only one table to contend with.)

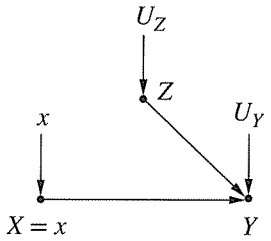


Figure 3.4 A modified graphical model representing an intervention on the model in Figure 3.3 that sets drug usage in the population, and results in the manipulated probability P_m

The key to computing the causal effect lies in the observation that P_m, the manipulated probability, shares two essential properties with P (the original probability function that prevails in the preintervention model of Figure 3.3). First, the marginal probability $P(Z = z)$ is invariant under the intervention, because the process determining Z is not affected by removing the arrow from Z to X. In our example, this means that the proportions of males and females remain the same, before and after the intervention. Second, the conditional probability $P(Y = y|Z = z, X = x)$ is invariant, because the process by which Y responds to X and Z, $Y = f(x, z, u_Y)$, remains the same, regardless of whether X changes spontaneously or by deliberate manipulation. We can therefore write two equations of invariance:

$$P_m(Y = y|Z = z, X = x) = P(Y = y|Z = z, X = x) \quad \text{and} \quad P_m(Z = z) = P(Z = z)$$

We can also use the fact that Z and X are d-separated in the modified model and are, therefore, independent under the intervention distribution. This tells us that $P_m(Z = z|X = x) = P_m(Z = z) = P(Z = z)$, the last equality following from above. Putting these considerations together, we have

$$\begin{aligned} &P(Y = y|do(X = x) \\ &\qquad = P_m(Y = y|X = x) \qquad \text{(by definition)} \end{aligned} \tag{3.2}$$

$$= \sum_z P_m(Y = y|X = x, Z = z)P_m(Z = z|X = x) \tag{3.3}$$

$$= \sum_z P_m(Y = y|X = x, Z = z)P_m(Z = z) \tag{3.4}$$

Equation (3.3) is obtained from Bayes' rule by conditioning on and summing over all values of $Z = z$ (as in Eq. (1.19)), while (Eq. 3.4) makes use of the independence of Z and X in the modified model.

Finally, using the invariance relations, we obtain a formula for the causal effect, in terms of preintervention probabilities:

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z) \tag{3.5}$$

Equation (3.5) is called the *adjustment formula*, and as you can see, it computes the association between X and Y for each value z of Z, then averages over those values. This procedure is referred to as "adjusting for Z" or "controlling for Z."

This final expression—the right-hand side of Eq. (3.5)—can be estimated directly from the data, since it consists only of conditional probabilities, each of which can be computed by the filtering procedure described in Chapter 1. Note also that no adjustment is needed in a randomized controlled experiment since, in such a setting, the data are generated by a model which already possesses the structure of Figure 3.4, hence, $P_m = P$ regardless of any factors Z that affect Y. Our derivation of the adjustment formula (3.5) constitutes therefore a formal proof that randomization gives us the quantity we seek to estimate, namely $P(Y = y|do(X = x))$. In practice, investigators use adjustments in randomized experiments as well, for the purpose of minimizing sampling variations (Cox 1958).

To demonstrate the working of the adjustment formula, let us apply it numerically to Simpson's story, with $X = 1$ standing for the patient taking the drug, $Z = 1$ standing for the patient being male, and $Y = 1$ standing for the patient recovering. We have

$$P(Y = 1|do(X = 1)) = P(Y = 1|X = 1, Z = 1)P(Z = 1) + P(Y = 1|X = 1, Z = 0)P(Z = 0)$$

Substituting the figures given in Table 1.1 we obtain

$$P(Y = 1|do(X = 1)) = \frac{0.93(87 + 270)}{700} + \frac{0.73(263 + 80)}{700} = 0.832$$

while, similarly,

$$P(Y = 1|do(X = 0)) = \frac{0.87(87 + 270)}{700} + \frac{0.69(263 + 80)}{700} = 0.7818$$

Thus, comparing the effect of drug-taking ($X = 1$) to the effect of nontaking ($X = 0$), we obtain

$$ACE = P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0)) = 0.832 - 0.7818 = 0.0502$$

giving a clear positive advantage to drug-taking. A more informal interpretation of ACE here is that it is simply the difference in the fraction of the population that would recover if everyone took the drug compared to when no one takes the drug.

We see that the adjustment formula instructs us to condition on gender, find the benefit of the drug separately for males and females, and only then average the result using the percentage of males and females in the population. It also thus instructs us to ignore the aggregated

population data $P(Y = 1|X = 1)$ and $P(Y = 1|X = 0)$, from which we might (falsely) conclude that the drug has a negative effect overall.

These simple examples might give readers the impression that whenever we face the dilemma of whether to condition on a third variable Z, the adjustment formula prefers the Z-specific analysis over the nonspecific analysis. But we know this is not so, recalling the blood pressure example of Simpson's paradox given in Table 1.2. There we argued that the more sensible method would be not to condition on blood pressure, but to examine the unconditional population table directly. How would the adjustment formula cope with situations like that?

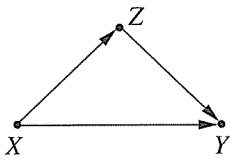


Figure 3.5 A graphical model representing the effects of a new drug, with X representing drug usage, Y representing recovery, and Z representing blood pressure (measured at the end of the study). Exogenous variables are not shown in the graph, implying that they are mutually independent

The graph in Figure 3.5 represents the causal story in the blood pressure example. It is the same as Figure 3.4, but with the arrow between X and Z reversed, reflecting the fact that the treatment has an effect on blood pressure and not the other way around. Let us try now to evaluate the causal effect $P(Y = 1|do(X = 1))$ associated with this model as we did with the gender example. First, we simulate an intervention and then examine the adjustment formula that emanates from the simulated intervention. In graphical models, an intervention is simulated by severing all arrows that enter the manipulated variable X. In our case, however, the graph of Figure 3.5 shows no arrow entering X, since X has no parents. This means that no surgery is required; the conditions under which data were obtained were such that treatment was assigned "as if randomized." If there was a factor that would make subjects prefer or reject treatment, such a factor should show up in the model; the absence of such a factor gives us the license to treat X as a randomized treatment.

Under such conditions, the intervention graph is equal to the original graph—no arrow need be removed—and the adjustment formula reduces to

$$P(Y = y|do(X = x)) = P(Y = y|X = x),$$

which can be obtained from our adjustment formula by letting the empty set be the element adjusted for. Obviously, if we were to adjust for blood pressure, we would obtain an incorrect assessment—one corresponding to a model in which blood pressure causes people to seek treatment.

3.2.1 *To Adjust or not to Adjust?*

We are now in a position to understand what variable, or set of variables, Z can legitimately be included in the adjustment formula. The intervention procedure, which led to the adjustment formula, dictates that Z should coincide with the parents of X, because it is the influence of

these parents that we neutralize when we fix X by external manipulation. Denoting the parents of X by $PA(X)$, we can therefore write a general adjustment formula and summarize it in a rule:

Rule 1 (The Causal Effect Rule) *Given a graph G in which a set of variables PA are designated as the parents of X, the causal effect of X on Y is given by*

$$P(Y = y|do(X = x) = \sum_z P(Y = y|X = x, PA = z)P(PA = z) \tag{3.6}$$

where z ranges over all the combinations of values that the variables in PA can take.

If we multiply and divide the summand in (3.6) by the probability $P(X = x|PA = z)$, we get a more convenient form:

$$P(y|do(x)) = \sum_z \frac{P(X = x, Y = y, PA = z)}{P(X = x|PA = z)} \tag{3.7}$$

which explicitly displays the role played by the parents of X in predicting the results of interventions. The factor $P(X = x|PA = z)$ is known as the "propensity score" and the advantages of expressing $P(y|do(x))$ in this form will be discussed in Section 3.5.

We can appreciate now what role the causal graph plays in resolving Simpson's paradox, and, more generally, what aspects of the graph allow us to predict causal effects from purely statistical data. We need the graph in order to determine the identity of X's parents—the set of factors that, under nonexperimental conditions, would be sufficient for determining the value of X, or the probability of that value.

This result alone is astounding; using graphs and their underlying assumptions, we were able to identify causal relationships in purely observational data. But, from this discussion, readers may be tempted to conclude that the role of graphs is fairly limited; once we identify the parents of X, the rest of the graph can be discarded, and the causal effect can be evaluated mechanically from the adjustment formula. The next section shows that things may not be so simple. In most practical cases, the set of X's parents will contain unobserved variables that would prevent us from calculating the conditional probabilities in the adjustment formula. Luckily, as we will see in future sections, we can adjust for other variables in the model to substitute for the unmeasured elements of $PA(X)$.

Study questions

Study questions 3.2.1

Referring to Study question 1.5.2 (Figure 1.10) and the parameters listed therein,

(a) Compute $P(y|do(x))$ for all values of x and y, by simulating the intervention $do(x)$ on the model.
(b) Compute $P(y|do(x))$ for all values of x and y, using the adjustment formula (3.5)
(c) Compute the ACE

$$ACE = P(y_1|do(x_1)) - P(y_1|do(x_0))$$

and compare it to the Risk Difference

$$RD = P(y_1|x_1) - P(y_1|x_0)$$

What is the difference between ACE and the RD? What values of the parameters would minimize the difference?

(d) *Find a combination of parameters that exhibit Simpson's reversal (as in Study question 1.5.2(c)) and show explicitly that the overall causal effect of the drug is obtained from the desegregated data.*

3.2.2 Multiple Interventions and the Truncated Product Rule

In deriving the adjustment formula, we assumed an intervention on a single variable, X, whose parents were disconnected, so as to simulate the absence of their influence after intervention. However, social and medical policies occasionally involve multiple interventions, such as those that dictate the value of several variables simultaneously, or those that control a variable over time. To represent multiple interventions, it is convenient to resort to the product decomposition that a graphical model imposes on joint distributions, as we have discussed in Section 1.5.2. According to the Rule of Product Decomposition, the preintervention distribution in the model of Figure 3.3 is given by the product

$$P(x, y, z) = P(z)P(x|z)P(y|x, z) \tag{3.8}$$

whereas the postintervention distribution, governed by the model of Figure 3.4 is given by the product

$$P(z, y|do(x)) = P_m(z)P_m(y|x, z) = P(z)P(y|x, z) \tag{3.9}$$

with the factor $P(x|z)$ purged from the product, since X becomes parentless as it is fixed at $X = x$. This coincides with the adjustment formula, because to evaluate $P(y|do(x))$ we need to marginalize (or sum) over z, which gives

$$P(y|do(x)) = \sum_z P(z)P(y|x, z)$$

in agreement with (3.5).

This consideration also allows us to generalize the adjustment formula to multiple interventions, that is, interventions that fix the values of a set of variables X to constants. We simply write down the product decomposition of the preintervention distribution, and strike out all factors that correspond to variables in the intervention set X. Formally, we write

$$P(x_1, x_2, \ldots, x_n|do(x)) = \prod_i P(x_i|pa_i) \qquad \text{for all } i \text{ with } X_i \text{ not in } X.$$

This came to be known as the *truncated product formula* or g-formula. To illustrate, assume that we intervene on the model of Figure 2.9 and set T to t and Z_3 to z_3. The postintervention distribution of the other variables in the model will be

$$P(z_1, z_2, w, y|do(T = t, Z_3 = z_3)) = P(z_1)P(z_2)P(w|t)P(y|w, z_3, z_2)$$

where we have deleted the factors $P(t|z_1, z_3)$ and $P(z_3|z_1, z_2)$ from the product.

It is interesting to note that combining (3.8) and (3.9), we get a simple relation between the pre- and postintervention distributions:

$$P(z, y|do(x)) = \frac{P(x, y, z)}{P(x|z)} \tag{3.10}$$

It tells us that the conditional probability $P(x|z)$ is all we need to know in order to predict the effect of an intervention $do(x)$ from nonexperimental data governed by the distribution $P(x, y, z)$.

3.3 The Backdoor Criterion

In the previous section, we came to the conclusion that we should adjust for a variable's parents, when trying to determine its effect on another variable. But often, we know, or believe, that the variables have unmeasured parents that, though represented in the graph, may be inaccessible for measurement. In those cases, we need to find an alternative set of variables to adjust for.

This dilemma unlocks a deeper statistical question: Under what conditions does a causal story permit us to compute the causal effect of one variable on another, from data obtained by passive observations, with no interventions? Since we have decided to represent causal stories with graphs, the question becomes a graph-theoretical problem: Under what conditions, is the structure of the causal graph sufficient for computing a causal effect from a given data set?

The answer to that question is long enough—and important enough—that we will spend the rest of the chapter addressing it. But one of the most important tools we use to determine whether we can compute a causal effect is a simple test called the *backdoor criterion*. Using it, we can determine whether, for any two variables X and Y in a causal model represented by a DAG, which set of variables Z in that model should be conditioned on when searching for the causal relationship between X and Y.

Definition 3.3.1 (The Backdoor Criterion) *Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the* backdoor criterion *relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.*

If a set of variables Z satisfies the backdoor criterion for X and Y, then the causal effect of X on Y is given by the formula

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

just as when we adjust for $PA(X)$. (Note that $PA(X)$ always satisfies the backdoor criterion.)

The logic behind the backdoor criterion is fairly straightforward. In general, we would like to condition on a set of nodes Z such that

1. We block all spurious paths between X and Y.
2. We leave all directed paths from X to Y unperturbed.
3. We create no new spurious paths.

When trying to find the causal effect of X on Y, we want the nodes we condition on to block any "backdoor" path in which one end has an arrow into X, because such paths may make X and Y dependent, but are obviously not transmitting causal influences from X, and if we do not block them, they will confound the effect that X has on Y. We condition on backdoor paths so as to fulfill our first requirement. However, we don't want to condition on any nodes that are descendants of X. Descendants of X would be affected by an intervention on X and might themselves affect Y; conditioning on them would block those pathways. Therefore, we don't condition on descendants of X so as to fulfill our second requirement. Finally, to comply with the third requirement, we should refrain from conditioning on any collider that would unblock a new path between X and Y. The requirement of excluding descendants of X also protects us from conditioning on children of intermediate nodes between X and Y (e.g., the collision node W in Figure 2.4.) Such conditioning would distort the passage of causal association between X and Y, similar to the way conditioning on their parents would.

To see what this means in practice, let's look at a concrete example, shown in Figure 3.6.

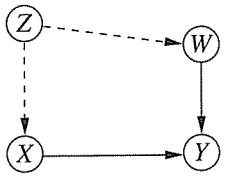


Figure 3.6 A graphical model representing the relationship between a new drug (X), recovery (Y), weight (W), and an unmeasured variable Z (socioeconomic status)

Here we are trying to gauge the effect of a drug (X) on recovery (Y). We have also measured weight (W), which has an effect on recovery. Further, we know that socioeconomic status (Z) affects both weight and the choice to receive treatment—but the study we are consulting did not record socioeconomic status.

Instead, we search for an observed variable that fits the backdoor criterion from X to Y. A brief examination of the graph shows that W, which is not a descendant of X, also blocks the backdoor path $X \leftarrow Z \rightarrow W \rightarrow Y$. Therefore, W meets the backdoor criterion. So long as the causal story conforms to the graph in Figure 3.6, adjusting for W will give us the causal effect of X on Y. Using the adjustment formula, we find

$$P(Y = y|do(X = x)) = \sum_w P(Y = y|X = x, W = w)P(W = w)$$

This sum can be estimated from our observational data, so long as W is observed.

With the help of the backdoor criterion, you can easily and algorithmically come to a conclusion about a pressing policy concern, even in complicated graphs. Consider the model in Figure 2.8, and assume again that we wish to evaluate the effect of X on Y. What variables should we condition on to obtain the correct effect? The question boils down to finding a set of variables that satisfy the backdoor criterion, but since there are no backdoor paths from X to Y, the answer is trivial: The empty set satisfies the criterion, hence no adjustment is needed. The answer is

$$P(y|do(x)) = P(y|x)$$

Suppose, however, that we were to adjust for W. Would we get the correct result for the effect of X on Y? Since W is a collider, conditioning on W would open the path $X \rightarrow W \leftarrow Z \leftrightarrow$

$T \rightarrow Y$. This path is spurious since it lies outside the causal pathway from X to Y. Opening this path will create bias and yield an erroneous answer. This means that computing the association between X and Y for each value of W separately will not yield the correct effect of X on Y, and it might even give the wrong effect for each value of W.

How then do we compute the causal effect of X on Y for a specific value w of W? W may represent, for example, the level of posttreatment pain of a patient, and we might be interested in assessing the effect of X on Y for only those patients who did not suffer any pain. Specifying the value of W amounts to conditioning on $W = w$, and this, as we have realized, opens a spurious path from X to Y by virtue of the fact the W is a collider.

The answer is that we still have the option of blocking that path using other variables. For example, if we condition on T, we would block the spurious path $X \rightarrow W \leftarrow Z \leftrightarrow T \rightarrow Y$, even if W is part of the conditioning set. Thus to compute the w-specific causal effect, written $P(y|do(x), w)$, we adjust for T, and obtain

$$P(Y = y|do(X = x), W = w) = \sum_t P(Y = y|X = x, W = w, T = t)P(T = t|W = w) \quad (3.11)$$

Computing such W-specific causal effects is an essential step in examining *effect modification* or *moderation*, that is, the degree to which the causal effect of X and Y is modified by different values of W. Consider, again, the model in Figure 3.6, and suppose we wish to test whether the causal effect for units at level $W = w$ is the same as for units at level $W = w'$ (W may represent any pretreatment variable, such as age, sex, or ethnicity). This question calls for comparing two causal effects,

$$P(Y = y|do(X = x), W = w) \quad \text{and} \quad P(Y = y|do(X = x), W = w')$$

In the specific example of Figure 3.6, the answer is simple, because W satisfies the backdoor criterion. So, all we need to compare are the conditional probabilities $P(Y = y|X = x, W = w)$ and $P(Y = y|X = x, W = w')$; no summation is required. In the more general case, where W alone does not satisfy the backdoor criterion, yet a larger set, $T \cup W$, does, we need to adjust for members of T, which yields Eq. (3.11). We will return to this topic in Section 3.5.

From the examples seen thus far, readers may get the impression that one should refrain from adjusting for colliders. Such adjustment is sometimes unavoidable, as seen in Figure 3.7. Here, there are four backdoor paths from X to Y, all traversing variable Z, which is a collider on the path $X \leftarrow E \rightarrow Z \leftarrow A \rightarrow Y$. Conditioning on Z will unblock this path and will violate the backdoor criterion. To block all backdoor paths, we need to condition on one of the following sets: $\{E, Z\}$, $\{A, Z\}$, or $\{E, Z, A\}$. Each of these contains Z. We see, therefore, that Z, a collider, must be adjusted for in any set that yields an unbiased estimate of the effect of X on Y.

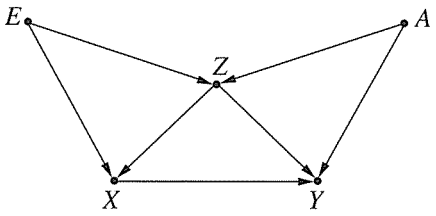


Figure 3.7 A graphical model in which the backdoor criterion requires that we condition on a collider (Z) in order to ascertain the effect of X on Y

The backdoor criterion has some further possible benefits. Consider the fact that $P(Y = y|do(X = x))$ is an empirical fact of nature, not a byproduct of our analysis. That means that any suitable variable or set of variables that we adjust on—whether it be $PA(X)$ or any other set that conforms to the backdoor criterion—must return the same result for $P(Y = y|do(X = x))$. In the case we looked at in Figure 3.6, this means that

$$\sum_w P(Y = y|X = x, W = w)P(W = w) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

This equality is useful in two ways. First, in the cases where we have multiple observed sets of variables suitable for adjustment (e.g., in Figure 3.6, if both W and Z had been observed), it provides us with a choice of which variables to adjust for. This could be useful for any number of practical reasons—perhaps one set of variables is more expensive to measure than the other, or more prone to human error, or simply has more variables and is therefore more difficult to calculate.

Second, the equality constitutes a testable constraint on the data when all the adjustment variables are observed, much like the rules of d-separation. If we are attempting to fit a model that leads to such an equality on a data set that violates it, we can discard that model.

Study questions

Study question 3.3.1

Consider the graph in Figure 3.8:

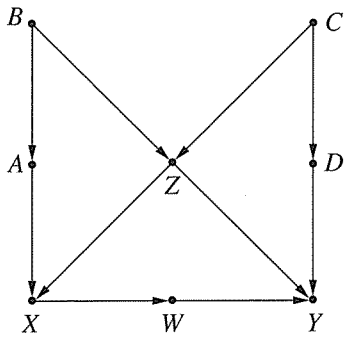


Figure 3.8 Causal graph used to illustrate the backdoor criterion in the following study questions

(a) *List all of the sets of variables that satisfy the backdoor criterion to determine the causal effect of X on Y.*
(b) *List all of the minimal sets of variables that satisfy the backdoor criterion to determine the causal effect of X on Y (i.e., any set of variables such that, if you removed any one of the variables from the set, it would no longer meet the criterion).*
(c) *List all minimal sets of variables that need be measured in order to identify the effect of D on Y. Repeat, for the effect of* $\{W, D\}$ *on Y.*

Study question 3.3.2 (Lord's paradox)

At the beginning of the year, a boarding school offers its students a choice between two meal plans for the year: Plan A and Plan B. The students' weights are recorded at the beginning and the end of the year. To determine how each plan affects students' weight gain, the school hired two statisticians who, oddly, reached different conclusions. The first statistician calculated the difference between each student's weight in June (W_F) and in September (W_I) and found that the average weight gain in each plan was zero.

The second statistician divided the students into several subgroups, one for each initial weight, W_I. He finds that for each initial weight, the final weight for Plan B is higher than the final weight for Plan A.

So, the first statistician concluded that there was no effect of diet on weight gain and the second concluded there was.

Figure 3.9 illustrates data sets that can cause the two statisticians to reach conflicting conclusions. Statistician-1 examined the weight gain $W_F - W_I$, which, for each student, is represented by the shortest distance to the 45° line. Indeed, the average gain for each diet plan is zero; the two groups are each situated symmetrically relative to the zero-gain line, $W_F = W_I$. Statistician-2, on the other hand, compared the final weights of plan A students to those of plan B students who entered school with the same initial weight W_0 and, as the vertical line in the figure indicates, plan B students are situated above plan A students along this vertical line. The same will be the case for any other vertical line, regardless of W_0.

(a) Draw a causal graph representing the situation.
(b) Determine which statistician is correct.
(c) How is this example related to Simpson's paradox?

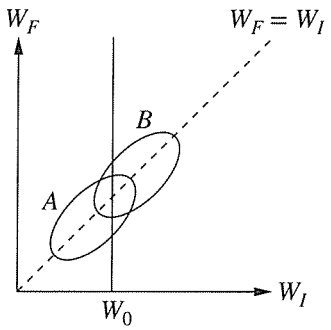


Figure 3.9 Scatter plot with students' initial weights on the x-axis and final weights on the y-axis. The vertical line indicates students whose initial weights are the same, and whose final weights are higher (on average) for plan B compared with plan A

Study questions 3.3.3

Revisit the lollipop story of Study question 1.2.4 and answer the following questions:

(a) Draw a graph that captures the story.
(b) Determine which variables must be adjusted for by applying the backdoor criterion.

(c) *Write the adjustment formula for the effect of the drug on recovery.*
(d) *Repeat questions (a)–(c) assuming that the nurse gave lollipops a day after the study, still preferring patients who received treatment over those who received placebo.*

3.4 The Front-Door Criterion

The backdoor criterion provides us with a simple method of identifying sets of covariates that should be adjusted for when we seek to estimate causal effects from nonexperimental data. It does not, however, exhaust *all* ways of estimating such effects. The *do*-operator can be applied to graphical patterns that do not satisfy the backdoor criterion to identify effects that on first sight seem to be beyond one's reach. One such pattern, called front-door, is discussed in this section.

Consider the century-old debate on the relation between smoking and lung cancer. In the years preceding 1970, the tobacco industry has managed to prevent antismoking legislation by promoting the theory that the observed correlation between smoking and lung cancer could be explained by some sort of carcinogenic genotype that also induces an inborn craving for nicotine.

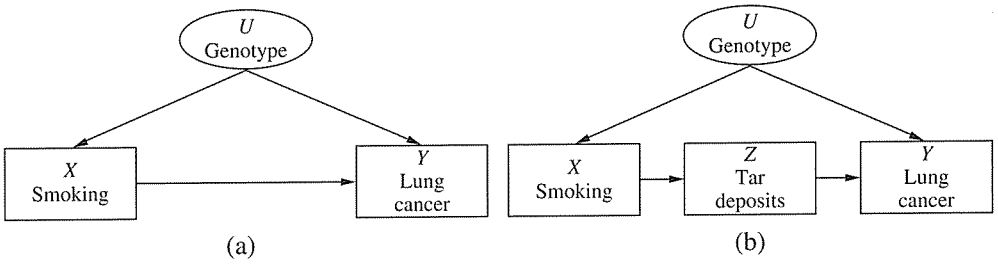


Figure 3.10 A graphical model representing the relationships between smoking (X) and lung cancer (Y), with unobserved confounder (U) and a mediating variable Z

A graph depicting this example is shown in Figure 3.10(a) This graph does not satisfy the backdoor condition because the variable U is unobserved and hence cannot be used to block the backdoor path from X to Y. The causal effect of smoking on lung cancer is not identifiable in this model; one can never ascertain which portion of the observed correlation between X and Y is spurious, attributable to their common effect, U, and what portion is genuinely causative. (We note, however, that even in these circumstances, much compelling work has been done to quantify how strong the (unobserved) associates between both U and X, and U and Y, must be in order to entirely explain the deserved association between X and Y.)

However, we can go much further by considering the model in Figure 3.10(b), where an additional measurement is available: the amount of tar deposits in patients lungs. This model does not satisfy the backdoor criterion, because there is still no variable capable of blocking the spurious path $X \leftarrow U \rightarrow Y$. We see, however, that the causal effect $P(Y = y|do(X = x))$ is nevertheless identifiable in this model, through two consecutive applications of the backdoor criterion.

How can the intermediate variable Z help us to assess the effect of X on Y? The answer is not at all trivial: as the following quantitative example shows, it may lead to heated debate.

Assume that a careful study was undertaken, in which the following factors were measured simultaneously on a randomly selected sample of 800,000 subjects considered to be at very high risk of cancer (because of environmental exposures such as smoking, asbestos, random, and the like).

1. Whether the subject smoked
2. Amount of tar in the subject's lungs
3. Whether lung cancer has been detected in the patient.

The data from this study are presented in Table 3.1, where, for simplicity, all three variables are assumed to be binary. All numbers are given in thousands.

Table 3.1 A hypothetical data set of randomly selected samples showing the percentage of cancer cases for smokers and nonsmokers in each tar category (numbers in thousands)

	Tar 400		No tar 400		All subjects 800	
	Smokers	Nonsmokers	Smokers	Nonsmokers	Smokers	Nonsmokers
	380	*20*	*20*	*380*	*400*	*400*
No cancer	323	1	18	38	341	39
	(85%)	(5%)	(90%)	(10%)	(85%)	(9.75%)
Cancer	57	19	2	342	59	361
	(15%)	(95%)	(10%)	(90%)	(15%)	(90.25%)

Two opposing interpretations can be offered for these data. The tobacco industry argues that the table proves the beneficial effect of smoking. They point to the fact that only 15% of the smokers have developed lung cancer, compared to 92.25% of the nonsmokers. Moreover, within each of two subgroups, tar and no tar, smokers show a much lower percentage of cancer than nonsmokers. (These numbers are obviously contrary to empirical observations but well illustrate our point that observations are not to be trusted.)

However, the antismoking lobbyists argue that the table tells an entirely different story—that smoking would actually increase, not decrease, one's risk of lung cancer. Their argument goes as follows: If you choose to smoke, then your chances of building up tar deposits are 95%, compared to 5% if you choose not to smoke (380/400 vs 20/400). To evaluate the effect of tar deposits, we look separately at two groups, smokers and nonsmokers, as done in Table 3.2. All numbers are given in thousands.

Table 3.2 Reorganization of the data set of Table 3.1 showing the percentage of cancer cases in each smoking-tar category (number in thousands)

	Smokers 400		Nonsmokers 400		All subjects 800	
	Tar	No tar	Tar	No tar	Tar	No tar
	380	*20*	*20*	*380*	*400*	*400*
No cancer	323	18	1	38	324	56
	(85%)	(90%)	(5%)	(10%)	(81%)	(19%)
Cancer	57	2	19	342	76	344
	(15%)	(10%)	(95%)	(90%)	(9%)	(81%)

It appears that tar deposits have a harmful effect in both groups; in smokers it increases cancer rates from 10% to 15%, and in nonsmokers it increases cancer rates from 90% to 95%. Thus, regardless of whether I have a natural craving for nicotine, I should avoid the harmful effect of tar deposits, and no-smoking offers very effective means of avoiding them.

The graph of Figure 3.10(b) enables us to decide between these two groups of statisticians. First, we note that the effect of X on Z is identifiable, since there is no backdoor path from X to Z. Thus, we can immediately write

$$P(Z = z|do(X = x)) = P(Z = z|X = x) \tag{3.12}$$

Next we note that the effect of Z on Y is also identifiable, since the backdoor path from Z to Y, namely $Z \leftarrow X \leftarrow U \rightarrow Y$, can be blocked by conditioning on X. Thus, we can write

$$P(Y = y|do(Z = z)) = \sum_x P(Y = y|Z = z, X = x) \tag{3.13}$$

Both (3.12) and (3.13) are obtained through the adjustment formula, the first by conditioning on the null set, and the second by adjusting for X.

We are now going to chain together the two partial effects to obtain the overall effect of X on Y. The reasoning goes as follows: If nature chooses to assign Z the value z, then the probability of Y would be $P(Y = y|do(Z = z))$. But the probability that nature would choose to do that, given that we choose to set X at x, is $P(Z = z|do(X = x))$. Therefore, summing over all states z of Z, we have

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|do(Z = z))P(Z = z|do(X = x)) \tag{3.14}$$

The terms on the right-hand side of (3.14) were evaluated in (3.12) and (3.13), and we can substitute them to obtain a *do*-free expression for $P(Y = y|do(X = x))$. We also distinguish between the x that appears in (3.12) and the one that appears in (3.13), the latter of which is merely an index of summation and might as well be denoted x'. The final expression we have is

$$P(Y = y|do(X = x)) =$$
$$\sum_z \sum_{x'} P(Y = y|Z = z, X = x')P(X = x')P(Z = z|X = x) \tag{3.15}$$

Equation (3.15) is known as the *front-door formula*.

Applying this formula to the data in Table 3.1, we see that the tobacco industry was wrong; tar deposits have a harmful effect in that they make lung cancer more likely and smoking, by increasing tar deposits, increases the chances of causing this harm.

The data in Table 3.1 are obviously unrealistic and were deliberately crafted so as to surprise readers with counterintuitive conclusions that may emerge from naive analysis of observational data. In reality, we would expect observational studies to show positive correlation between smoking and lung cancer. The estimand of (3.15) could then be used for confirming and quantifying the harmful effect of smoking on cancer.

The preceding analysis can be generalized to structures, where multiple paths lead from X to Y.

Definition 3.4.1 (Front-Door) *A set of variables Z is said to satisfy the front-door criterion relative to an ordered pair of variables (X, Y) if*

1. *Z intercepts all directed paths from X to Y.*
2. *There is no unblocked path from X to Z.*
3. *All backdoor paths from Z to Y are blocked by X.*

Theorem 3.4.1 (Front-Door Adjustment) *If Z satisfies the front-door criterion relative to (X, Y) and if $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by the formula*

$$P(y|do(x)) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x') \tag{3.16}$$

The conditions stated in Definition 3.4.1 are overly conservative; some of the backdoor paths excluded by conditions (ii) and (iii) can actually be allowed provided they are blocked by some variables. There is a powerful symbolic machinery, called the *do-calculus*, that allows analysis of such intricate structures. In fact, the *do*-calculus uncovers *all* causal effects that can be identified from a given graph. Unfortunately, it is beyond the scope of this book (see Pearl 2009 and Shpitser and Pearl 2008 for details). But the combination of the adjustment formula, the backdoor criterion, and the front-door criterion covers numerous scenarios. It proves the enormous, even revelatory, power that causal graphs have in not merely representing, but actually discovering causal information.

Study questions

Study question 3.4.1

Assume that in Figure 3.8, only X, Y, and one additional variable can be measured. Which variable would allow the identification of the effect of X on Y? What would that effect be?

Study question 3.4.2

I went to a pharmacy to buy a certain drug, and I found that it was available in two different bottles: one priced at $1, the other at $10. I asked the druggist, "What's the difference?" and he told me, "The $10 bottle is fresh, whereas the $1 bottle one has been on the shelf for 3 years. But, you know, data shows that the percentage of recovery is much higher among those who bought the cheap stuff. Amazing isn't it?" I asked if the aged drug was ever tested. He said, "Yes, and this is even more amazing; 95% of the aged drug and only 5% of the fresh drug has lost the active ingredient, yet the percentage of recovery among those who got bad bottles, with none of the active ingredient, is still much higher than among those who got good bottles, with the active ingredient."

Before ordering a cheap bottle, it occurred to me to have a good look at the data. The data were, for each previous customer, the type of bottle purchased (aged or fresh), the concentration of the active ingredient in the bottle (high or low), and whether the customer recovered from the illness. The data perfectly confirmed the druggist's story. However, after making some additional calculations, I decided to buy the expensive bottle after all; even without testing its

content, I could determine that a fresh bottle would offer the average patient a greater chance of recovery.

Based on two very reasonable assumptions, the data show clearly that the fresh drug is more effective. The assumptions are as follows:

(i) Customers had no information about the chemical content (high or low) of the specific bottle of the drug that they were buying; their choices were influenced by price and shelf-age alone.
(ii) The effect of the drug on any given individual depends only on its chemical content, not on its shelf age (fresh or aged).

(a) Determine the relevant variables for the problem, and describe this scenario in a causal graph.
(b) Construct a data set compatible with the story and the decision to buy the expensive bottle.
(c) Determine the effect of choosing the fresh versus the aged drug by using assumptions (i) and (ii), and the data given in (b).

3.5 Conditional Interventions and Covariate-Specific Effects

The interventions considered thus far have been limited to actions that merely force a variable or a group of variables X to take on some specified value x. In general, interventions may involve dynamic policies in which a variable X is made to respond in a specified way to some set Z of other variables—say, through a functional relationship $x = g(z)$ or through a stochastic relationship, whereby X is set to x with probability $P^*(x|z)$. For example, suppose a doctor decides to administer a drug only to patients whose temperature Z exceeds a certain level, $Z = z$. In this case, the action will be *conditional* upon the value of Z and can be written $do(X = g(Z))$, where $g(Z)$ is equal to one when $Z > z$ and zero otherwise (where $X = 0$ represents no drug). Since Z is a random variable, the value of X chosen by the action will similarly be a random variable, tracking variations in Z. The result of implementing such a policy is a probability distribution written $P(Y = y|do(X = g(Z)))$, which depends only on the function g and the set Z of variables that drive X.

In order to estimate the effect of such a policy, let us take a closer look at another concept, the "z-specific effect" of X, which we encountered briefly in Section 3.3 (Eq. (3.11)). This effect, written $P(Y = y|do(X = x), Z = z)$ measures the distribution of Y in a subset of the population for which Z achieves the value z after the intervention. For example, we may be interested in how a treatment affects a specific age group, $Z = z$, or people with a specific feature, $Z = z$, which may be measured after the treatment.

The z-specific effect can be identified by a procedure similar to the backdoor adjustment. The reasoning goes as follows: When we aim to estimate $P(Y = y|do(X = x))$, an adjustment for a set S is justified if S blocks all backdoor paths from X to Y. Now that we wish to identify $P(Y = y|do(X = x), Z = z)$, we need to ensure that those paths remain blocked when we add one more variable, Z, to the conditioning set. This yield a simple criterion for the identification of z-specific effect:

Rule 2 *The z-specific effect $P(Y = y|do(X = x), Z = z)$ is identified whenever we can measure a set S of variables such that $S \cup Z$ satisfies the backdoor criterion. Moreover, the z-specific*

effect is given by the following adjustment formula

$$P(Y = y|do(X = x), Z = z) = \sum_s P(Y = y|X = x, S = s, Z = z)P(S = s)$$

This modified adjustment formula is similar to Eq. (3.5) with two exceptions. First, the adjustment set is $S \cup Z$, not just S and, second, the summation goes only over S, not including Z. The $\cup$ symbol in the expression $S \cup Z$ stands for set addition (or union), which means that, if Z is a subset of S, we have $S \cup Z = S$, and S alone need satisfy the backdoor criterion.

Note that the identifiability criterion for z-specific effects is somewhat stricter than that for nonspecific effect. Adding Z to the conditioning set might create dependencies that would prevent the blocking of all backdoor paths. A simple example occurs when Z is a collider; conditioning on Z will create new dependency between Z's parents and thus violate the backdoor requirement.

We are now ready to tackle our original task of estimating conditional interventions. Suppose a policy maker contemplates an age-dependent policy whereby an amount x of drug is to be administered to patients, depending on their age Z. We write it as $do(X = g(Z))$. To find out the distribution of outcome Y that results from this policy, we seek to estimate $P(Y = y|do(X = g(Z)))$.

We now show that identifying the effect of such policies is equivalent to identifying the expression for the z-specific effect $P(Y = y|do(X = x), Z = z)$.

To compute $P(Y = y|do(X = g(Z)))$, we condition on $Z = z$ and write

$$\begin{aligned} &P(Y = y|do(X = g(Z))) \\ &= \sum_z P(Y = y|do(X = g(Z)), Z = z)P(Z = z|do(X = g(Z))) \\ &= \sum_z P(Y = y|do(X = g(z)), Z = z)P(Z = z) \end{aligned} \tag{3.17}$$

The equality

$$P(Z = z|do(X = g(Z))) = P(Z = z)$$

stems, of course, from the fact that Z occurs before X; hence, any control exerted on X can have no effect on the distribution of Z. Equation (3.17) can also be written as

$$\sum_z P(Y = y|do(X = x), z)|_{x=g(z)}P(Z = z)$$

which tells us that the causal effect of a conditional policy $do(X = g(Z))$ can be evaluated directly from the expression of $P(Y = y|do(X = x), Z = z)$ simply by substituting $g(z)$ for x and taking the expectation over Z (using the observed distribution $P(Z = z)$).

Study question 3.5.1

Consider the causal model of Figure 3.8.

(a) Find an expression for the c-specific effect of X on Y.

(b) *Identify a set of four variables that need to be measured in order to estimate the z-specific effect of X on Y, and find an expression for the size of that effect.*
(c) *Using your answer to part (b), determine the expected value of Y under a Z-dependent strategy, where X is set to 0 when Z is smaller or equal to 2 and X is set to 1 when Z is larger than 2. (Assume Z takes on integer values from 1 to 5.)*

3.6 Inverse Probability Weighing

By now, the astute reader may have noticed a problem with our intervention procedures. The backdoor and front-door criteria tell us whether it is possible to predict the results of hypothetical interventions from data obtained in an observational study. Moreover, they tell us that we can make this prediction without simulating the intervention and without even thinking about it. All we need to do is identify a set Z of covariates satisfying one of the criteria, plug this set into the adjustment formula, and we're done: the resulting expression is guaranteed to provide a valid prediction of how the intervention will affect the outcome.

This is lovely in theory, but in practice, adjusting for Z may prove problematic. It entails looking at each value or combination of values of Z separately, estimating the conditional probability of Y given X in that stratum and then averaging the results. As the number of strata increases, adjusting for Z will encounter both computational and estimational difficulties. Since the set Z can be comprised of dozens of variables, each spanning dozens of discrete values, the summation required by the adjustment formula may be formidable, and the number of data samples falling within each $Z = z$ cell may be too small to provide reliable estimates of the conditional probabilities involved.

All of our work in this chapter has not been for naught, however. The adjustment procedure is straightforward, and, therefore, easy to use in the explanation of intervention criteria. But there is another, more subtle procedure that overcomes the practical difficulties of adjustment.

In this section, we discuss one way of circumventing this problem, provided only that we can obtain a reliable estimate of the function $g(x, z) = P(X = x|Z = z)$, often called the "propensity score," for each x and z. Such an estimate can be obtained by fitting the parameters of a flexible function $g(x, z)$ to the data at hand, in much the same way that we fitted the coefficients of a linear regression function, so as to minimize the mean square error with respect to a set of samples (Figure 1.4). The method used will depend on the nature of the random variable X, whether it is continuous, discrete, binary, or a caveat, for example.

Assuming that the function $P(X = x|Z = z)$ is available to us, we can use it to generate artificial samples that act as though they were drawn from the postintervention probability P_m, rather than $P(x, y, z)$. Once we obtain such fictitious samples, we can evaluate $P(Y = y|do(x))$ by simply counting the frequency of the event $Y = y$, for each stratum $X = x$ in the sample. In this way, we skip the labor associated with summing over all strata $Z = z$; we essentially let nature do the summation for us.

The idea of estimating probabilities using fictitious samples is not new to us; it was used all along, though implicitly, whenever we estimated conditional probabilities from finite samples.

In Chapter 1, we characterized conditioning as a process of filtering—that is, ignoring all cases for which the condition $X = x$ does not hold, and normalizing the surviving cases, so that their total probabilities would add up to one. The net result of this operation is that the probability of each surviving case is boosted by a factor $1/P(X = x)$. This can be seen directly

from Bayes' rule, which tells us that

$$P(Y = y, Z = z|X = x) = \frac{P(Y = y, Z = z, X = x)}{P(X = x)}$$

In other words, to find the probability of each row in the surviving table, we multiply the unconditional probability, $P(Y = y, Z = z, X = x)$ by the constant $1/P(X = x)$.

Let us now examine the population created by the $do(X = x)$ operation and ask how the probability of each case changes as a result of this operation. The answer is given to us by the adjustment formula, which reads

$$P(y|do(x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

Multiplying and dividing the expression inside the sum by the propensity score $P(X = x|Z = z)$, we get

$$P(y|do(x)) = \frac{\sum_z P(Y = y|X = x, Z = z)P(X = x|Z = z)P(Z = z)}{P(X = x|Z = z)}$$

Upon realizing the numerator is none other but the pretreatment distribution of (X, Y, Z), we can write

$$P(y|do(x)) = \sum_z \frac{P(Y = y, X = x, Z = z)}{P(X = x|Z = z)}$$

and the answer becomes clear: each case $(Y = y, X = x, Z = z)$ in the population should boost its probability by a factor equals to $1/P(X = x|Z = z)$. (Hence the name "inverse probability weighting.")

This provides us with a simple procedure of estimating $P(Y = y|do(X = x))$ when we have finite samples. If we weigh each available sample by a factor $= 1/P(X = x|Z = z)$, we can then treat the reweighted samples as if they were generated from P_m, not P, and proceed to estimate $P(Y = y|do(x))$ accordingly.

This is best demonstrated in an example.

Table 3.3 returns to our Simpson's paradox example of the drug that seems to help men and women but to hurt the general population. We'll use the same data we used before but presented

Table 3.3 Joint probability distribution $P(X, Y, Z)$ for the drug-gender-recovery story of Chapter 1 (Table 1.1)

X	Y	Z	% of population
Yes	Yes	Male	0.116
Yes	Yes	Female	0.274
Yes	No	Male	0.01
Yes	No	Female	0.101
No	Yes	Male	0.334
No	Yes	Female	0.079
No	No	Male	0.051
No	No	Female	0.036

Table 3.4 Conditional probability distribution $P(Y, Z|X)$ for drug users ($X = yes$) in the population of Table 3.3

X	Y	Z	% of population
Yes	Yes	Male	0.232
Yes	Yes	Female	0.547
Yes	No	Male	0.02
Yes	No	Female	0.202

this time as a weighted table. In this case, X represents whether or not the patient took the drug, Y represents whether the patient recovered, and Z represents the patient's gender.

If we condition on "$X = Yes$," we get the data set shown in Table 3.4, which was formed in two steps. First, all rows with $X = No$ were excluded. Second, the weights given to the remaining rows were "renormalized," that is, multiplied by a constant so as to make them sum to one. This constant, according to Bayes' rule, is $1/P(X = yes)$, and $P(X = yes)$ in our example, is the combined weight of the first four rows of Table 3.3, which amounts to

$$P(X = yes) = 0.116 + 0.274 + 0.01 + 0.101 = 0.49$$

The result is the weight distribution in the four top rows of Table 3.4; the weight of each row has been boosted by a factor $1/0.49 = 2.041$.

Let us now examine the population created by the $do(X = yes)$ operation, representing a deliberate decision to administer the drug to the same population.

To calculate the distribution of weights in this population, we need to compute the factor $P(X = yes|Z = z)$ for each z, which, according to Table 3.3, is given by

$$P(X = yes|Z = Male) = \frac{(0.116 + 0.01)}{(0.116 + 0.01 + 0.334 + 0.051)} = 0.233$$

$$P(X = yes|Z = Female) = \frac{(0.274 + 0.101)}{(0.274 + 0.101 + 0.079 + 0.036)} = 0.765$$

Multiplying the gender-matching rows by $1/0.233$ and $1/0.765$, respectively, we obtain Table 3.5, which represents the postintervention distribution of the population of Table 3.3. The probability of recovery in this distribution can now be computed directly from the data, by summing the first two rows:

$$P(Y = yes|do(X = yes)) = 0.476 + 0.357 = 0.833$$

Table 3.5 Probability distribution for the population of Table 3.3 under the intervention $do(X = Yes)$, determined via the inverse probability method

X	Y	Z	% of population
Yes	Yes	Male	0.476
Yes	Yes	Female	0.357
Yes	No	Male	0.041
Yes	No	Female	0.132

Three points are worth noting about this procedure. First, the redistribution of weight is no longer proportional but quite discriminatory. Row #1, for instance, boosted its weight from 0.116 to 0.476, a factor of 4.1, whereas Row #2 is boosted from 0.274 to 0.357, a factor of only 1.3. This redistribution renders X independent of Z, as in a randomized trial (Figure 3.4).

Second, an astute reader would notice that in this example no computational savings were realized; to estimate $P(Y = yes|do(X = yes))$ we still needed to sum over all values of Z, males and females. Indeed, the savings become significant when the number of Z values is in the thousands or millions, and the sample size is in the hundreds. In such cases, the number of Z values that the inverse probability method would encounter is equal to the number of samples available, not to the number of possible Z values, which is prohibitive.

Finally, an important word of caution. The method of inverse probability weighing is only valid when the set Z entering the factor $1/P(X = x|Z = z)$ satisfies the backdoor criterion. Lacking this assurance, the method may actually introduce more bias than the one obtained through naive conditioning, which produces Table 3.4 and the absurdities of Simpson's paradox.

Up to this point, and in the following, we focus on unbiased estimation of causal effects. In other words, we focus on estimates that will converge to the true causal effects as the number of samples increases indefinitely.

This is obviously important, but it is not the *only* issue relevant to estimation. In addition, we must also address *precision*. Precision refers to the variability of our causal estimates if the number of samples is finite, and, in particular, how much our estimate would vary from experiment to experiment. Clearly, all other things being equal, we prefer estimation procedures with high precision in addition to their possessing little or no bias. Practically, high-precision estimates lead to shorter confidence intervals that quantify our level of certainty as to how our sample estimates describe the causal effect of interest. Most of our discussion does not address the "best," or most precise, way to estimate relevant causal means and effects but focuses on whether it is possible to estimate such quantities from observed data distributions, when the number of samples goes to infinity.

For example, suppose we wish to estimate the causal effect of X on Y (in a causal graph as above), where X and Y both reflect continuous variables. Suppose the effect of Z is to make both high and low values of X most commonly observed, with values close to the middle of the range of X much less common. Then, inverse probability weighting down-weights the extreme values of X on both ends of its range (since these are observed most frequently due to Z) and essentially focuses entirely on the "middle" values of X. If we then use a regression model to estimate the causal effect of X on Y (see Section 3.8, for example) using the reweighed observations to account for the role of Z, the resulting estimates will be very imprecise. In such cases, we usually seek for alternative estimation strategies that are more precise. While we do not pursue these alternatives in this book, it is important to emphasize that, in addition to seeing that causal effects be identified from the data, we must also devise effective strategies of using finite data to estimate effect sizes.

3.7 Mediation

Often, when one variable causes another, it does so both directly and indirectly, through a set of mediating variables. For instance, in our blood pressure/treatment/recovery example of Simpson's paradox, treatment is both a direct (negative) cause of recovery, and an indirect

(positive) cause, through the mediator of blood pressure—treatment decreases blood pressure, which increases recovery. In many cases, it is useful to know how much of variable X's effect on variable Y is direct and how much is mediated. In practice, however, separating these two avenues of causation has proved difficult.

Suppose, for example, we want to know whether and to what degree a company discriminates by gender (X) in its hiring practices (Y). Such discrimination would constitute a direct effect of gender on hiring, which is illegal in many cases. However, gender also affects hiring practices in other ways; often, for instance, women are more or less likely to go into a particular field than men, or to have achieved advanced degrees in that field. So gender may also have an indirect effect on hiring through the mediating variable of qualifications (Z).

In order to find the direct effect of gender on hiring, we need to somehow hold qualifications steady, and measure the remaining relationship between gender and hiring; with qualifications unchanging, any change in hiring would have to be due to gender alone. Traditionally, this has been done by conditioning on the mediating variable. So if $P(Hired|Female, Highly\ Qualified)$ is different from $P(Hired|Male, Highly\ Qualified)$, the reasoning goes, then there is a direct effect of gender on hiring.

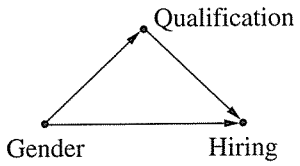


Figure 3.11 A graphical model representing the relationship between gender, qualifications, and hiring

In the example in Figure 3.11, this is correct. But consider what happens if there are confounders of the mediating variable and the outcome variable. For instance, income: People from higher income backgrounds are more likely to have gone to college and more likely to have connections that would help them get hired.

Now, if we condition on qualifications, we are conditioning on a collider. So if we don't condition on qualifications, indirect dependence can pass from gender to hiring through the path *Gender* → *Qualifications* → *Hiring*. But if we do condition on qualifications, indirect dependence can pass from gender to hiring through the path *Gender* → *Qualifications* ← *Income* → *Hiring*. (To understand the problem intuitively, note that by conditioning on qualification, we will be comparing men and women at different levels of income, because income must change to keep qualification constant.) No matter how you look at it, we're not getting the true direct effect of gender on hiring. Traditionally, therefore, statistics has had to abandon a huge class of potential mediation problems, where the concept of "direct effect" could not be defined, let alone estimated.

Luckily, we now have a conceptual way of holding the mediating variable steady without conditioning on it: We can intervene on it. If, instead of conditioning, we fix the qualifications, the arrow between gender and qualifications (and the one between income and qualifications) disappears, and no spurious dependence can pass through it. (Of course, it would be impossible for us to literally change the qualifications of applicants, but recall, this is a theoretical intervention of the kind discussed in the previous section, accomplished by choosing a proper adjustment.) So for any three variables X, Y, and Z, where Z is a mediator between X and Y,

the *controlled direct effect* (CDE) on Y of changing the value of X from x to x' is defined as

$$CDE = P(Y = y|do(X = x), do(Z = z)) - P(Y = y|do(X = x'), do(Z = z)) \quad (3.18)$$

The obvious advantage of this definition over the one based on conditioning is its generality; it captures the intent of "keeping Z constant" even in cases where the $Z \rightarrow Y$ relationship is confounded (the same goes for the $X \rightarrow Z$ and $X \rightarrow Y$ relationships). Practically, this definition assures us that in any case where the intervened probabilities are identifiable from the observed probabilities, we can estimate the direct effect of X on Y. Note that the direct effect may differ for different values of Z; for instance, it may be that hiring practices discriminate against women in jobs with high qualification requirements, but they discriminate against men in jobs with low qualifications. Therefore, to get the full picture of the direct effect, we'll have to perform the calculation for every relevant value z of Z. (In linear models, this will not be necessary; for more information, see Section 3.8.)

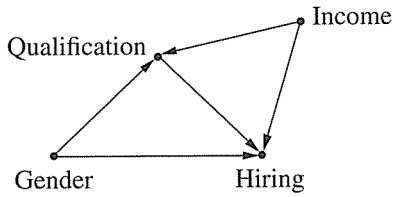


Figure 3.12 A graphical model representing the relationship between gender, qualifications, and hiring, with socioeconomic status as a mediator between qualifications and hiring

How do we estimate the direct effect when its expression contains two *do*-operators? The technique is more or less the same as the one employed in Section 3.2, where we dealt with a single *do*-operator by adjustment. In our example of Figure 3.12, we first notice that there is no backdoor path from X to Y in the model, hence we can replace $do(x)$ with simply conditioning on x (this essentially amounts to adjusting for all confounders). This results in

$$P(Y = y|X = x, do(Z = z)) - P(Y = y|X = x', do(Z = z))$$

Next, we attempt to remove the $do(z)$ term and notice that two backdoor paths exist from Z to Y, one through X and one through I. The first is blocked (since X is conditioned on) and the second can be blocked if we adjust for I. This gives

$$\sum_i [P(Y = y|X = x, Z = z, I = i) - P(Y = y|X = x', Z = z, I = i)]P(I = i)$$

The last formula is *do*-free, which means it can be estimated from nonexperimental data.

In general, the CDE of X on Y, mediated by Z, is identifiable if the following two properties hold:

1. There exists a set S_1 of variables that blocks all backdoor paths from Z to Y.
2. There exists a set S_2 of variables that blocks all backdoor paths from X to Y, after deleting all arrows entering Z.

If these two properties hold in a model M, then we can determine $P(Y = y|do(X = x), do(Z = z))$ from the data set by adjusting for the appropriate variables, and estimating the conditional probabilities that ensue. Note that condition 2 is not necessary in randomized trials, because randomizing X renders X parentless. The same is true in cases where X is judged to be exogenous (i.e., "as if" randomized), as in the aforementioned gender discrimination example.

It is even trickier to determine the indirect effect than the direct effect, because there is simply no way to condition away the direct effect of X on Y. It's easy enough to find the total effect and the direct effect, so some may argue that the indirect effect should just be the difference between those two. This may be true in linear systems, but in nonlinear systems, differences don't mean much; the change in Y might, for instance, depend on some interaction between X and Z—if, as we posited above, women are discriminated against in high-qualification jobs and men in low-qualification jobs, subtracting the direct effect from the total effect would tell us very little about the effect of gender on hiring as mediated by qualifications. Clearly, we need a definition of indirect effect that does not depend on the total or direct effects.

We will show in Chapter 4 that these difficulties can be overcome through the use of *counterfactuals*, a more refined type of intervention that applies at the individual level and can be computed from structural models.

3.8 Causal Inference in Linear Systems

One of the advantages of the causal methods we have introduced in this book is that they work regardless of the type of equations that make up the model in question. d-separation and the backdoor criterion make no assumptions about the form of the relationship between two variables—only that the relationship exists.

However, showcasing and explaining causal methods from a nonparametric standpoint has limited our ability to present the full power of these methods as they play out in linear systems—the arena where traditional causal analysis has primarily been conducted in the social and behavioral sciences. This is unfortunate, as many statisticians work extensively in linear systems, and nearly all statisticians are very familiar with them.

In this section, we examine in depth what causal assumptions and implications look like in systems of linear equations and how graphical methods can help us answer causal questions posed in those systems. This will serve as both a reinforcement of the methods we applied in nonparametric models and as a useful aid for those hoping to apply causal inference specifically in the context of linear systems.

For instance, we might want to know the effect of birth control use on blood pressure after adjusting for confounders; the total effect of an after-school study program on test scores; the direct effect, unmediated by other variables, of the program on test scores; or the effect of enrollment in an optional work training program on future earnings, when enrollment and earnings are confounded by a common cause (e.g., motivation). Such questions, invoking continuous variables, have traditionally been formulated as linear equation models with only minor attention to the unique causal character of those equations; we make this character unambiguous.

In all models used in this section, we make the strong assumption that the relationships between variables are linear, and that all error terms have Gaussian (or "normal") distributions (in some cases, we only need to assume symmetric distributions). This assumption provides an

enormous simplification of the procedure needed for causal analysis. We are all familiar with the bell-shaped curve that characterizes the normal distribution of one variable. The reason it is so popular in statistics is that it occurs so frequently in nature whenever a phenomenon is a byproduct of many noisy microprocesses that add up to produce macroscopic measurements such as height, weight, income, or mortality. Our interest in the normal distribution, however, stems primarily from the way several normally distributed variables combine to shape their joint distribution. The assumption of normality gives rise to four properties that are of enormous use when working with linear systems:

1. Efficient representation
2. Substitutability of expectations for probabilities
3. Linearity of expectations
4. Invariance of regression coefficients.

Starting with two normal variables, X and Y, we know that their joint density forms a three-dimensional cusp (like a mountain rising above the X–Y plane) and that the planes of equal height on that cusp are ellipses like those shown in Figure 1.2. Each such ellipse is characterized by five parameters: $\mu_X, \mu_Y, \sigma_X, \sigma_Y$, and ρ_{XY}, as defined in Sections 1.3.8 and 1.3.9. The parameters μ_X and μ_Y specify the location (or the center of gravity) of the ellipse in the X–Y plane, the variances σ_X and σ_Y specify the spread of the ellipse along the X and Y dimensions, respectively, and the correlation coefficient ρ_{XY} specifies its orientation. In three dimensions, the best way to depict the joint distribution is to imagine an oval football suspended in the X–Y–Z space (Figure 1.2); every plane of constant Z would then cut the football in a two-dimensional ellipse like the ones shown in Figure 1.1.

As we go to higher dimensions, and consider a set of N normally distributed variables $X_1, X_2, \ldots, X_N$, we need not concern ourselves with additional parameters; it is sufficient to specify those that characterize the $N(N-1)/2$ pairs of variables, (X_i, X_j). In other words, the joint density of $(X_1, X_2, \ldots, X_N)$ is fully specified once we specify the bivariate density of (X_i, X_j), with i and j $(i \neq j)$ ranging from 1 to N. This is an enormously useful property, as it offers an extremely parsimonious way of specifying the N-variable joint distribution. Moreover, since the joint distribution of each pair is specified by five parameters, we conclude that the joint distribution requires at most $5 \times N(N-1)/2$ parameters (means, variances, and covariances), each defined by expectation. In fact, the total number of parameters is even smaller than this, namely $2N + N(N-1)/2$; the first term gives the number of mean and variance parameters, and the second the number of correlations.

This brings us to another useful feature of multivariate normal distributions: they are fully defined by expectations, so we need not concern ourselves with probability tables as we did when dealing with discrete variables. Conditional probabilities can be expressed as conditional expectations, and notions such as conditional independence that define the structure of graphical models can be expressed in terms of equality relationships among conditional expectations. For instance, to express the conditional independence of Y and X, given Z,

$$P(Y|X,Z) = P(X|Z)$$

we can write

$$E[Y|X,Z] = E[Y|Z]$$

(where Z is a set of variables).

This feature of normal systems gives us an incredibly useful ability: Substituting expectations for probabilities allows us to use regression (a predictive method) to determine causal information. The next useful feature of normal distributions is their linearity: every conditional expectation $E[Y|X_1, X_2, \ldots, X_n]$ is given by a linear combination of the conditioning variables. Formally,

$$E[Y|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = r_0 + r_1x_1 + r_2x_2 + \cdots + r_nx_n$$

where each of the slopes $r_1, r_2, \ldots, r_n$ is a partial regression coefficient of as defined in Sections 1.3.10 and 1.3.11.

The magnitudes of these slopes do not depend on the values $x_1, x_2, \ldots, x_n$ of the conditioning variables, called *regressors*; they depend only on which variables are chosen as regressors. In other words, the sensitivity of Y to the measurement $X_i = x_i$ does not depend on the measured values of the other variables in the regression; it depends only on which variables we choose to measure. It doesn't matter whether $X_i = 1, X_i = 2$, or $X_i = 312.3$; as long as we regress Y on X_i, the coefficient r_j of X_j will remain the same.

This unique and useful feature of normal distributions is illustrated in Figures 1.1 and 1.2 of Chapter 1. Figure 1.1 shows that regardless of what level of age we choose, the slope of Y on X at that level is the same. If, however, we do not hold age constant (i.e., we do not regress on it), the slope becomes vastly different, as is shown in Figure 1.2.

The linearity assumption also permits us to fully specify the functions in the model by annotating the causal graph with a *path coefficient* (or structural coefficient) along each edge. The path coefficient β along the edge $X \rightarrow Y$ quantifies the contribution of X in the function that defines Y in the model. For instance, if the function defines $Y = 3X + U$, the path coefficient of $X \rightarrow Y$ will be 3. The path coefficients $\beta_1, \beta_2, \ldots, \beta_n$ are fundamentally different from the regression coefficients $r_1, r_2, \ldots, r_n$ that we discussed in Section 1.3. The former are "structural" or "causal," whereas the latter are statistical. The difference is explained in the next section.

Many of the regression methods we discuss are far more general, applying in situations where the variables $X_1, \ldots, X_k$ follow distribution far from multivariate Normal; for example, when some of the X_i's are categorical or even binary. Such generalizations also therefore allow the conditional mean $E(Y|X_1 = x_1, \ldots, X_k = x_k)$ to include nonlinear combinations of the X_i's, including such terms as X_1X_2, for example, to allow for effect modification, or interaction. Since we are conditioning on the values of the X_i's, it is usually not necessary to enforce a distributional assumption for such variables. Nevertheless, the full multivariate Normal scenario provides considerable insight into structural causal models.

3.8.1 *Structural versus Regression Coefficients*

As we are now about to deal with linear models, and thus, as a matter of course, with regression-like equations, it is of paramount importance to define the difference between regression equations and the structural equations we have used in SCMs throughout the book. A regression equation is descriptive; it makes no assumptions about causation. When we write $y = r_1x + r_2z + \epsilon$, as a regression equation, we are not saying that X and Z cause Y. We merely confess our need to know which values of r_1 and r_2 would make the equation $y = r_1x + r_2z$

the best linear approximation to the data, or, equivalently, the best linear approximation of $E(y|x, z)$.

Because of this fundamental difference between structural and regression equations, some books distinguish them by writing an arrow, instead of equality sign, in structural equations, and some distinguish the coefficients by using a different font. We distinguish them by denoting structural coefficients as α, β, and so on, and regression coefficients as r_1, r_2, and so on. In addition, we distinguish between the stochastic "error terms" that appear in these equations. Errors in regression equations are denoted ϵ_1, ϵ_2, and so on, as in Eq. (1.24), and those in structural equations by U_1, U_2, and so on. as in SCM 1.5.2. The former denote the residual errors in observation, after fitting the equation $y = r_1x + r_2z$ to data, whereas the latter represent latent factors (sometimes called "disturbances" or "omitted variables") that influence Y and are not themselves affected by X. The former are human-made (due to imperfect fitting); the latter are nature-made.

Though they are not causally binding themselves, regression equations are of significant use in the study of causality as it pertains to linear systems. Consider: In Section 3.2, we were able to express the effects of interventions in terms of conditional probabilities, as, for example, in the adjustment formula of Eq. (3.5). In linear systems, the role of conditional probabilities will be taken over by regression coefficients, since these coefficients represent the dependencies induced by the model and, in addition, they are easily estimable using least square analyses. Similarly, whereas the testable implications of nonparametric models are expressed in the form of conditional independencies, these independencies are signified in linear models by vanishing correlation coefficients, like those discussed in Section 1.3.11. Specifically, given the regression equation

$$y = r_0 + r_1x_1 + r_2x_2 + \cdots + r_nx_n + \epsilon$$

if $r_i = 0$, then Y is independent of X_i conditional on all the other regression variables.

3.8.2 *The Causal Interpretation of Structural Coefficients*

In a linear system, every path coefficient stands for the direct effect of the independent variable, X, on the dependent variable, Y. To see why this is so, we refer to the interventional definition of direct effect given in Section 3.7 (Eq. (3.18)), which calls for computing the change in Y as X increases by one unit whereas all other parents of Y are held constant. When we apply this definition to any linear system, regardless of whether the disturbances are correlated or not, the result will be the path coefficient on the arrow $X \rightarrow Y$.

Consider, for example, the model in Figure 3.13, and assume we wish to estimate the direct effect of Z on Y. The structural equations in the fully specified model read:

$$\begin{aligned} X &= U_X \\ Z &= aX + U_Z \\ W &= bX + cZ + U_W \\ Y &= dZ + eW + U_Y \end{aligned}$$

Writing Eq. (3.18) in expectation form, we obtain

$$DE = E[Y|do(Z = z + 1), do(W = w)] - E[Y|do(Z = z), do(W = w)]$$

since W is the only other parent of Y in the graph. Applying the *do* operators by deleting the appropriate equations from the model, the postincrease term in DE becomes $d(z + 1) + ew$ and the preincrease term becomes $dz + ew$. As expected, the difference between the two is d—the path coefficient between Z and Y. Note that the license to reduce the equation in this way comes directly from the definition of the *do*-operator (Eq. (3.18)) making no assumption about correlations among the U factors; the equality $DE = d$ would be valid even if the error term U_Y were correlated with U_Z, though this would have made d nonidentifiable. The same goes for the other direct effects; every structural coefficient represents a direct effect, regardless of how the error terms are distributed. Note also that variable X, as well as the coefficients a, b, and c, do not enter into this computation, because the "surgeries" required by the *do* operators remove them from the model.

That is all well and good for the direct effect. Suppose, however, we wish to calculate the *total* effect of Z on Y.

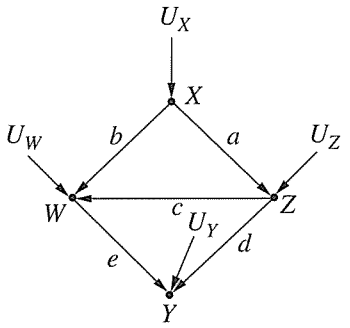


Figure 3.13 A graphical model illustrating the relationship between path coefficients and total effects

In a linear system, the total effect of X on Y is simply the sum of the products of the coefficients of the edges on every nonbackdoor path from X to Y.

That's a bit of a mouthful, so think of it as a process: To find the total effect of X on Y, first find every nonbackdoor path from X to Y; then, for each path, multiply all coefficients on the path together; then add up all the products.

The reason for this identity lies in the nature of SCMs. Consider again the graph of Figure 3.13. Since we want to find the total effect of Z on Y, we should first intervene on Z, removing all arrows going into Z, then express Y in terms of Z in the remaining model. This we can do with a little algebra:

$$\begin{aligned} Y &= dZ + eW + U_Y \\ &= dZ + e(bX + cZ) + U_Y + eU_W \\ &= (d + ec)Z + ebX + U_Y + eU_W \end{aligned}$$

The final expression is in the form $Y = \tau Z + U$, where $\tau = d + ec$ and U contains only terms that do not depend on Z in the modified model. An increase of a single unit in Z, therefore, will increase Y by τ—the definition of the total effect. A quick examination will show that τ

is the sum of the products of the coefficients on the two nonbackdoor paths from Z to Y. This will be the case in all linear models; algebra demands it. Moreover, the sum of product rule will be valid regardless of the distributions of the U variables and regardless of whether they are dependent or independent.

3.8.3 *Identifying Structural Coefficients and Causal Effect*

Thus far, we have expressed the total and direct effects in terms of path coefficients, assuming that the latter are either known to us a priori or estimated from interventional experiments. We now tackle a much harder problem; estimating total and direct effects from nonexperimental data. This problem is known as "identifiability" and, mathematically, it amounts to expressing the path coefficients associated with the total and direct effects in terms of the covariances σ_{XY} or regression coefficients $R_{YX \cdot Z}$, where X and X are any two variables in the model, and Z a set of variables in the model (Eqs. (1.27) and (1.28) and Section 1.3.11).

In many cases, however, it turns out that to identify direct and total effects, we do not need to identify each and every structural parameter in the model. Let us first demonstrate with the total effect, τ. The backdoor criterion gives us the set Z of variables we need to adjust for in order to determine the causal effect of X on Y. How, though, do we make use of the criterion to determine effects in a linear system? In principle, once we obtain the set, Z, we can estimate the conditional expectation of Y given X and Z and, then, averaging over Z, we can use the resultant dependence between Y and X to measure the effect of X on Y. We need only translate this procedure to the language of regression.

The translation is rather simple. First, we find a set of covariates Z that satisfies the backdoor criterion from X to Y in the model. Then, we regress Y on X and Z. The coefficient of X in the resulting equation represents the true causal effect of X on Y. The reasoning for this is similar to the reasoning we used to justify the backdoor criterion in the first place—regressing on Z adds those variables into the equation, blocking all backdoor paths from X and Y, thus preventing the coefficient of X from absorbing the spurious information those paths contain.

For example, consider a linear model that complies with the graph in Figure 3.14. If we want to find the total causal effect of X on Y, we first determine, using the backdoor criterion, that we must adjust for T. So we regress Y on X and T, using the regression equation $y = r_X X +$

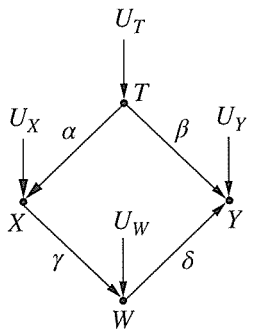


Figure 3.14 A graphical model in which X has no direct effect on Y, but a total effect that is determined by adjusting for T

$r_T T + \epsilon$. The coefficient r_X represents the total effect of X on Y. Note that this identification was possible without identifying any of the model parameters and without measuring variable W; the graph structure in itself gave us the license to ignore W, regress Y on T and X only, and identify the total effect (of X on Y) with the coefficient of X in that regression.

Suppose now that instead of the total causal effect, we want to find X's direct effect on Y. In a linear system, this direct effect is the structural coefficient α in the function $y = \alpha x + \beta z + \cdots + U_Y$ that defines Y in the system. We know from the graph of Figure 3.14 that $\alpha = 0$, because there is no direct arrow from X to Y. So, in this particular case, the answer is trivial: the direct effect is zero. But in general, how do we find the magnitude of α from data, if the model does not determine its value?

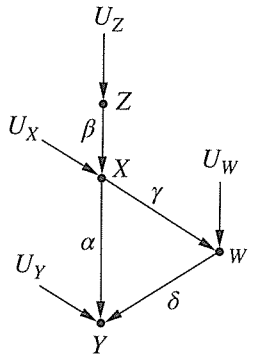


Figure 3.15 A graphical model in which X has direct effect α on Y

We can invoke a procedure similar to backdoor, except that now, we need to block not only backdoor paths but also indirect paths going from X to Y. First, we remove the edge from X to Y (if such an edge exists), and call the resulting graph G_α. If, in G_α, there is a set of variables Z that d-separates X and Y, then we can simply regress Y on X and Z. The coefficient of X in the resulting equation will equal the structural coefficient α.

The procedure above, which we might as well call "The Regression Rule for Identification" provides us with a quick way of determining whether any given parameter (say α) can be identified by ordinary least square (OLS) regression and, if so, what variables should to into the regression equation. For example, in the linear model of Figure 3.15, we can find the direct effect of X on Y by this method. First, we remove the edge between X and Y and get the graph G_α shown in Figure 3.16. It's easy to see that in this new graph, W d-separates X and Y. So we regress Y on X and W, using the regression equation $Y = r_X X + r_W W + \epsilon$. The coefficient r_X is the direct effect of X on Y.

Summarizing our observations thus far, two interesting features emerge. First, we see that, in linear systems, regression serves as the major tool for the identification and estimation of causal effects. To estimate a given effect, all we need to do is to write down a regression equation and specify (1) what variables should be included in the equation and (2) which of the coefficients in that equation represents the effect of interest. The rest is routine least square analysis on the sampled data which, as we remarked before, is facilitated by a variety of extremely efficient software packages. Second, we see that, as long as the U variables are independent of each

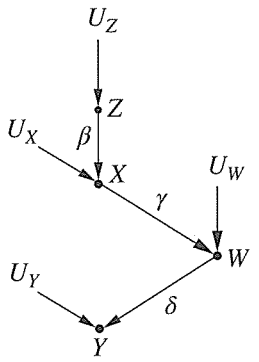


Figure 3.16 By removing the direct edge from X to Y and finding the set of variables $\{W\}$ that d-separate them, we find the variables we need to adjust for to determine the direct effect of X on Y

others, and all variables in the graph are measured, every structural parameter can be identified in this manner, namely, there is at least one identifying regression equation in which one of the coefficients corresponds to the parameter we seek to estimate. One such equation is obviously the structural equation itself, with the parents of Y serving as regressors. But there may be several other identifying equations, with possibly better features for estimation and graphical analysis can reveal them all (see Study question 3.8.1(c)). Moreover, when some variables are not measured, or when some error terms are correlated, the task of finding an identifying regression from the structural equations themselves would normally be insurmountable; the G_α procedure then becomes indispensable (see Study question 3.8.1(d)).

Remarkably, the regression rule procedure has eluded investigators for almost a century, possibly because it is extremely difficult to articulate in algebraic, nongraphical terms.

Suppose, however, there is no set of variables that d-separates X and Y in G_α. For instance, in Figure 3.17, X and Y have an unobserved common cause represented by the dashed

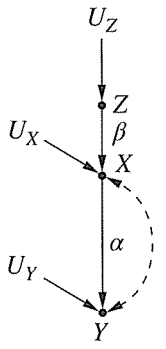


Figure 3.17 A graphical model in which we cannot find the direct effect of X on Y via adjustment, because the dashed double-arrow arc represents the presence of a backdoor path between X and Y, consisting of unmeasured variables. In this case, Z is an instrument with regard to the effect of X on Y that enables the identification of α

double-arrowed arc. Since it hasn't been measured, we can't condition on it, so X and Y will always be dependent through it. In this particular case, we may use an *instrumental variable* to determine the direct effect. A variable is called an "instrument" if it is d-separated from Y in G_α and, it is d-connected to X. To see why such a variable enables us to identify structural coefficients, we take a closer look at Figure 3.17.

In Figure 3.17, Z is an instrument with regard to the effect of X on Y because it is d-connected to X and d-separated from Y in G_α. We regress X and Y on Z separately, yielding the regression equations $y = r_1 z + \epsilon$ and $x = r_2 z + \epsilon$, respectively. Since Z emits no backdoors, r_2 equals β and r_1 equals the total effect of Z on Y, $\beta\alpha$. Therefore, the ratio r_1/r_2 provides the desired coefficient α. This example illustrates how direct effects can be identified from total effects but not the other way around.

Graphical models provide us with a procedure for finding all instrumental variables in a system, though the procedure for enumerating them is beyond the scope of this book. Those interested in learning more can (see Chen and Pearl 2014; Kyono 2010).

Study questions

Study question 3.8.1

Model 3.1

$$Y = aW_3 + bZ_3 + cW_2 + U \qquad X = t_1 W_1 + t_2 Z_3 + U'$$
$$W_3 = c_3 X + U_3' \qquad W_1 = a_1' Z_1 + U_1'$$
$$Z_3 = a_3 Z_1 + b_3 Z_2 + U_3 \qquad Z_1 = U_1$$
$$W_2 = c_2 Z_2 + U_2' \qquad Z_2 = U_2$$

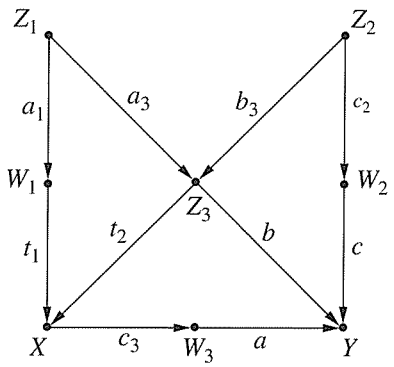


Figure 3.18 Graph corresponding to Model 3.1 in Study question 3.8.1

Given the model depicted above, answer the following questions:

(All answers should be given in terms of regression coefficients in specified regression equations.)

(a) *Identify three testable implications of this model.*
(b) *Identify a testable implication assuming that only X, Y, W_3, and Z_3 are observed.*
(c) *For each of the parameters in the model, write a regression equation in which one of the coefficients is equal to that parameter. Identify the parameters for which more than one such equation exists.*
(d) *Suppose X, Y, and W_3 are the only variables observed. Which parameters can be identified from the data? Can the total effect of X on Y be estimated?*
(e) *If we regress Z_1 on all other variables in the model, which regression coefficient will be zero?*
(f) *The model in Figure 3.18 implies that certain regression coefficients will remain invariant when an additional variable is added as a regressor. Identify five such coefficients with their added regressors.*
(g) *Assume that variables Z_2 and W_2 cannot be measured. Find a way to estimate b using regression coefficients. [Hint: Find a way to turn Z_1 into an instrumental variable for b.]*

3.8.4 Mediation in Linear Systems

When we can assume linear relationships between variables, mediation analysis becomes much simpler than the analysis conducted in nonlinear or nonparametric systems (Section 3.7). Estimating the direct effect of X on Y, for instance, amounts to estimating the path coefficient between the two variables, and this reduces to estimating correlation coefficients, using the techniques introduced in Section 3.8.3. The indirect effect, similarly, is computed via the difference $IE = \tau - DE$, where τ, the total effect, can be estimated by regression in the manner shown in Figure 3.14. In nonlinear systems, on the other hand, the direct effect is defined through expressions such as (3.18), or

$$DE = E[Y|do(x, z)] - E[Y|do(x', z)]$$

where $Z = z$ represents a specific stratum of all other parents of Y (besides X). Even when the identification conditions are satisfied, and we are able to reduce the *do*() operators (by adjustments) to ordinary conditional expectations, the result will still depend on the specific values of x, x', and z. Moreover, the indirect effect cannot be given a definition in terms as *do*-expressions, since we cannot disable the capacity of Y to respond to X by holding variables constant. Nor can the indirect effect be defined as the difference between the total and direct effects, since differences do not faithfully reflect operations in nonlinear systems to X.

Such an operation will be introduced in Chapter 4 (Sections 4.4.5 and 4.5.2) using the language of counterfactuals.

Bibliographical Notes for Chapter 3

Study question 3.3.2 is a version of Lord's paradox (Lord 1967), and is described in Glymour (2006), Hernández-Díaz et al. (2006), Senn (2006), and Wainer (1991). A unifying treatment is given in Pearl (2014b). The definition of the *do*-operator and "ACE" in terms of a modified model, has its conceptual origin with the economist Trygve Haavelmo (1943), who was the first

to simulate interventions by modifying equations in the model (see Pearl (2014b) for historical account). Strotz and Wold (1960) later advocated "wiping out" the equation determining X, and Spirtes et al. (1993) gave it a graphical representation in a form of a "manipulated graph." The "adjustment formula" of Eq. (3.5) as well as the "truncated product formula" first appeared in Spirtes et al. (1993), though these are implicit in the G-computation formula of Robins (1986), which was derived using counterfactual assumptions (see Chapter 4). The backdoor criterion of Definition 3.3.1 and its implications for adjustments were introduced in Pearl (1993). The front-door criterion and a general calculus for identifying causal effects (named *do*-calculus) were introduced in Pearl (1995) and further improved in Tian and Pearl (2002) and Shpitser and Pearl (2007). Section 3.7, and the identification of conditional interventions and c-specific effects is based on (Pearl 2009, pp. 113–114). Its extension to dynamic, time-varying policies is described in Pearl and Robins (1995) and (Pearl 2009, pp. 119–126). The role of covariate-specific effects in assessing interaction, moderation or effect modification is described in Morgan and Winship 2014 and Vanderweele (2015), whereas applications of Rule 2 to the detection of latent heterogeneity are described in Pearl (2015b). Additional discussions on the use of inverse probability weighting (Section 3.6) can be found in Hernán and Robins (2006). Our discussion of mediation (Section 3.7) and the identification of CDEs are based on Pearl (2009, pp. 126–130), whereas the fallibility of "conditioning" on a mediator to assess direct effects is demonstrated in Pearl (1998) as well as Cole and Hernán (2002).

The analysis of mediation has become extremely active in the past 15 years, primarily due to the advent of counterfactual logic (see Section 4.4.5); a comprehensive account of this progress is given in Vanderweele (2015). A tutorial survey of causal inference in linear systems (Section 3.8), focusing on parameter identification, is provided by Chen and Pearl (2014). Additional discussion on the confusion of regression versus structural equations can be found in Bollen and Pearl (2013).

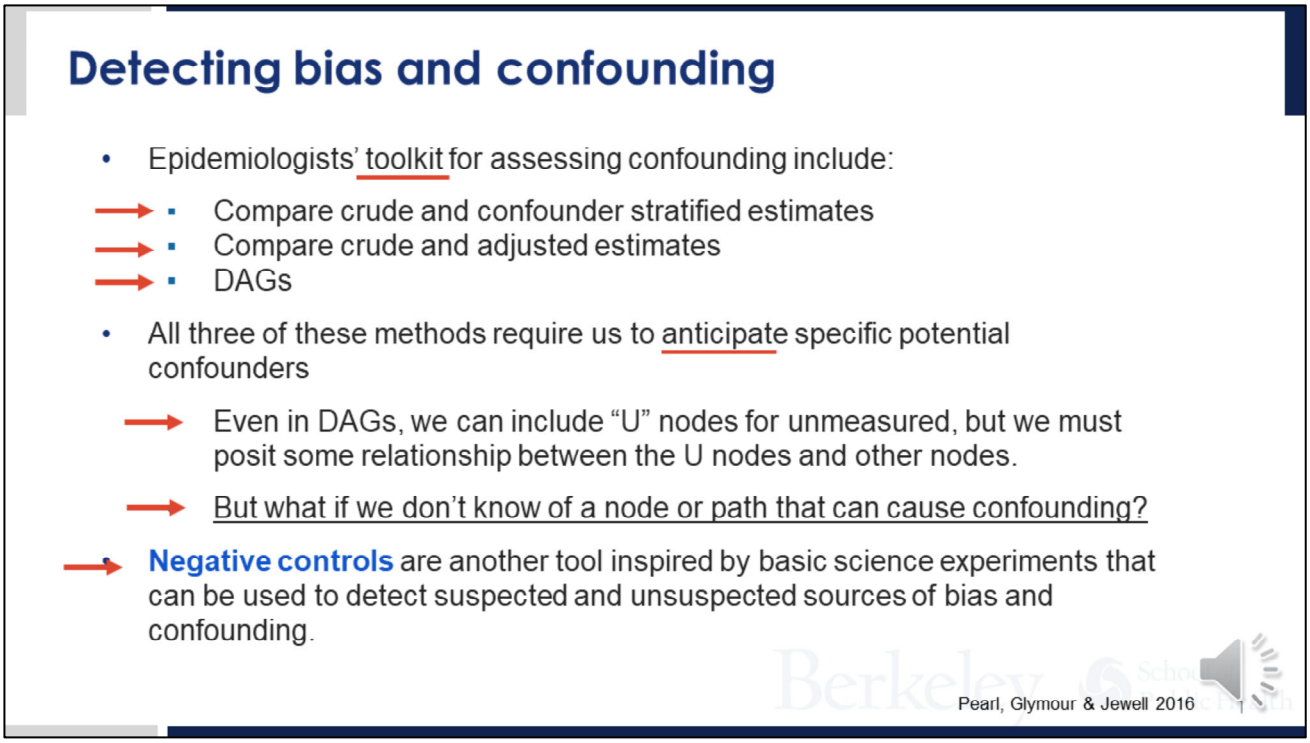
A classic, and still the best textbook on the relationships between structural and regession coefficients is Heise (1975) (available online: http://www.indiana.edu/~socpsy/public_files/CausalAnalysis.zip). Other classics are Duncan (1975), Kenny (1979), and Bollen (1989). Classical texts, however, fall short of providing graphical tools of identification, such as those invoking backdoor and G_α (see Study question 3.8.1). A recent exception is Kline (2016).

Introductions to instrumental variables can be found in Greenland (2000) and in many textbooks of econometrics (e.g., Bowden and Turkington 1984, Wooldridge 2013). Generalized instrumental variables, extending the classical definition of Section 3.8.3 were introduced in Brito and Pearl (2002).

The program DAGitty (which is available online: http://www.dagitty.net/dags.html), permits users to search the graph for generalized instrumental variables, and reports the resulting IV estimators (Textor et al. 2011).

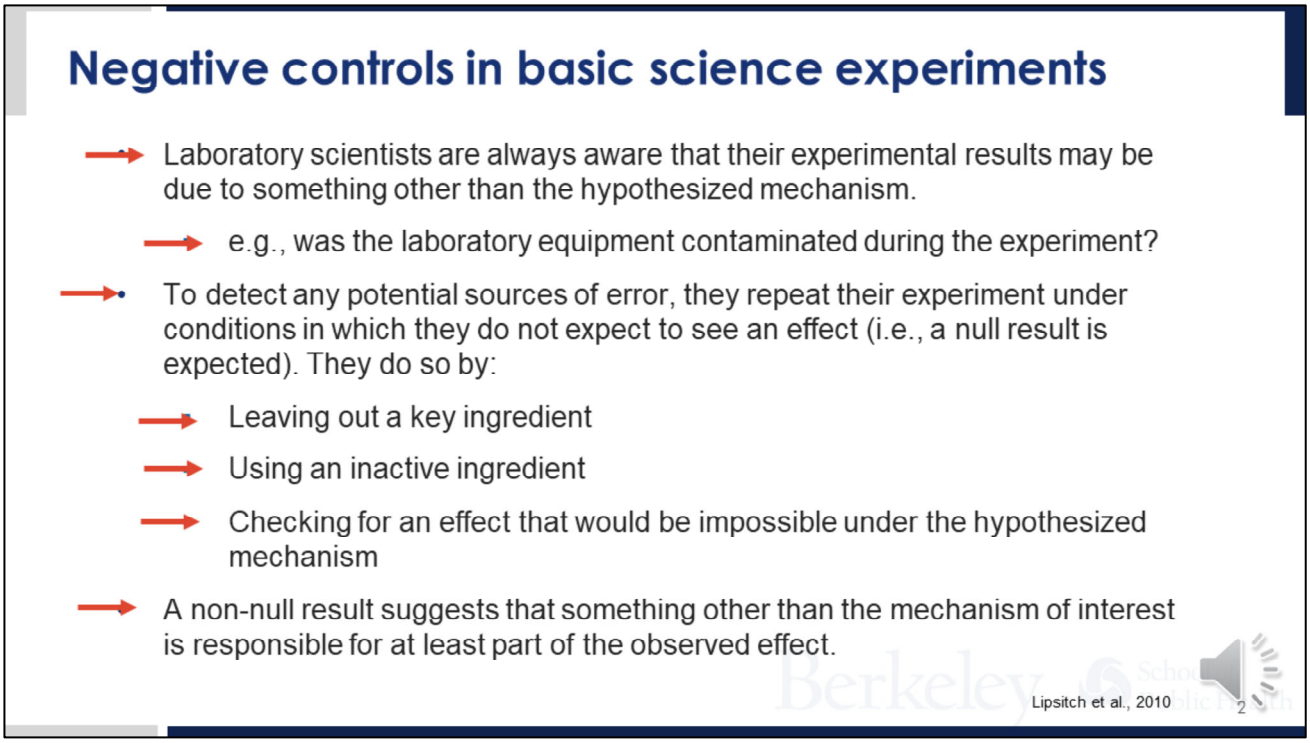
Lecture: Negative Controls

In this video, I'll talk about how we can use something called negative controls to detect bias and confounding in epidemiologic studies.



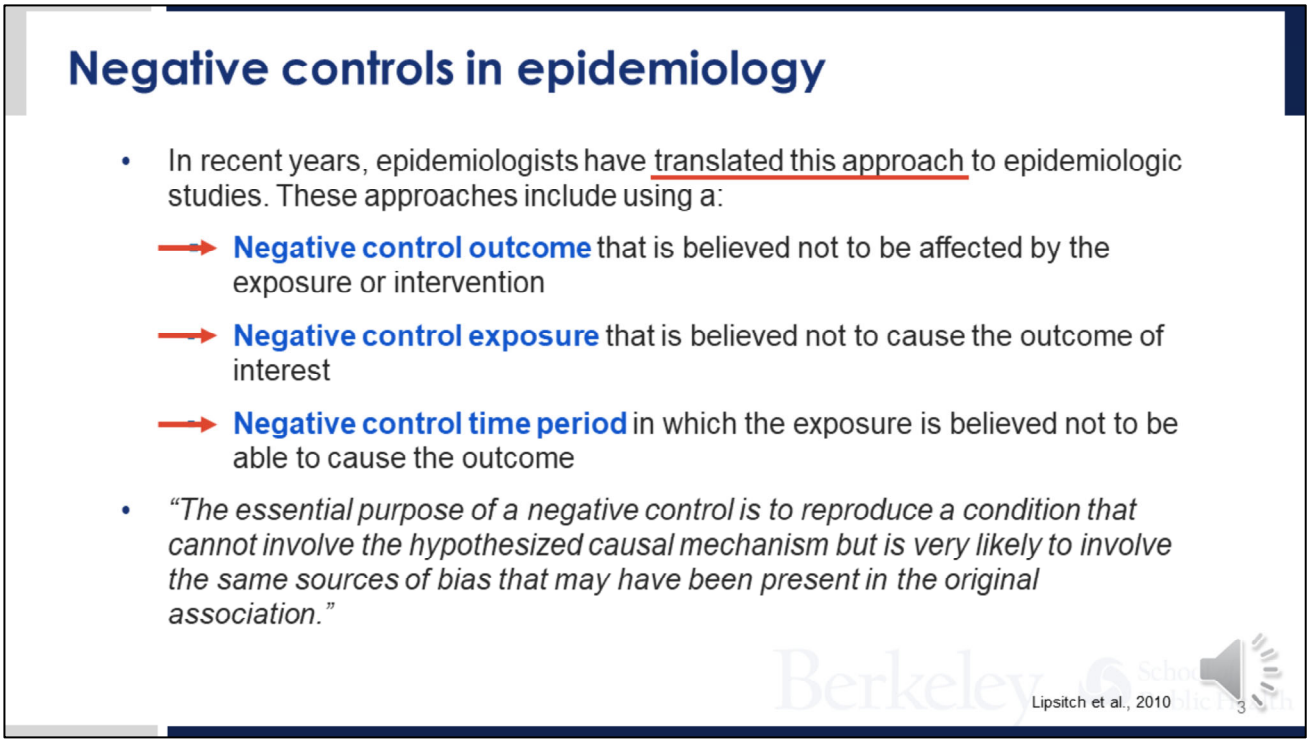
In epidemiology, our *toolkit for assessing confounding, includes *comparing crude and confounder stratified estimates, *comparing crude and adjusted estimates, and using *directed acyclic graphs. All three of these methods require us to *anticipate potential confounders. In other words, we have to think of particular variables that might confound the relationship between exposure and outcome.

*In DAGs, we can use "U" nodes that indicate unmeasured potential confounders. But even with such U nodes, we have to posit a relationship between U nodes and other nodes in the DAG. *What if you don't know of a node or a path or, speaking outside of the DAG context, of a variable that can cause confounding? Well, this is where negative controls are potentially, extremely powerful. *They're a tool that was inspired by basic science experiments. And they can be used to detect suspected and unsuspected sources of bias and confounding in our studies.



Let's start with talking about how negative controls are used in basic science experiments. *So lab scientists are always aware that their experimental results may not be due to the mechanism that they hypothesize. *There may have been some contamination of laboratory equipment or something else that went wrong in the process of conducting the experiment that caused them to see the finding that they saw. It might not be due to the mechanism that they suspected. *And so in lab science, it's very common to repeat an experiment under different conditions to see if the result conforms with expectation.

And, most of the time, what they're doing is they're running an experiment such that they don't expect to see any finding. In other words, they want to get a null result. *And so they do this by leaving out a key ingredient that would cause them to see a result, *using an inactive version of an ingredient, or *checking for an effect that would not be possible under the hypothesized mechanism of interest. And what they want to see when trying these three things is null results-- null result. *If they find that the result is not null, it suggests that something other than the mechanism of interest could be responsible for at least part of the observed effect.

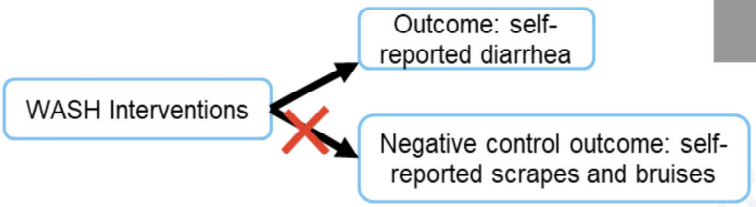


In recent years, epidemiologists have *translated this approach to their studies. And there are three primary different ways that epidemiologists have adapted this. First is to use a negative control outcome. So that's an outcome, or disease, that's believed to not be affected by the exposure or intervention under study. You can also use a negative control exposure that's believed not to cause the outcome of interest. And then, we can also use a negative control time period. So if the exposure outcome relationship is suspected to only occur during a certain time of year, you could look for that relationship during another time of year when it's not suspected to occur.

This paper cited down here at the bottom by Lipsitch et al. in 2010, introduced this concept. And here's a quote from the paper. "The essential purpose of a negative control is to reproduce a condition that cannot involve the hypothesis causal mechanism but is very likely to involve the same sources of bias that may have been present in the original association." And this is the key part here-- a good negative control needs to involve the same sources of bias. If that's not the case, then the negative control analysis will not achieve the goal.

Example: negative control outcome

- In a trial of the effects of water and sanitation interventions on self-reported diarrhea, measure the effect of interventions on self-reported scrapes and bruises
- Both outcomes are self-reported and may involve similar sources of bias. However, scrapes and bruises are not affected by the intervention.



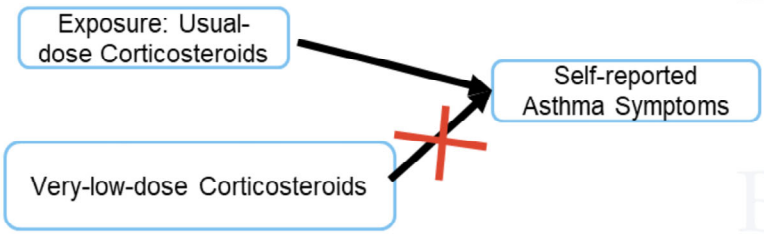
Lipsitch et al., 2010

Let's go over a few examples. So the first is looking at a negative control outcome. In a *trial of water and sanitation interventions, the *outcome is self-reported diarrhea, and *the negative control outcome could then be self-reported scrapes and bruises. Now, we don't expect the water and sanitation intervention to affect scrapes and bruises.

Both the real outcome and the negative control outcome are self-reported. The purpose of this is to try to capture the same source of bias associated with self-reporting of diarrhea and the self-reporting of scrapes and bruises.

Example: negative control exposure

- In a study of corticosteroids and self-reported asthma symptoms, repeat the analysis among a very low dose of the drug (so low that no effect on asthma is possible).
- An observed effect among patients who took a very low dose would suggest that there was bias in the measurement of asthma symptoms.



Lipsitch et al., 2010

Here's an example of a negative control exposure. A study *looks at the effect of corticosteroids on *self-reported asthma symptoms. And the purpose of the drug is to reduce the asthma symptoms. The corticosteroids can be delivered at different dosages. *So the study also looked at the effect of the exposure with a very low dose of the drug, *so low that no effect on asthma should be possible. So if patients report reduced asthma symptoms after taking this very low dose of the drug, it would suggest that there was some error in the measurement of these asthma symptoms or *potentially some recall bias because no effect would be suspected at that low dose of the drug.

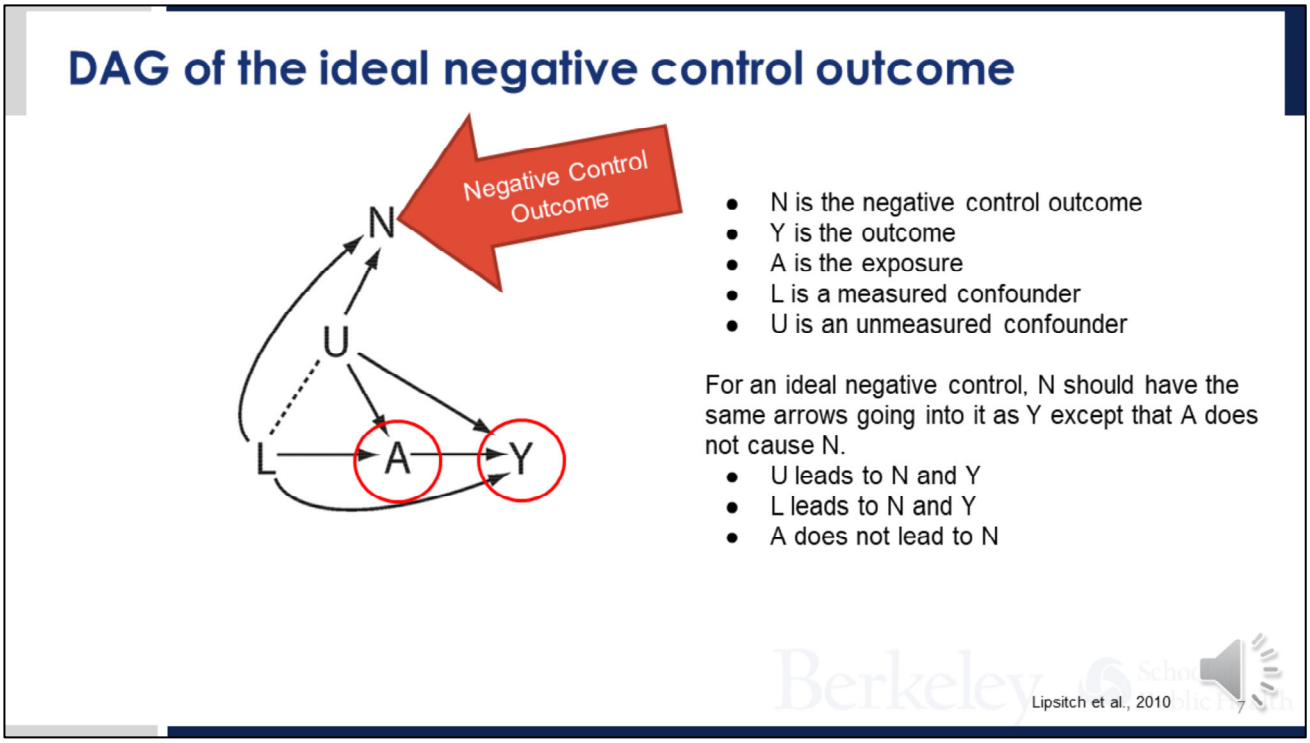
Example: negative control time period

- In a cohort study of the effect of influenza vaccination on mortality, measure the effect in the exposed vs. unexposed group in the period after immunization but before the start of influenza season, when no influenza is circulating.
- An observed difference in mortality between the exposed and unexposed in the period when influenza was not circulating locally would suggest other systematic differences between the exposed and unexposed explained differences in mortality, not influenza vaccination.

Lipsitch et al., 2010

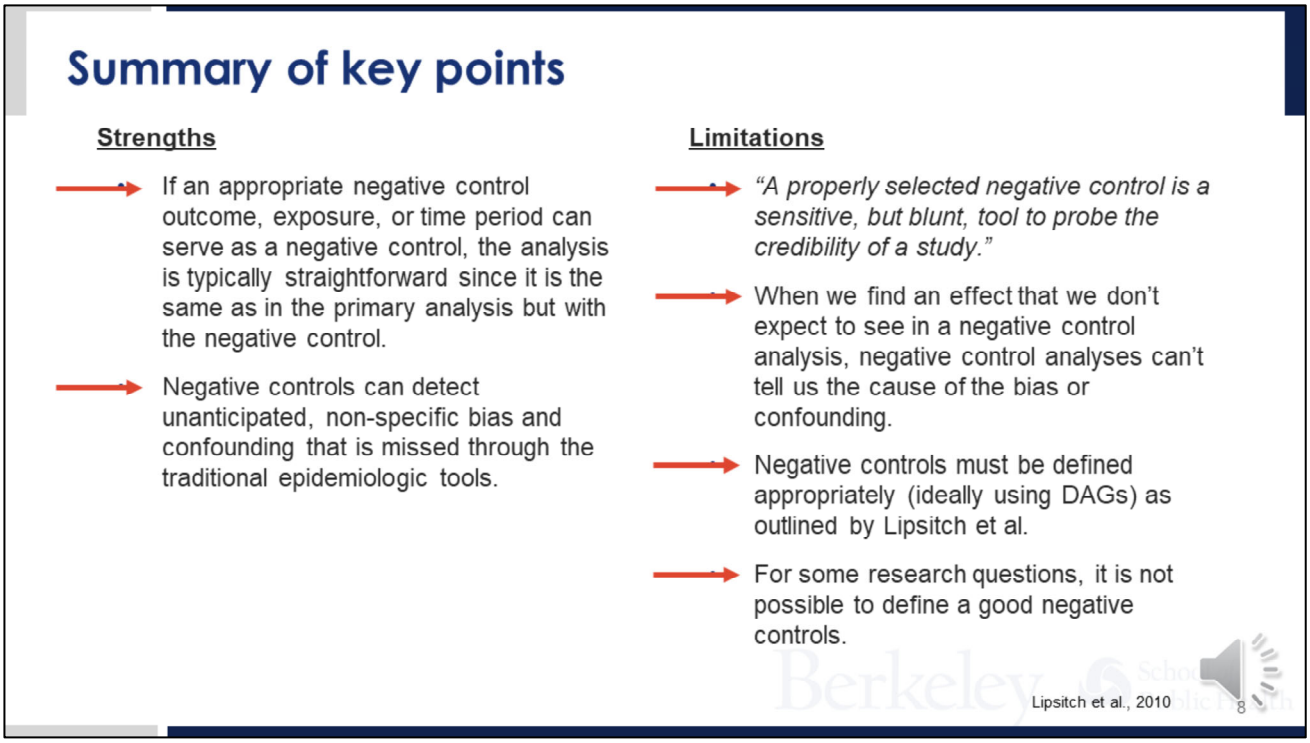
And here's an example of a negative control time period. *So we could conduct a cohort study of the effect of influenza vaccination on mortality. And we could compare the effect in the exposed versus unexposed group in the period *after immunization but *before the start of the flu season. So influenza vaccines are altered year to year, depending on the circulating strains of influenza, which vary from season to season.

In the period before the season starts, *we wouldn't expect being vaccinated for influenza to reduce mortality because there just isn't much influenza at that particular time. So if we did see a reduction in mortality before the start of flu season, it may suggest that there was some bias in our study design.



We can use DAGs to assess whether a potential negative control outcome or exposure of interest fits the ideal for a negative control. So, in this DAG, *A is our exposure, *Y is our outcome, U is an unmeasured confounder, L is a measured confounder, and *N is our negative control outcome. And this is the ideal structure for a negative control outcome. And this is because N should have the same arrows going into it as Y, except that A does not cause N. So we have an arrow from U to Y. We also have an arrow from U to N. We have an arrow from L to N. And we have an arrow from L to Y. Everything is the same upstream of Y as it is upstream of N, except that A does not cause N.

So if you're interested in one of your studies in using a negative control outcome, draw a DAG and see if it has this particular structure for the negative control outcome of interest. And you can look at the Lipsitch paper for an example of a DAG for a negative control exposure. It's a little more complicated, so I'm not going to go into it here. But we can use DAGs for all kinds of things, and so I just want to underscore that point. And if this is something of more interest, I recommend that you read the full paper.



To summarize, the strengths of negative control analyzes are that if you *define an appropriate negative control, whether that's a negative control exposure, outcome, or time period, the analysis is often pretty straightforward. You're basically doing the same analysis you would otherwise do. You're just switching out of variables. You're switching the real exposure for a negative control exposure, the real outcome for the negative control outcome. You do have to collect data on an additional variable, and so this is something that you have to think of in advance of data collection.

*And a major strength is that these negative controls can detect unanticipated, nonspecific sources of bias and confounding that would otherwise be missed through our traditional epidemiologic tools.

*But a limitation, and this is a nice quote from this Lipsitch paper, "A limitation is that a properly selected negative control is a sensitive but blunt tool to probe the credibility of a study." *So another way of saying this is that when we find an effect in a negative control analysis, it doesn't tell us the cause of the bias or confounding. So it may be quite frustrating because you see this effect that you don't really want to see or didn't expect to see, but you're not sure what to do about it because you don't know what the cause of the problem is.

*This is under the limitation column, but it's not totally a limitation. It's just to say that we need to define negative controls appropriately. So if you use an improper

negative control, one that doesn't adhere to the sort of ideal DAG structure, then you may get a non-null finding, but that doesn't mean that there's necessarily a bias or confounding. In other words, it's really important to do a good job defining your negative control in order to be able to trust the result.

*And then finally, for some research questions, it just really isn't possible to define a good negative control. And that's because the exposure may just cause all kinds of different outcomes that are not of interest, but, as a result, it's difficult to come up with a negative control outcome, for example.

Despite these limitations, we do think that this is a really useful tool and strongly encourage you to think about it in your future research.

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,

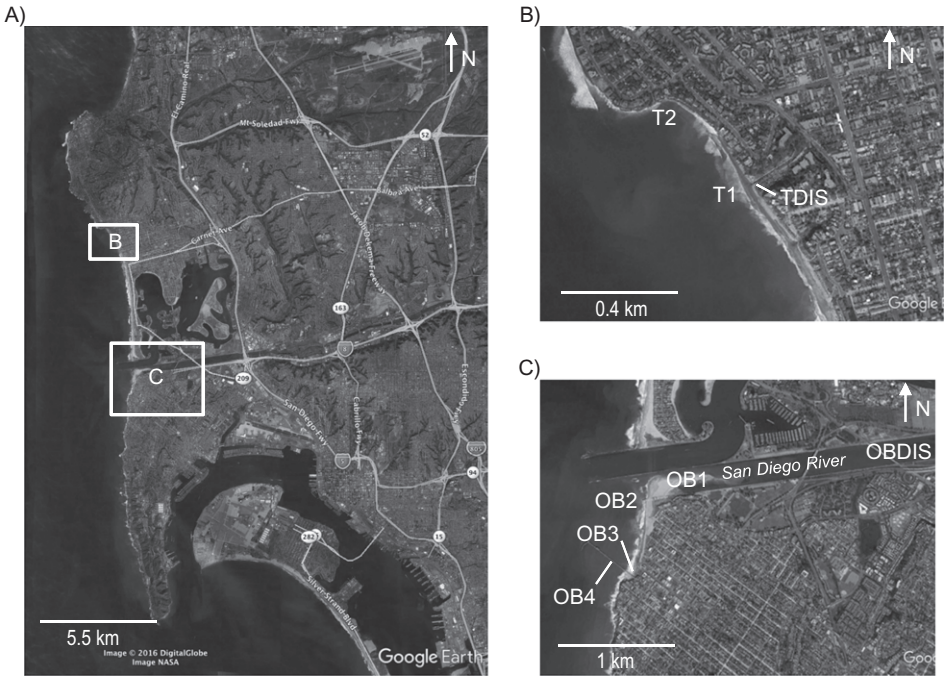


Figure 1. Monitoring beach water quality sampling locations in San Diego, California, winters of 2013–2014 and 2014–2015. Shown are the locations of the 2 monitored beaches along the San Diego coastline (A) and the water quality sampling sites at Tourmaline Surfing Park (B) and Ocean Beach (C). Samples 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).

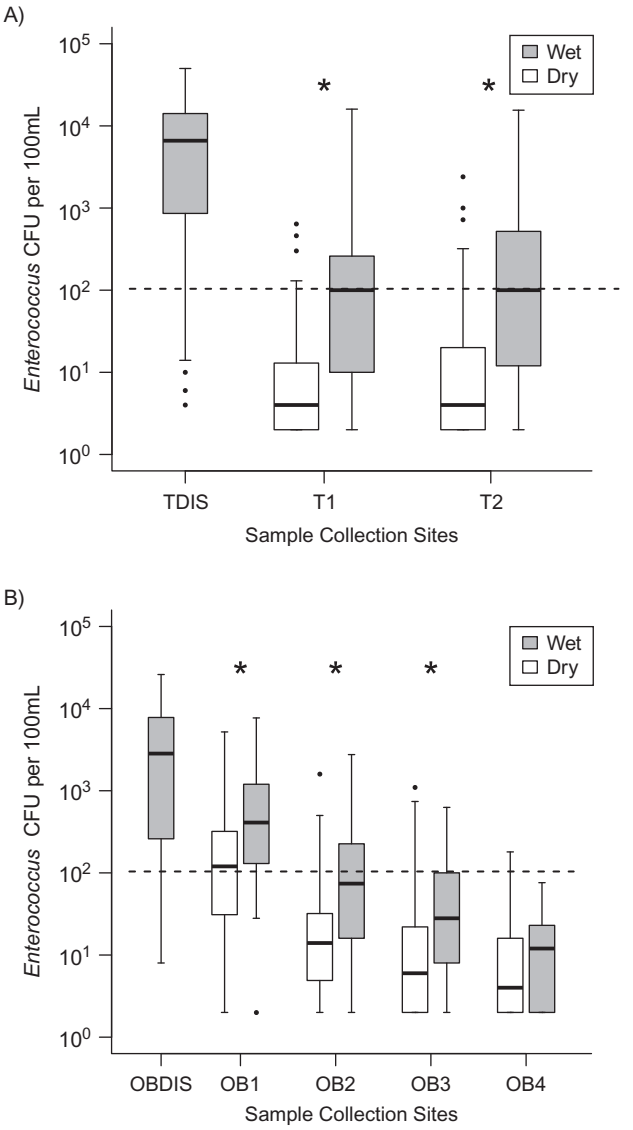


Figure 2. *Enterococcus* levels during dry and wet weather at the sampling locations at Tourmaline Surfing Park (A) and Ocean Beach (B) mapped in Figure 1. Boxes mark interquartile ranges, vertical lines mark 1.5 times the interquartile range, and points mark outliers. Horizontal dashed lines mark the 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.

Acute illness among surfers following dry and wet weather seawater exposure

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, John M. Colford, Jr.

American Journal of Epidemiology

Web Material

Web Table 1. STROBE Checklist

Web Appendix 1. Statistical details

Web Appendix 2. Power and sample size calculations

Web Appendix 3. Conversion of incidence rates into cumulative incidence

Web Appendix 4. Sensitivity analyses

Web Appendix 5. Negative control exposure analyses

Web Appendix 6. Comparison of gastrointestinal illness rates with summer cohorts

Web Appendix References

Additional Web Tables and Figures

Web Fig 1, Web Fig 2, Web Table 2, Web Fig 3, Web Fig 4, Web Table 3, Web Fig 5, Web Table 4, Web Fig 6, Web Table 5, Web Fig 7

Web Table 1. STROBE Checklist

Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)

Item	DESCRIPTION	REPORTED IN SECTION
Title and Abstract		
1a	Indicate the study's design with a commonly used term in the title or the abstract	Abstract (longitudinal cohort study)
1b	Provide in the abstract an informative and balanced summary of what was done and what was found	Abstract
Introduction		
Background/rationale		
2	Explain the scientific background and rationale for the investigation being reported	Introduction
Objectives		
3	State specific objectives, including any prespecified hypotheses	Introduction
Methods		
Study Design		
4	Present key elements of study design early in the paper	Methods (Study Design and Enrollment)
Setting		
5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	Methods (Setting, Study Design and Enrollment)
Participants		
6a	Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	Methods (Study Design and Enrollment)
6b	For matched studies, give matching criteria and number of exposed and unexposed.	Not applicable
Variables		
7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	Methods (Outcome Definition and Measurement, Exposure Definition and Measurement, Statistical Analysis)
Data Sources and Management		
8	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	Methods (Outcome Definition and Measurement, Exposure Definition and Measurement)
Bias		

9	Describe any efforts to address potential sources of bias	Methods (Statistical Analysis)
Study Size		
10	Explain how the study size was arrived at	Web Appendix 2 (Power and sample size calculations)
Quantitative Variables		
11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	Methods (Statistical Analysis)
Statistical Methods		
12a	Describe all statistical methods, including those used to control for confounding	Methods (Statistical Analysis) Web Appendix 1 (Detailed stat. notation)
12b	Describe any methods used to examine subgroups and interactions	Methods (Statistical Analysis) Web Appendix 1 (Detailed stat. notation)
12c	Explain how missing data were addressed	Methods (Statistical Analysis)
12d	If applicable, explain how loss to follow-up was addressed	Methods (Statistical Analysis)
12e	Describe any sensitivity analyses	Methods (Statistical Analysis) Web Appendix 4 (Sensitivity analyses)
Results		
Participants		
13a	Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	Results (Study Population)
13b	Give reasons for non-participation at each stage	Refusal rates were not measurable
13c	Consider use of a flow diagram	Not applicable
Descriptive data		
14a	Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	Results (Study Population, Water Quality and Surfer Exposure) Table 1 Figs 1-2 Web Figs 1-2
14b	Indicate number of participants with missing data for each variable of interest	Table 1
14c	Summarise follow-up time (eg, average and total amount)	Results (Study Population, Water Quality and Surfer Exposure)
Outcome Data		

15	Report numbers of outcome events or summary measures over time	Table 2 Table 4 Web Table 2
Main Results		
16a	Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	Table 3 Table 4
16b	Report category boundaries when continuous variables were categorized	Not applicable
16c	If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	Web Fig 3 Web Fig 6
Other Analyses		
17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	Web Figs 3 - 7 Web Tables 2 - 5
Discussion		
Key Results		
18	Summarise key results with reference to study objectives	Discussion (Key Results)
Limitations		
19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	Discussion (Limitations)
Interpretation		
20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	Discussion (Interpretation)
Generalisability		
21	Discuss the generalisability (external validity) of the study results	Discussion (Interpretation)
Other Information		
Funding		
22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	Acknowledgements

Web Appendix 1. Statistical details

Analyses were pre-specified through the Open Science Framework (see repository for pre-registered analysis plan along with full replication files: https://osf.io/hvn7s). We defined incident episodes as the onset of symptoms preceded by ≥6 symptom-free days to increase the likelihood that separate episodes represented distinct infections (1,2). We treated an individual's first 6 days of follow-up time as "at risk" under the assumption that the individual did not have incident illness in the days immediately before the start of their recorded symptom history. 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 ocean exposure and subsequent illness using an incidence rate ratio (IRR). Let Y_{it} be a binary indicator equal to 1 if individual i is ill on day t (0 otherwise), let T_{it} be an indicator that participant i is at risk of illness on day t. Let E_{it} be a binary indicator of equal to 1 if individual i entered the ocean on day t (0 otherwise). Define $E^*_{it} = \max(E_{i,t-1}, \ldots, E_{i,t-3})$, which is a binary indicator of whether the individual entered the ocean in the three days prior to the outcome measurement on day t. Our first parameter of interest was the IRR associated with ocean exposure in the past three days ($E^* = 1$), averaged over potentially confounding covariates (X). We modeled illness for individual i on day t using the following log-linear rate model (3), subset to days at risk ($T_{it} = 1$):

$$\log \mathrm{E}[Y_{it} \mid E^*_{it}, X_{it}] = \alpha + \beta E^*_{it} + \gamma X_{it} \qquad (1)$$

where X_{it} is a vector of potential confounders included in adjusted analyses (details below). We estimated the IRR associated with ocean exposure from the model, exp(β), and used robust standard errors that accounted for repeated observations within individuals (4).

Our second research question examined whether ocean exposure increased illness rates more if exposure took place within three days of wet weather compared with exposure during dry weather. Let D_t be a count of days since it rained >0.25 cm in 24 hours, with $D_t = \{0, 1, 2, \ldots\}$. Let R_{it} be a binary indicator equal to 1 if individual i entered the ocean on day t and $D_t \leq 3$ (0 otherwise), indicating that a surf session took place within three days of rain. Define R^*_{it} = max($R_{i,t-1}$, $R_{i,t-2}$, $R_{i,t-3}$), a binary indicator of whether an individual had a wet weather exposure in the past three days. With E^*_{it} (an indicator of any ocean exposure in the past three days), we created a three level categorical exposure:

$$W_{it} = \begin{cases} E^*_{it} = 0, R^*_{it} = 0\text{: unexposed} \\ E^*_{it} = 1, R^*_{it} = 0\text{: dry} \\ E^*_{it} = 1, R^*_{it} = 1\text{: wet} \end{cases}$$

We estimated a log-linear model, subset to days at risk ($T_{it} = 1$):

$$\log \mathrm{E}[Y_{it} \mid W_{it}, X_{it}] = \alpha + \beta_1 I(W_{it}=dry) + \beta_2 I(W_{it}=wet) + \gamma X_{it} \qquad (2)$$

where X_{it} are covariates in adjusted models. We estimated separate IRRs from the model for surf exposure during dry versus unexposed periods, exp(β_1), for surf exposure during wet versus unexposed periods, exp(β_2), and for wet versus dry periods, exp($\beta_2 - \beta_1$). For each outcome, we calculated a test of trend in the IRRs for dry and wet weather exposures (not pre-specified), in which the test for log-linear trend in incidence rates was significant if the coefficient β_2 differed from zero (5).

We estimated the association between fecal indicator bacteria levels and illness using the subset of surf sessions matched to water quality indicator measurements at the sentinel beaches using $\log_{10}$ continuous indicator levels, F_{it} . For surfers with a single day of exposure matched to indicator levels in the past three days, F_{it} equaled the daily geometric mean value on the exposed day. For surfers with multiple exposures matched to indicator levels in the past three days, we calculated the mean concentration weighted by the number of hours spent in the water on each day. We modeled the relationship between indicator levels and illness for individual *i* on day *t* using a log-linear model, subset to days at risk ($T_{it} = 1$):

$$\log \mathrm{E}[Y_{it} \mid F_{it}, X_{it}] = \alpha + \delta F_{it} + \gamma \mathbf{X}_{it} \quad (3)$$

where $\exp(\delta)$ estimates the IRR associated with a 1-$\log_{10}$ increase in indicator level. We also estimated the IRR associated with values above versus below USEPA *Enterococcus* regulatory guidelines (6) by replacing F_{it} in equations 3 and 4 with an indicator equal to 1 if F_{it} exceeded 35 CFU/100ml or, in a second definition, if any single sample on the exposure day exceeded 104 CFU/100ml. In the water quality analysis, we also hypothesized that the relationship between fecal indicator bacteria and illness could be modified by whether it was dry or wet weather exposure. We allowed the exposure-response relationship to vary by exposure during dry and wet weather by including an indicator for wet weather exposure in the past three days, R^*_{it} , and an interaction term in the model:

$$\log \mathrm{E}[Y_{it} \mid F_{it}, W^*_{it}, X_{it}] = \alpha + \delta_1 F_{it} + \delta_2 R^*_{it} + \delta_3 F_{it} R^*_{it} + \gamma \mathbf{X}_{it} \quad (4)$$

Potential confounders: We selected potential confounders that could be either a cause of seawater exposure, a cause of illness, or both (7). We controlled for the following time-invariant

potential confounders measured at enrollment: age, sex, education, employment status, household income, years the individual had surfed, reported behavior of typically avoiding the ocean following wet weather, surfboard length, mode of enrollment (beach vs. web), and chronic health conditions included only for the corresponding outcomes: ear problems, sinus problems, gastrointestinal conditions, respiratory conditions, skin conditions. We also controlled for 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 three days. In the adjusted overall seawater exposure analysis and water quality analysis we considered an indicator of wet weather in the past three days; in the water quality analysis we also considered an indicator for sentinel beach and an indicator for whether the individual surfed at beaches other than our two sentinel beaches in the same three-day period as their sentinel beach exposure. From this set of potential confounders, we retained those that had a univariate association with the outcome, defined as a likelihood ratio test *P*-value <0.20 in an unadjusted model. For categorical variables, we included a "missing" category for missing values.

Web Appendix 2. Power and sample size calculations

The sample size for the study was developed in two stages because little was known about outcome or exposure prevalence in the surfer population. In year 1, we aimed to enroll 100-200 surfers and follow them for up to 12 weeks to collect exposure and illness information, as well as fecal indicator bacteria levels. Using exposure and outcome information from the initial, smaller cohort in year 1, we then calculated sample size and power for the full study and this informed enrollment targets for year 2.

During the first year of the study, we enrolled 162 individuals and observed 12 incident cases of gastrointestinal illness from 2,310 days at risk -- an incidence rate of 5 episodes per 1,000 person-days. We used a standard sample size equation for the comparison of two incidence rates (8): $y = (z_{\alpha/2} + z_{\beta})^2 (\lambda_0 + \lambda_1) / (\lambda_0 - \lambda_1)^2$, where y is the number of person-days required in each exposure category, $z_{\alpha/2}$ and z_{β} are standard normal distribution values corresponding to upper tail probability values $\alpha/2$ and β (we set α=0.05 and β=0.2), and λ_0 and λ_1 are incidence rates in the unexposed and exposed periods. Assuming a rate of 5 episodes per 1,000 person-days during unexposed periods (λ_0=0.005), the Table below summarizes the number of person-days of observation in each exposure group required to detect different magnitudes of effect, as measured by the incidence rate ratio (IRR).

In year 1 of the study, 55% of the days of observation were exposed because surfers entered the ocean frequently. However, only 13% of the days of observation were classified as wet weather exposure because it was a drought year.

Given this, we expected that a total of 2,408 / 0.13 = 18,520 person-days of observation would be sufficient to detect an IRR of 1.50 or greater in wet weather exposed versus unexposed periods.

IRR	Person-days of observation in each exposure group (y)	Total person-days required, assuming 13% of days are wet weather exposure (y / 0.13)
1.2	13,242	101,859
1.3	6,153	47,328
1.4	3,611	27,780
1.5	2,408	18,520

For associations between $\log_{10}$ *Enterococcus* and incident illness we used a simulation-based approach (9). The simulation resampled the empirical distribution of water quality measurements from year 1 for 200 surfers with different lengths of follow-up and calculated a predicted probability of incident gastrointestinal illness on each day using the rate in the unexposed periods from year 1 and an increased rate following exposure that corresponded to different effect sizes. For a given strength of association (IRR), we then increased the length of follow-up until >80% of the 1,000 simulations had a $P<0.05$ (equivalent to $\alpha=0.05$ and $\beta=0.2$). Simulations showed that 3,000 days of follow-up matched to water quality indicator measurements at sentinel beaches would provide >80% power to estimate an IRR of 1.75 or greater for a $\log_{10}$ increase in *Enterococcus* levels.

Web Appendix 3. Conversion of incidence rates into cumulative incidence

The longitudinal design with varying lengths of follow-up and varying exposure periods meant that the natural measure of illness was incidence rates (episodes / person-days) (3). However, federal water quality guidelines and quantitative microbial risk assessment models measure illness in units of cumulative incidence or "risk" (episodes / person) for gastrointestinal illness (6). We converted marginally adjusted incidence rate estimates from log-linear models described above into 3-day cumulative incidence using the density method (10). We compared cumulative incidence during dry and wet weather exposure periods to unexposed periods using the difference in cumulative incidence ("risk difference" [RD]), and estimated standard errors and 95% confidence intervals for the RD using the delta method (11). We used a 3-day cumulative incidence because incidence rates were measured over 3-day periods following exposure -- the high frequency of exposure made longer follow-up periods infeasible. In California swimmer cohorts, the majority of excess cases of gastrointestinal illness occurred in the 1-2 days following ocean exposure; for this reason, a 3-day RD should be a reasonable approximation of the RD calculated over a longer 10-12 day period, as measured in past swimmer cohort studies (12–14).

Web Appendix 4. Sensitivity analyses

The primary analysis defined wet weather exposure as periods within 0-3 days following rainfall, consistent with current beach posting guidelines in California that warn recreators to avoid water contact for 72 hours after rainfall. In a sensitivity analysis we changed the length of the wet weather window in daily increments from 0 to 5 days following rainfall to determine if shorter windows were associated with larger increase in illness rates.

The technical report associated with this study (https://osf.io/hvn7s) includes additional sensitivity analyses. First, we stratified incidence rates by storm size. We found evidence for higher incidence rates associated with ocean exposure following larger storms compared with smaller storms, but the analysis relied upon relatively small sample sizes with the higher level of wet weather stratification and therefore should be viewed as supportive but exploratory. Second, we examined the effect of excluding individuals from the analysis who submitted >1 survey per day (to reduce potential measurement bias) or who did not report the precise location of their ocean exposure (to reduce potential exposure misclassification). Excluding these subgroups of the population from the analysis did not change our inference.

Web Appendix 5. Negative control exposure analyses

We matched fecal indicator bacteria levels (*Enterococcus*, fecal coliforms, total coliforms) measured at one of the sentinel beaches to illness measurements by date, randomly assigning either the Ocean Beach value or the Tourmaline Surfing Park value to each matched observation. We then excluded from the dataset observations that were within five days of any seawater exposure. The negative control exposure analysis was designed to ensure that the person-time included could not plausibly be influenced by pathogens in seawater, but could be influenced by unmeasured sources of confounding -- such as seasonal epidemics of enteric pathogens like norovirus.

We estimated the association between fecal indicator bacteria levels and incident illness, allowing the association to vary by dry and wet weather exposure by including an interaction term between the indicator levels and a wet weather period indicator, as in the main analysis. In this negative control exposure analysis, there should be no plausible relationship between fecal indicator bacteria levels and subsequent illness unless there is bias from unobserved confounding or measurement error (15,16).

Web Appendix 6. Comparison of gastrointestinal illness rates with summer cohorts

To help contextualize the gastrointestinal illness rates measured in this study we completed a supplemental analysis to more directly compare illness rates with those measured in summer swimmer cohorts. The present study used a different design that past recreational swimmer cohorts conducted in California (12–14,17,18). Past swimmer cohorts enrolled beachgoers and then measured cumulative incident illness over 10-12 days after their single exposure. Such measurement was infeasible among surfers because of their frequent exposure (median of 2 times per week). For this reason, the present study estimated daily incidence rates during unexposed and exposed periods of follow-up -- the most natural measure of disease given the design. This difference in design and measure of illness complicates direct comparisons of illness between this study and past swimmer cohorts. A second difference in design that complicates direct comparison is that the present study limited enrollment to adults, whereas past swimmer cohorts included many children who have higher rates of gastrointestinal illness (18).

We had access to participant data from four California swimmer cohorts (18), and this enabled us to derive estimates of gastrointestinal illness rates from the past studies that were more comparable to those estimated in the surfer cohort. We subset the four California cohorts (Avalon, Doheny, Malibu, Mission Bay) to adults (18 years or older) and calculated incidence rates over the first 3 days of follow-up -- a period after exposure comparable to the present study. We calculated incidence rates separately for non-swimmers (individuals with no water contact) and swimmers with head immersion exposure. We compared these gastrointestinal incidence rates with rates among surfers during unexposed and exposed periods in the present study.

Web Appendix References

1. 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.

2. 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.

3. Rothman KJ, Sander Greenland, Lash TL. Modern Epidemiology. 3rd ed. Philadelphia: Lippincott Williams and Wilkins; 2008.

4. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. books.google.com; 1967:221–233.

5. Vittinghoff E, Glidden DV, Shiboski SC, et al. Regression methods in biostatistics: linear, logistic, survival, and repeated measures models. 2nd ed. Springer Science & Business Media; 2012.

6. USEPA. Recreational Water Quality Criteria. United States Environmental Protection Agency Office of Water; 2012.

7. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. *Biometrics*. 2011;67(4):1406–1413.

8. Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. *Int. J. Epidemiol.* 1999;28(2):319–326.

9. Arnold B, Hogan D, Colford J, et al. Simulation methods to estimate design power: an overview for applied research. *BMC Med. Res. Methodol.* 2011;11(1):94.

10. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research: Principles and Quantitative Methods. New York: John Wiley & Sons, Inc.; 1982.

11. Wasserman L. All of statistics: a concise course in statistical inference. Springer; 2004.

12. 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.

13. 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.

14. 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.

15. Lipsitch M, Tchetgen ET, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*. 2010;21(3):383–388.

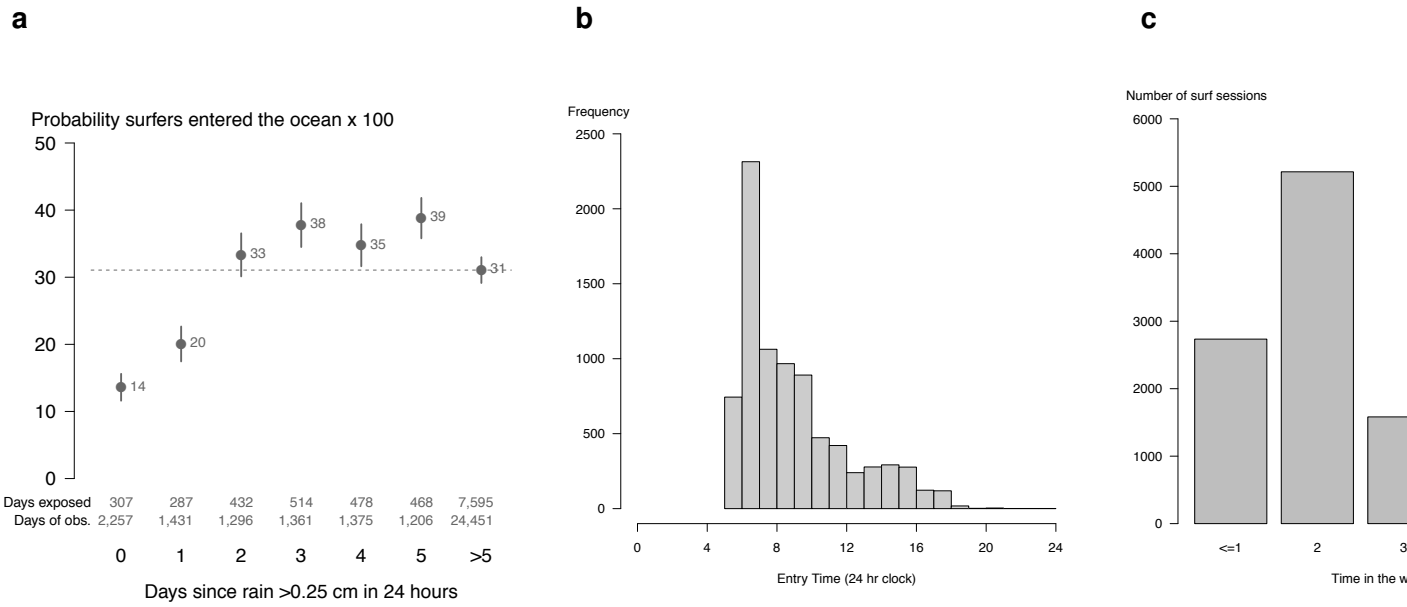
16. Arnold BF, Ercumen A, Benjamin-Chung J, et al. Negative controls to detect selection bias and

measurement bias in epidemiologic studies. *Epidemiology*. 2016;27(5):637–641.

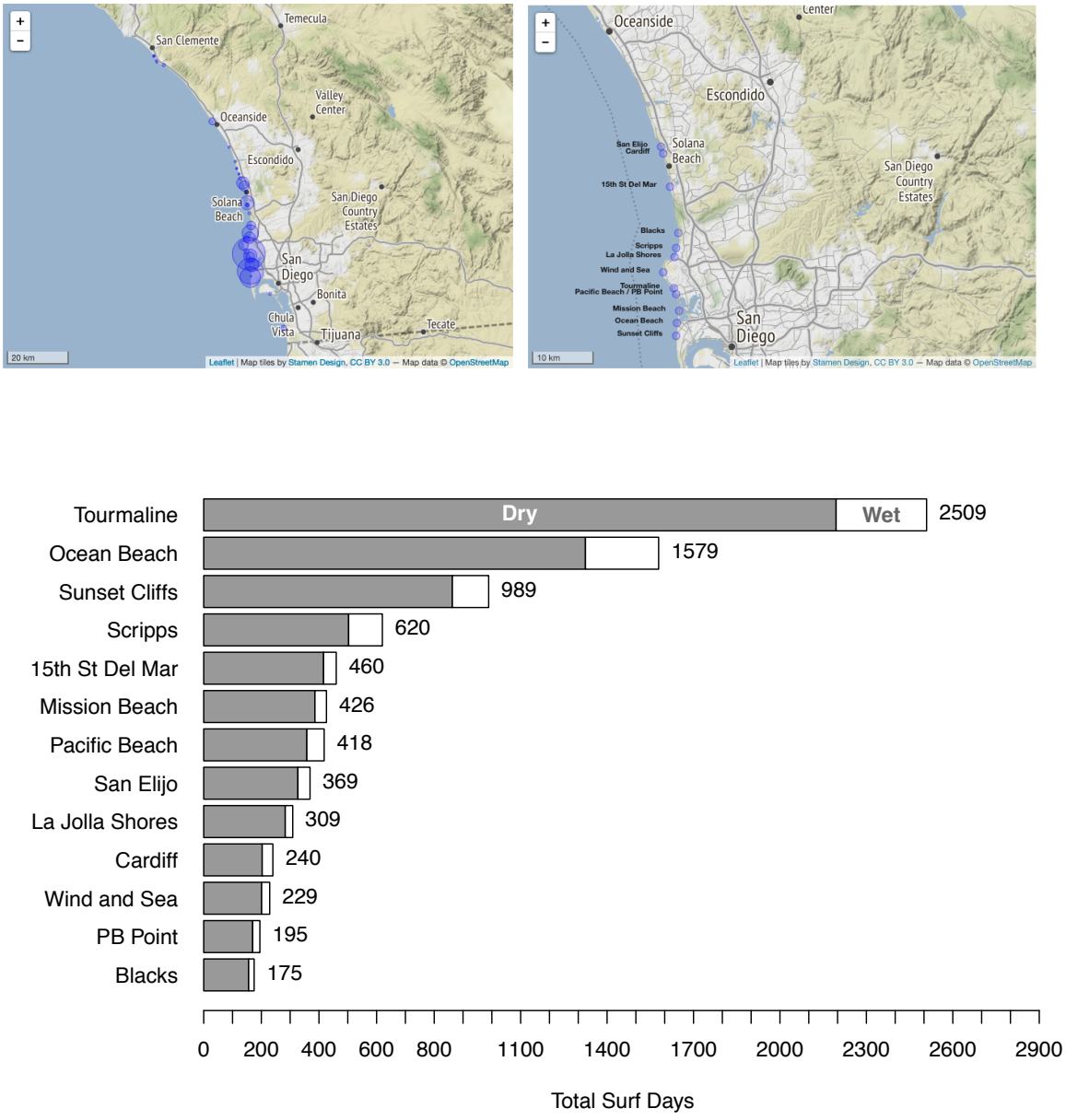
17. Colford JM, Wade TJ, Schiff KC, et al. Water quality indicators and the risk of illness at beaches with nonpoint sources of fecal contamination. *Epidemiology*. 2007;18(1):27–35.

18. 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.

Additional Web Tables and Figures



Web Figure 1: Surfer exposure during follow-up, summarized from 654 surfers (10,081 surf session Diego, CA region during the winters of 2013-14 and 2014-15. **a**, Probability that surfers entered the oc by days since precipitation >0.25 cm in 24 hours. Vertical lines indicate robust 95% confidence inte dashed line marks the probability for >5 days after rain. **b**, Distribution of ocean entry times. **c**, Distri spent in the ocean, rounded to hours.



Web Figure 2: Distribution of surf days for surf locations in San Deigo county (top left panel), where the area of each circle is scaled by the total surf days. The 13 most common surf locations (top right panel) represented 85% (8,518/10,081) of surf days observed in the study. The bottom panel shows the distribution of surf days at the 13 most popular locations. Wet weather was defined as >0.25 cm of rain in 24 hours. An interactive version of the maps is located in the study's open science framework repository: `https://osf.io/bfxq4`

Web Table 2: Incident illness and incidence rate ratios (IRR) associated with ocean exposure in the past three days among surfers in San Diego, CA (2013-14 and 2014-15 winters).

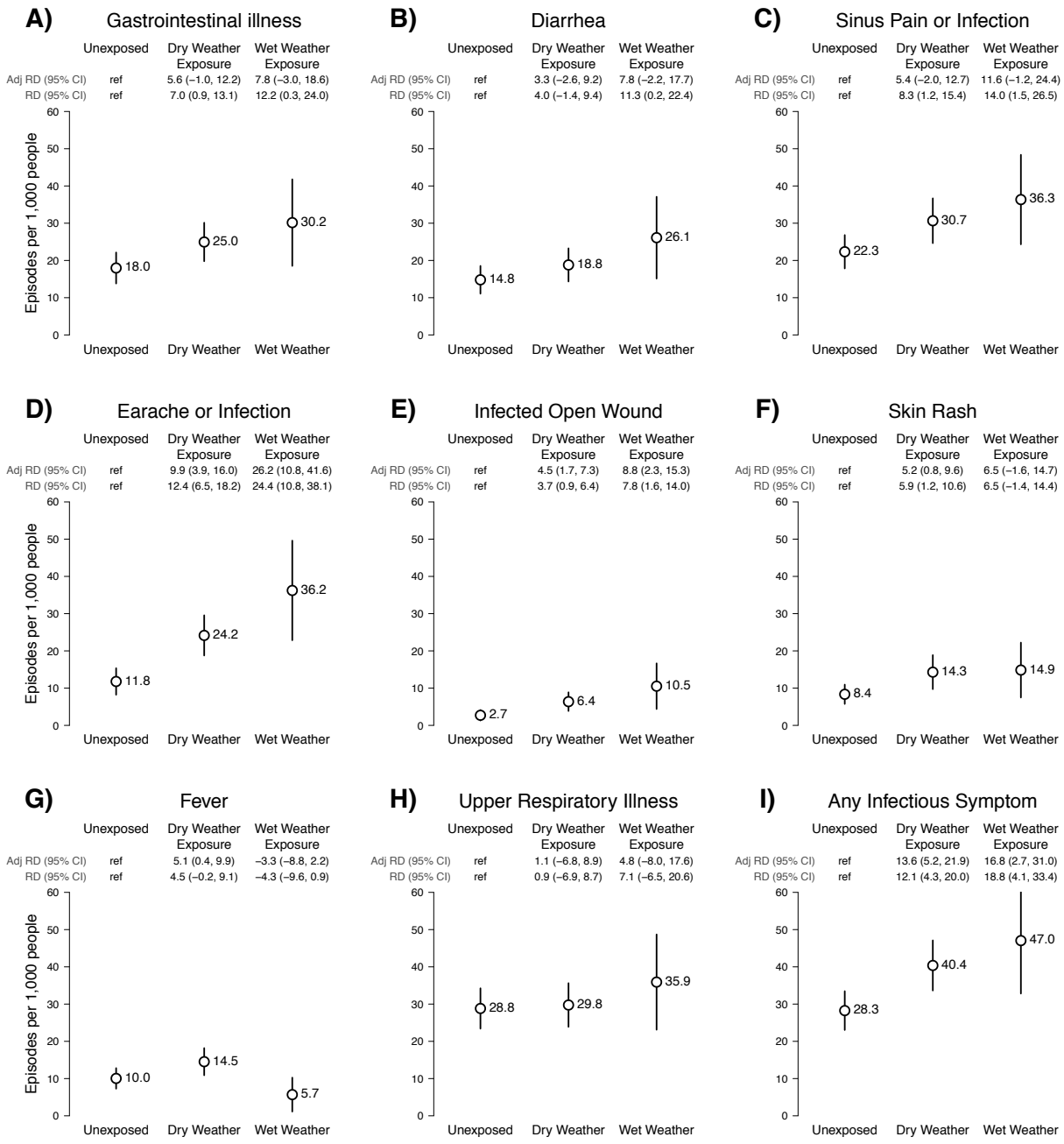
	Unexposed Periods		Ocean Exposure, Past 3 Days		Unadjusted
	Episodes/days at risk	Rate [a]	Episodes/days at risk	Rate [a]	IRR (95% CI)
Gastrointestinal illness	90/14,884	6.0	147/16,806	8.7	1.45 (1.10, 1.91)
Diarrhea	75/15,086	5.0	115/16,970	6.8	1.36 (1.00, 1.86)
Sinus pain or infection	109/14,475	7.5	176/16,389	10.7	1.43 (1.11, 1.83)
Earache or infection	59/14,931	4.0	148/16,626	8.9	2.25 (1.60, 3.16)
Infection of open wound	14/15,456	0.9	41/17,199	2.4	2.63 (1.45, 4.77)
Skin rash	42/15,024	2.8	81/16,757	4.8	1.73 (1.19, 2.51)
Fever	51/15,156	3.4	75/17,290	4.3	1.29 (0.89, 1.87)
Upper respiratory illness [c]	117/12,001	9.7	142/13,568	10.5	1.07 (0.83, 1.39)
Any infectious symptom [d]	138/14,445	9.6	228/16,102	14.2	1.48 (1.19, 1.85)

a Episodes per 1,000 person-days

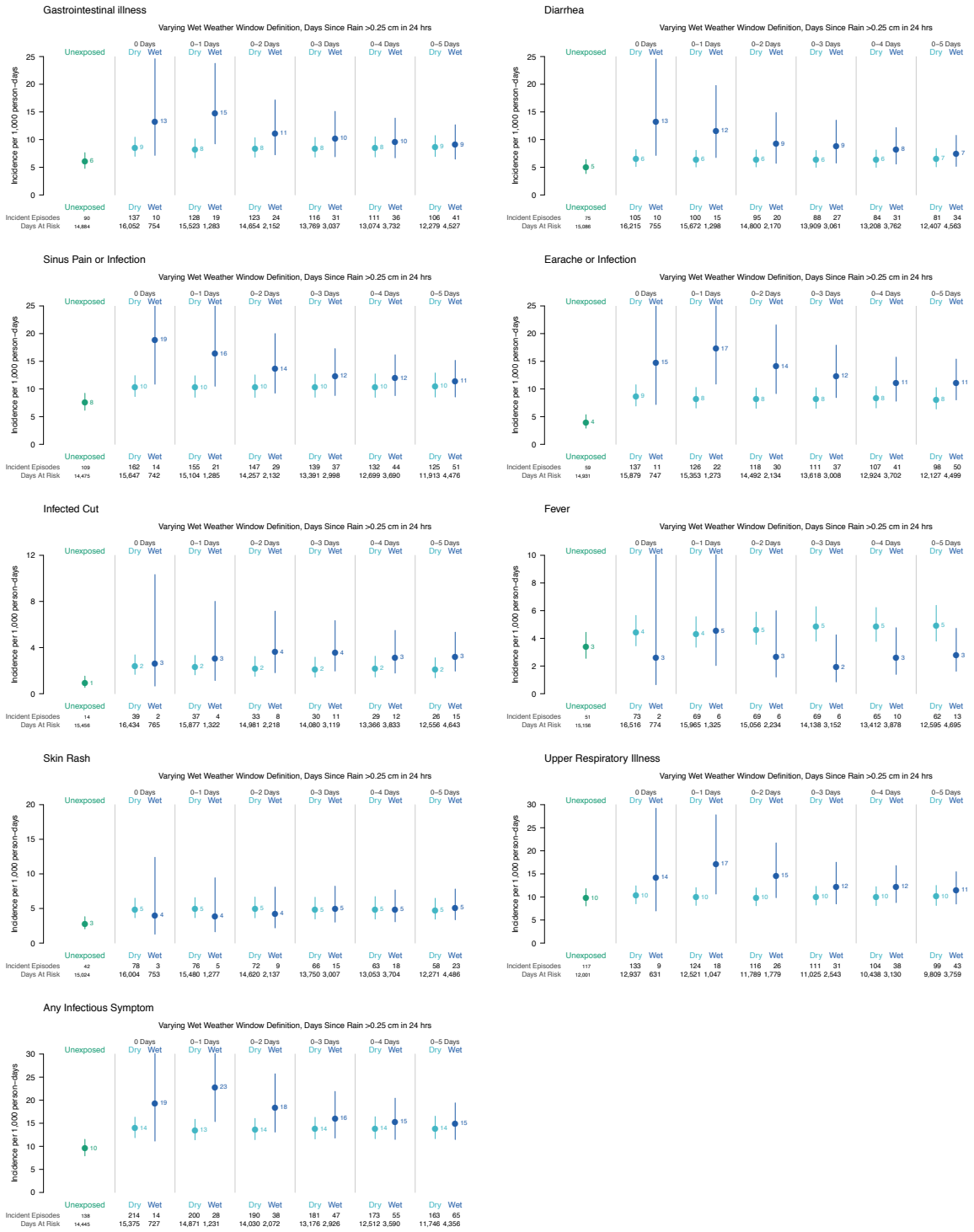
b Adjusted for a range of covariates (see statsistical methods for details)

c Only measured in year 2

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



Web Figure 3: Three-day cumulative incidence of illness among surfers associated with dry and wet weather exposure in San Diego, CA during the winters of 2013-14 and 2014-15. Unadjusted and adjusted risk differences (RD) compare cumulative incidence in the three days following ocean exposure during dry or wet weather with three-day cumulative incidence during unexposed periods. Wet weather was defined as >0.25 cm of rain in 24 hours.



Web Figure 4: Sensitivity analysis of wet weather exposure period definition on illness incidence rates among surfers in San Diego, CA (2013-14 and 2014-15 winters). Wet weather was defined as >0.25 cm of rain in 24 hours. Incidence rates for dry and wet weather were re-calculated for varying lengths of wet weather window. The primary analysis used a period of 0-3 days.

Web Table 3: Incidence rate ratios (IRR) associated with fecal indicator bacteria among surfers exposed at Tourmaline Surfing Park and Ocean Beach in San Diego, CA (2013-14 and 2014-15 winters). The IRR is associated with either a log10 increase in indicator bacteria (*Enterococcus* , fecal coliforms, total coliforms), or exposure above versus below *Enterococcus* regulatory guidelines of 35 or 104 colony forming units (CFU) per 100/ml

		Episodes/ days at risk	Unadjusted IRR (95% CI)	Adjusted [e] IRR (95% CI)
Enterococcus	Gastrointestinal illness	40 / 5548	1.23 (0.74 , 2.03)	1.04 (0.63 , 1.72)
log_{10}	Diarrhea	33 / 5590	1.54 (0.94 , 2.51)	1.36 (0.83 , 2.21)
	Sinus pain or infection	63 / 5392	1.61 (1.13 , 2.28)	1.27 (0.85 , 1.90)
	Earache or infection	52 / 5507	0.94 (0.54 , 1.62)	0.91 (0.53 , 1.56)
	Infection of open wound	15 / 5692	2.67 (1.39 , 5.13)	3.24 (1.66 , 6.31)
	Skin rash	24 / 5497	1.23 (0.65 , 2.31)	0.88 (0.39 , 1.98)
	Fever	24 / 5708	1.28 (0.74 , 2.22)	1.29 (0.74 , 2.24)
	Upper respiratory illness [c]	52 / 4769	1.28 (0.80 , 2.03)	1.12 (0.71 , 1.78)
	Any infectious symptom [d]	67 / 5344	1.51 (1.04 , 2.19)	1.42 (0.95 , 2.11)
Enterococcus	Gastrointestinal illness	40 / 5548	1.51 (0.76 , 3.02)	1.20 (0.59 , 2.44)
> 35 CFU [a]	Diarrhea	33 / 5590	2.08 (1.01 , 4.32)	1.77 (0.86 , 3.65)
	Sinus pain or infection	63 / 5392	1.80 (1.05 , 3.10)	1.35 (0.72 , 2.55)
	Earache or infection	52 / 5507	1.32 (0.71 , 2.47)	1.31 (0.72 , 2.37)
	Infection of open wound	15 / 5692	1.46 (0.52 , 4.09)	1.81 (0.64 , 5.15)
	Skin rash	24 / 5497	1.76 (0.74 , 4.16)	1.22 (0.43 , 3.47)
	Fever	24 / 5708	1.44 (0.58 , 3.55)	1.49 (0.62 , 3.62)
	Upper respiratory illness [c]	52 / 4769	0.97 (0.52 , 1.80)	0.80 (0.43 , 1.47)
	Any infectious symptom [d]	67 / 5344	1.47 (0.89 , 2.45)	1.35 (0.76 , 2.39)
Enterococcus	Gastrointestinal illness	40 / 5548	1.27 (0.66 , 2.44)	1.05 (0.53 , 2.06)
> 104 CFU [b]	Diarrhea	33 / 5590	1.75 (0.87 , 3.49)	1.53 (0.75 , 3.13)
	Sinus pain or infection	63 / 5392	1.85 (1.09 , 3.14)	1.39 (0.75 , 2.57)
	Earache or infection	52 / 5507	1.28 (0.72 , 2.28)	1.25 (0.71 , 2.20)
	Infection of open wound	15 / 5692	2.76 (1.03 , 7.44)	3.28 (1.16 , 9.23)
	Skin rash	24 / 5497	1.21 (0.51 , 2.85)	0.74 (0.27 , 2.07)
	Fever	24 / 5708	1.42 (0.58 , 3.49)	1.45 (0.59 , 3.55)
	Upper respiratory illness [c]	52 / 4769	1.09 (0.60 , 1.98)	0.89 (0.50 , 1.59)
	Any infectious symptom [d]	67 / 5344	1.79 (1.10 , 2.92)	1.69 (1.01 , 2.84)
Fecal coliform	Gastrointestinal illness	40 / 5548	1.40 (0.79 , 2.46)	1.14 (0.64 , 2.05)
log_{10}	Diarrhea	33 / 5590	1.76 (1.00 , 3.07)	1.56 (0.89 , 2.73)
	Sinus pain or infection	63 / 5392	1.85 (1.22 , 2.81)	1.31 (0.83 , 2.08)
	Earache or infection	52 / 5507	1.09 (0.61 , 1.95)	1.02 (0.57 , 1.84)
	Infection of open wound	15 / 5692	2.99 (1.47 , 6.07)	4.16 (1.84 , 9.41)
	Skin rash	24 / 5497	1.39 (0.69 , 2.81)	0.91 (0.36 , 2.28)
	Fever	24 / 5708	1.21 (0.57 , 2.60)	1.27 (0.61 , 2.65)
	Upper respiratory illness [c]	52 / 4769	1.46 (0.79 , 2.70)	1.18 (0.66 , 2.13)
	Any infectious symptom [d]	67 / 5344	1.73 (1.15 , 2.60)	1.66 (1.07 , 2.57)
Total coliform	Gastrointestinal illness	40 / 5548	1.17 (0.68 , 2.00)	1.03 (0.63 , 1.66)
log_{10}	Diarrhea	33 / 5590	1.15 (0.61 , 2.17)	1.04 (0.58 , 1.88)
	Sinus pain or infection	63 / 5392	1.73 (1.17 , 2.57)	1.50 (0.95 , 2.35)
	Earache or infection	52 / 5507	1.29 (0.75 , 2.21)	1.26 (0.72 , 2.21)
	Infection of open wound	15 / 5692	2.78 (1.35 , 5.73)	3.46 (1.78 , 6.75)
	Skin rash	24 / 5497	1.29 (0.72 , 2.29)	1.07 (0.51 , 2.27)
	Fever	24 / 5708	1.38 (0.75 , 2.52)	1.47 (0.81 , 2.67)
	Upper respiratory illness [c]	52 / 4769	1.31 (0.80 , 2.16)	1.20 (0.74 , 1.93)
	Any infectious symptom [d]	67 / 5344	1.70 (1.14 , 2.52)	1.65 (1.10 , 2.49)

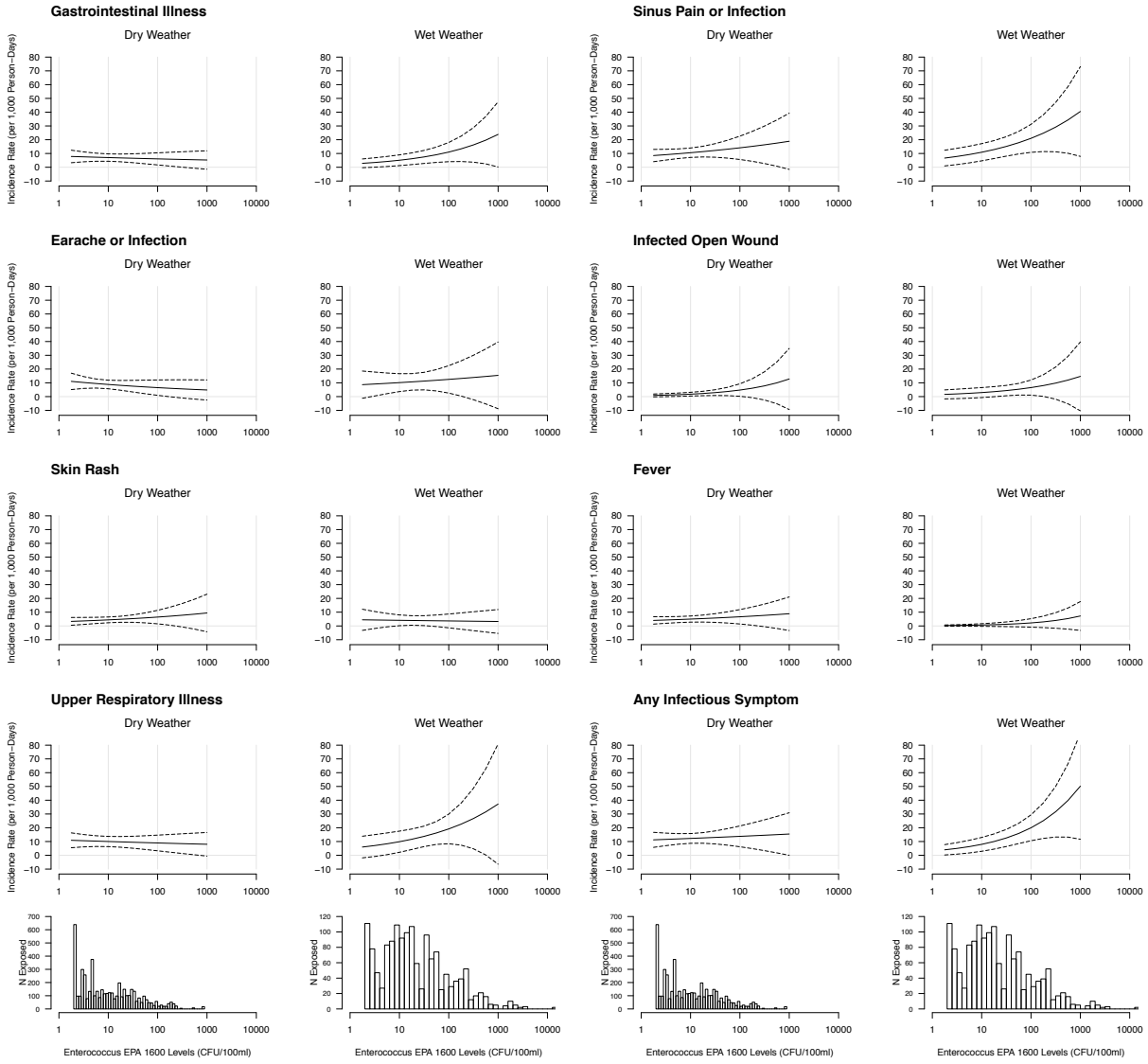
[a] Daily geomean > 35 CFU at any point during three-day follow-up

[b] Any single sample > 104 CFU at any point during three-day follow-up

[c] Only measured in year 2

[d] Includes gastrointestinal illness, diarrhea, vomiting, eye infections, infected cuts and fever

[e] Adjusted for a range of covariates (see statistical methods for details)



Web Figure 5: Incidence rates associated with *Enterococcus* levels measured during dry and wet weather periods, predicted from a log-linear model among surfers at Tourmaline Surfing Park and Ocean Beach, San Diego, CA (2013-14 and 2014-15 winters). Wet weather was defined as >0.25 cm of rain in 24 hours. Dashed lines indicate model-based 95% confidence intervals and histograms show the distribution of *Enterococcus* exposure in the population during dry and wet weather periods.

Web Table 4: Incidence rate ratios (IRR) associated with exposure above versus below *Enterococcus* regulatory guideline values, stratified by exposure during dry and wet weather, among surfers exposed at Tourmaline Surfing Park and Ocean Beach in San Diego, CA (2013-14 and 2014-15 winters).

		Dry Weather			Wet Weather				
		Episodes/ days at risk	Unadjusted IRR (95% CI)	Adjusted [e] IRR (95% CI)	Episodes/ days at risk	Unadjusted IRR (95% CI)	p [f]	Adjusted [e] IRR (95% CI)	p [f]
***Enterococcus* > 35 CFU** [a]	Gastrointestinal illness	30 / 4251	1.17 (0.49 , 2.80)	1.11 (0.46 , 2.69)	10 / 1297	2.90 (0.75 , 11.16)	0.29	1.81 (0.46 , 7.09)	0.56
	Diarrhea	24 / 4285	1.59 (0.64 , 3.96)	1.61 (0.64 , 4.04)	9 / 1305	4.39 (0.92 , 20.98)	0.30	2.97 (0.67 , 13.21)	0.52
	Sinus pain or infection	44 / 4130	1.52 (0.70 , 3.31)	0.99 (0.38 , 2.54)	19 / 1262	2.16 (0.85 , 5.48)	0.57	1.87 (0.73 , 4.81)	0.36
	Earache or infection	38 / 4233	1.28 (0.59 , 2.80)	1.26 (0.57 , 2.77)	14 / 1274	1.26 (0.42 , 3.79)	0.98	1.30 (0.44 , 3.81)	0.96
	Infection of open wound	9 / 4360	1.14 (0.22 , 5.87)	1.26 (0.23 , 6.91)	6 / 1332	1.28 (0.26 , 6.26)	0.92	1.90 (0.39 , 9.23)	0.74
	Skin rash	19 / 4230	2.38 (0.90 , 6.34)	1.77 (0.54 , 5.85)	5 / 1267	0.82 (0.14 , 4.89)	0.30	0.38 (0.04 , 3.76)	0.22
	Fever	22 / 4366	1.48 (0.55 , 4.01)	1.50 (0.57 , 3.94)	2 / 1342	-- [g]		-- [g]	
	Upper respiratory illness [c]	37 / 3679	0.64 (0.26 , 1.60)	0.48 (0.19 , 1.19)	15 / 1090	1.33 (0.48 , 3.72)	0.31	1.37 (0.52 , 3.60)	0.13
	Any infectious symptom [d]	50 / 4080	1.08 (0.54 , 2.17)	1.01 (0.47 , 2.15)	17 / 1264	3.08 (1.10 , 8.56)	0.11	2.76 (0.94 , 8.16)	0.14
***Enterococcus* > 104 CFU** [b]	Gastrointestinal illness	30 / 4251	0.93 (0.38 , 2.25)	0.92 (0.38 , 2.22)	10 / 1297	2.69 (0.69 , 10.39)	0.23	1.78 (0.44 , 7.26)	0.46
	Diarrhea	24 / 4285	1.26 (0.50 , 3.18)	1.32 (0.52 , 3.34)	9 / 1305	4.06 (0.85 , 19.50)	0.25	2.90 (0.59 , 14.21)	0.44
	Sinus pain or infection	44 / 4130	1.85 (0.94 , 3.63)	1.29 (0.58 , 2.85)	19 / 1262	1.63 (0.65 , 4.10)	0.83	1.33 (0.50 , 3.53)	0.96
	Earache or infection	38 / 4233	1.29 (0.68 , 2.47)	1.25 (0.63 , 2.49)	14 / 1274	1.14 (0.38 , 3.43)	0.84	1.11 (0.40 , 3.11)	0.85
	Infection of open wound	9 / 4360	2.53 (0.58 , 11.02)	2.72 (0.60 , 12.28)	6 / 1332	2.32 (0.43 , 12.47)	0.94	3.00 (0.54 , 16.65)	0.93
	Skin rash	19 / 4230	1.47 (0.56 , 3.91)	1.03 (0.33 , 3.16)	5 / 1267	0.74 (0.13 , 4.37)	0.51	0.27 (0.03 , 2.71)	0.28
	Fever	22 / 4366	1.44 (0.55 , 3.79)	1.45 (0.55 , 3.83)	2 / 1342	-- [g]		-- [g]	
	Upper respiratory illness [c]	37 / 3679	0.92 (0.43 , 1.96)	0.71 (0.34 , 1.51)	15 / 1090	1.22 (0.44 , 3.39)	0.66	1.08 (0.40 , 2.90)	0.51
	Any infectious symptom [d]	50 / 4080	1.56 (0.83 , 2.94)	1.52 (0.81 , 2.86)	17 / 1264	2.79 (0.99 , 7.82)	0.44	2.51 (0.84 , 7.47)	0.45

CFU: colony forming units

[a] Daily geometric mean > 35 CFU at any point during three-day follow-up

[b] Any single sample > 104 CFU at any point during three-day follow-up

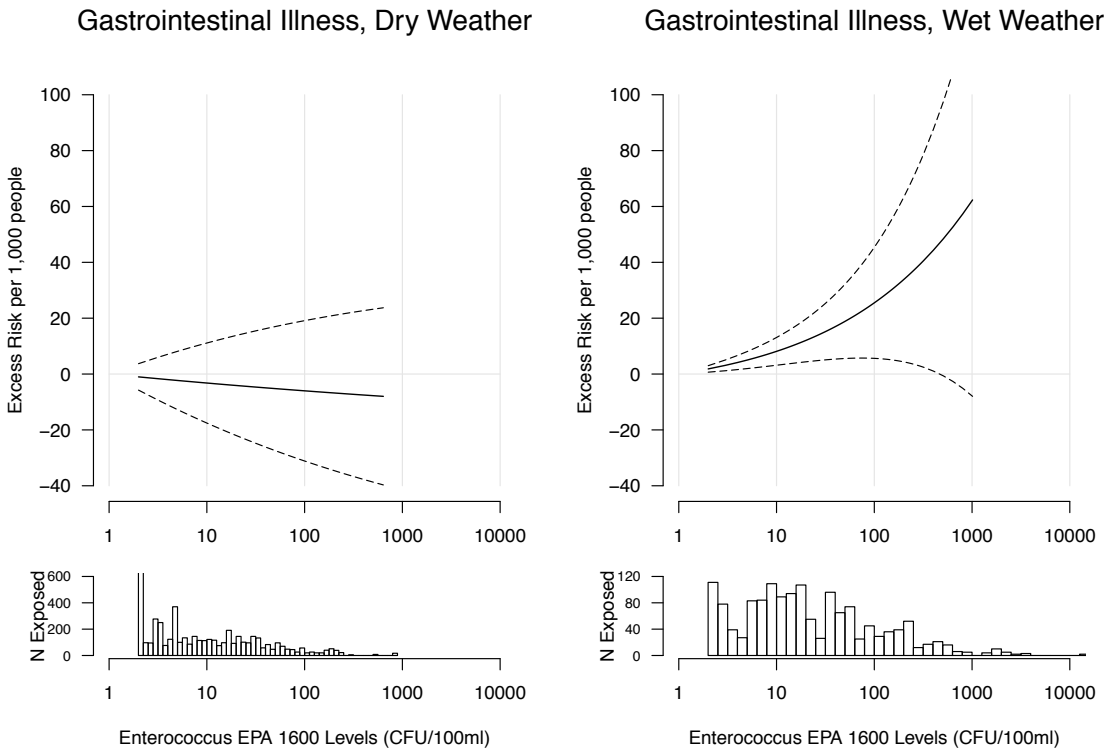
[c] Only measured in year 2

[d] Includes gastrointestinal illness, diarrhea, vomiting, eye infections, infected cuts and fever

[e] Adjusted for a range of covariates (see statistical methods for details)

[f] p-value for interaction term between water quality indicator and dry vs. wet weather

[g] Could not estimate due to sparse data



Web Figure 6: Excess risk of gastrointestinal illness associated with *Enterococcus* levels measured during dry and wet weather periods, predicted from a log-linear model among surfers at Tourmaline Surfing Park and Ocean Beach, San Diego, CA during the winters of 2013-14 and 2014-15. Dashed lines indicate model-based 95% confidence intervals and histograms show the distribution of *Enterococcus* exposure in the population during dry and wet weather periods.

Web Table 5: Negative control exposure analysis that matched daily mean log10 fecal indicator bacteria levels to surfers who were enrolled on that day, but had no seawater exposure for the past 5 days. Adjusted incidence rate ratios (IRRs) estimate the association between a 1-log10 increase in fecal indicator bacteria levels and incidence illness, stratified by dry and wet weather periods.

		Dry weather		Wet weather		
		IRR [c]	**(95% CI)**	**IRR [c]**	**(95% CI)**	**p [e]**
Enterococcus	Gastrointestinal illness	1.69	(0.80 , 3.56)	1.13	(0.39 , 3.25)	0.56
log_{10}	Diarrhea	1.43	(0.68 , 3.04)	0.82	(0.26 , 2.64)	0.46
	Sinus pain or infection	1.27	(0.64 , 2.50)	1.24	(0.35 , 4.42)	0.97
	Earache or infection	2.07	(0.60 , 7.10)	1.17	(0.31 , 4.32)	0.54
	Infection of open wound	-- [d]		-- [d]		
	Skin rash	0.88	(0.25 , 3.13)	-- [d]		
	Fever	1.21	(0.56 , 2.62)	0.82	(0.14 , 4.66)	0.69
	Upper respiratory illness [a]	1.34	(0.73 , 2.46)	0.46	(0.12 , 1.68)	0.15
	Any infectious symptom [b]	1.62	(0.86 , 3.04)	0.97	(0.33 , 2.84)	0.43
Fecal coliform	Gastrointestinal illness	1.69	(0.54 , 5.31)	0.46	(0.09 , 2.34)	0.22
log_{10}	Diarrhea	1.19	(0.32 , 4.36)	0.64	(0.13 , 3.07)	0.58
	Sinus pain or infection	1.08	(0.39 , 2.98)	0.72	(0.16 , 3.29)	0.67
	Earache or infection	2.43	(0.47 , 12.66)	1.77	(0.32 , 9.83)	0.79
	Infection of open wound	-- [d]		-- [d]		
	Skin rash	0.35	(0.05 , 2.75)	-- [d]		
	Fever	0.75	(0.17 , 3.25)	0.32	(0.06 , 1.67)	0.44
	Upper respiratory illness [a]	0.78	(0.28 , 2.20)	0.34	(0.10 , 1.13)	0.33
	Any infectious symptom [b]	1.29	(0.47 , 3.53)	0.34	(0.08 , 1.54)	0.17
Total coliform	Gastrointestinal illness	1.15	(0.49 , 2.70)	0.84	(0.22 , 3.16)	0.70
log_{10}	Diarrhea	0.81	(0.31 , 2.11)	0.80	(0.21 , 3.11)	0.99
	Sinus pain or infection	1.35	(0.71 , 2.55)	1.28	(0.31 , 5.25)	0.95
	Earache or infection	2.64	(1.17 , 5.97)	0.76	(0.32 , 1.79)	0.04
	Infection of open wound	-- [d]		-- [d]		
	Skin rash	0.96	(0.22 , 4.17)	-- [d]		
	Fever	1.61	(0.74 , 3.51)	0.90	(0.11 , 7.24)	0.61
	Upper respiratory illness [a]	1.69	(0.88 , 3.21)	0.47	(0.14 , 1.53)	0.06
	Any infectious symptom [b]	1.14	(0.57 , 2.28)	1.02	(0.28 , 3.70)	0.88

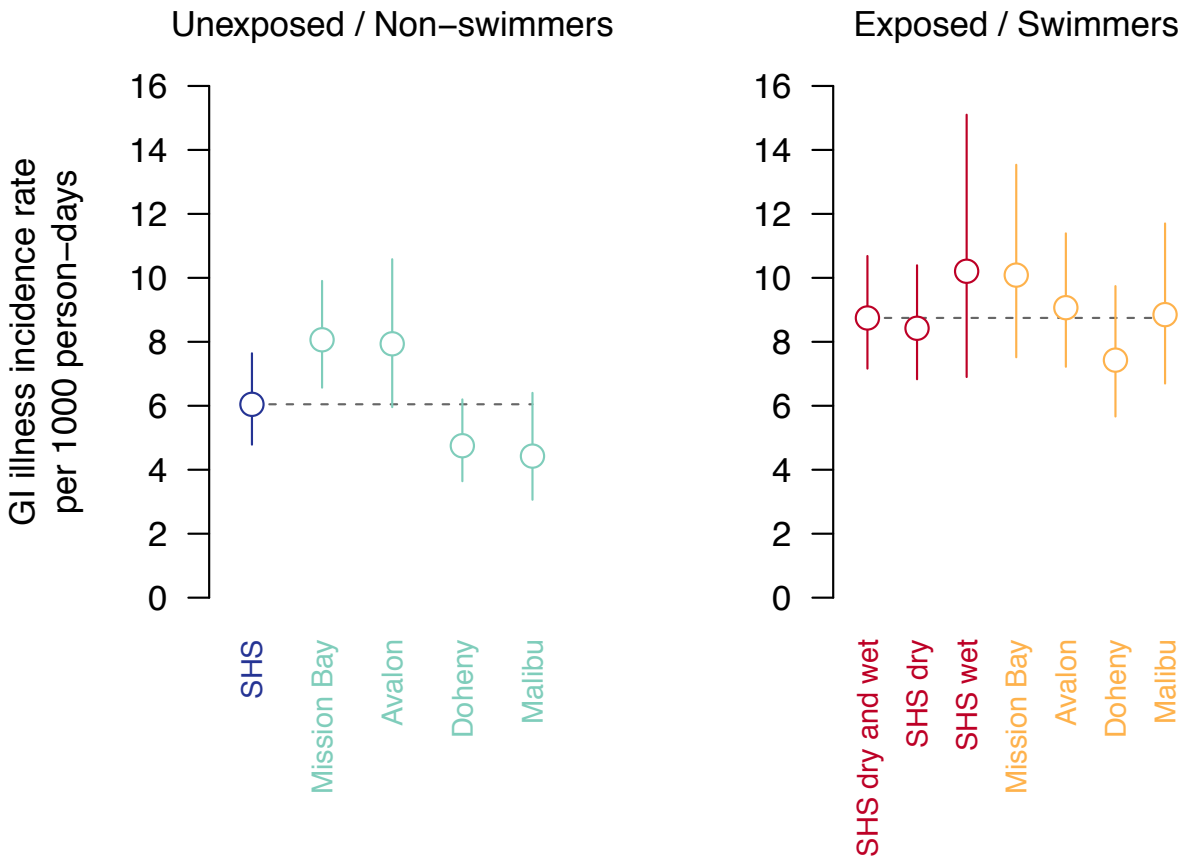
[a] Only measured in year 2

[b] Includes gastrointestinal illness, diarrhea, vomiting, eye infections, infected cuts and fever

[c] Adjusted for a range of covariates (see statistical methods for details)

[d] Could not estimate due to sparse data

[e] p-value for interaction term between water quality indicator and dry vs. wet weather



Web Figure 7: Incidence rates of gastrointestinal (GI) illness per 1,000 person-days in the present study (denoted SHS) and four previous summer swimmer cohort studies conducted in California. Swimmer cohorts were limited to adults (18 years or older) and swimmers included those with head immersion exposure. Rates in the present study are presented overall (dry and wet weather periods) and stratified by dry and wet weather. Vertical lines mark 95% confidence intervals. Horizontal dashed lines mark the present study rates to facilitate comparison with other estimates. Mission Bay = Mission Bay, San Diego [1]; Avalon = Avalon beach, Catalina island [2]; Doheny = Doheny State Beach [3]; Malibu = Malibu Surfrider State Beach [4].

References

1. 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;18:27–35.
2. 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.
3. 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;46:2176–2186.
4. 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;24:845–853.

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)	98	115 (63)	71*	104 (56)	60	219 (59)	66
Multiple brands and/or styles	18 (17)		15 (8)		7 (4)		22 (6)	
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.