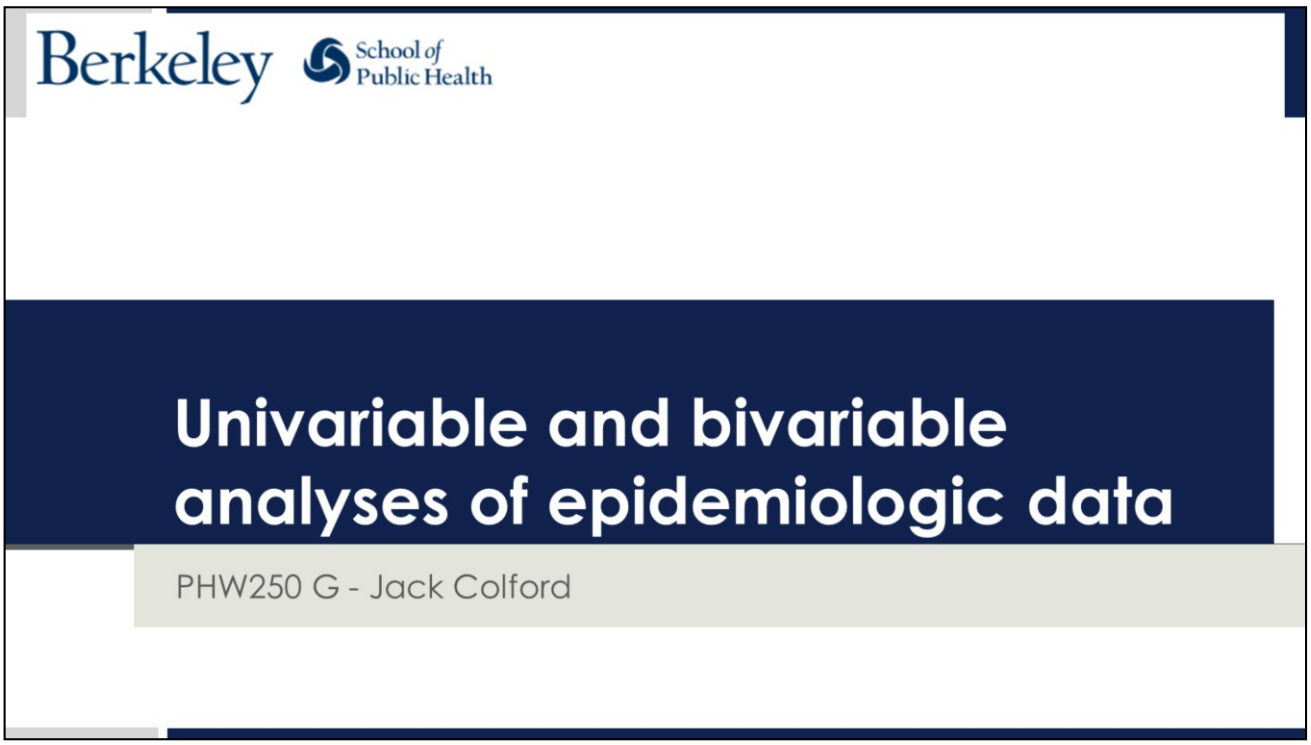
PHW 250B Week 13 Reader

Topic 1: Univariable and Bivariable Analyses

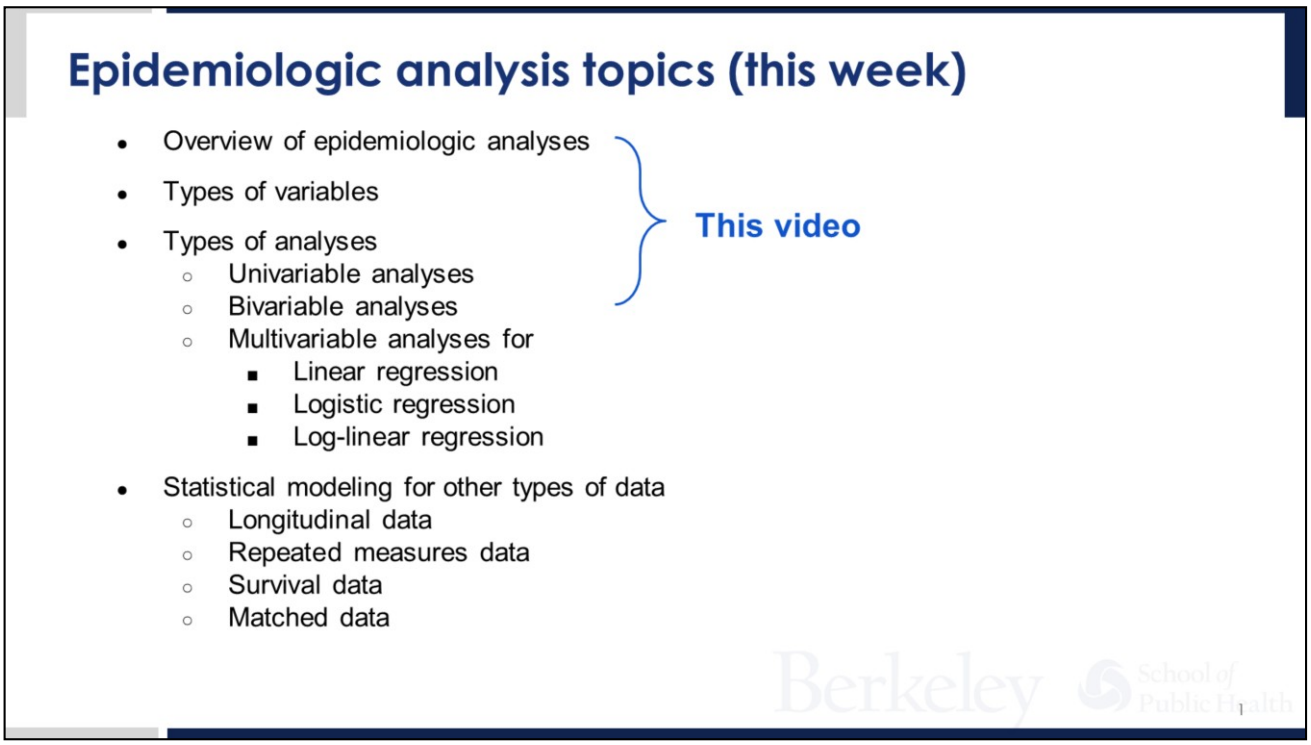
Topic 2: Multivariable Linear, Log-linear and Logistic Regression Models

Topic 3: Survival and Matched Case Control Analyses

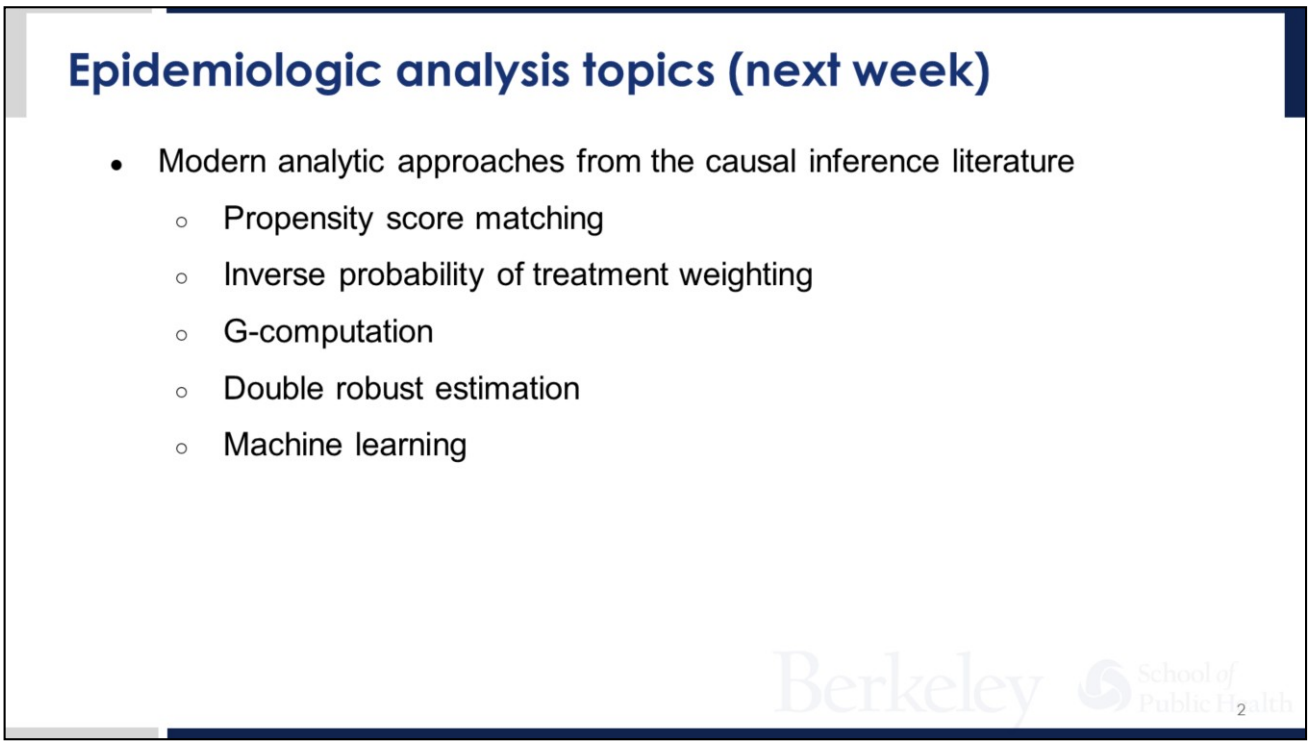
Journal Club



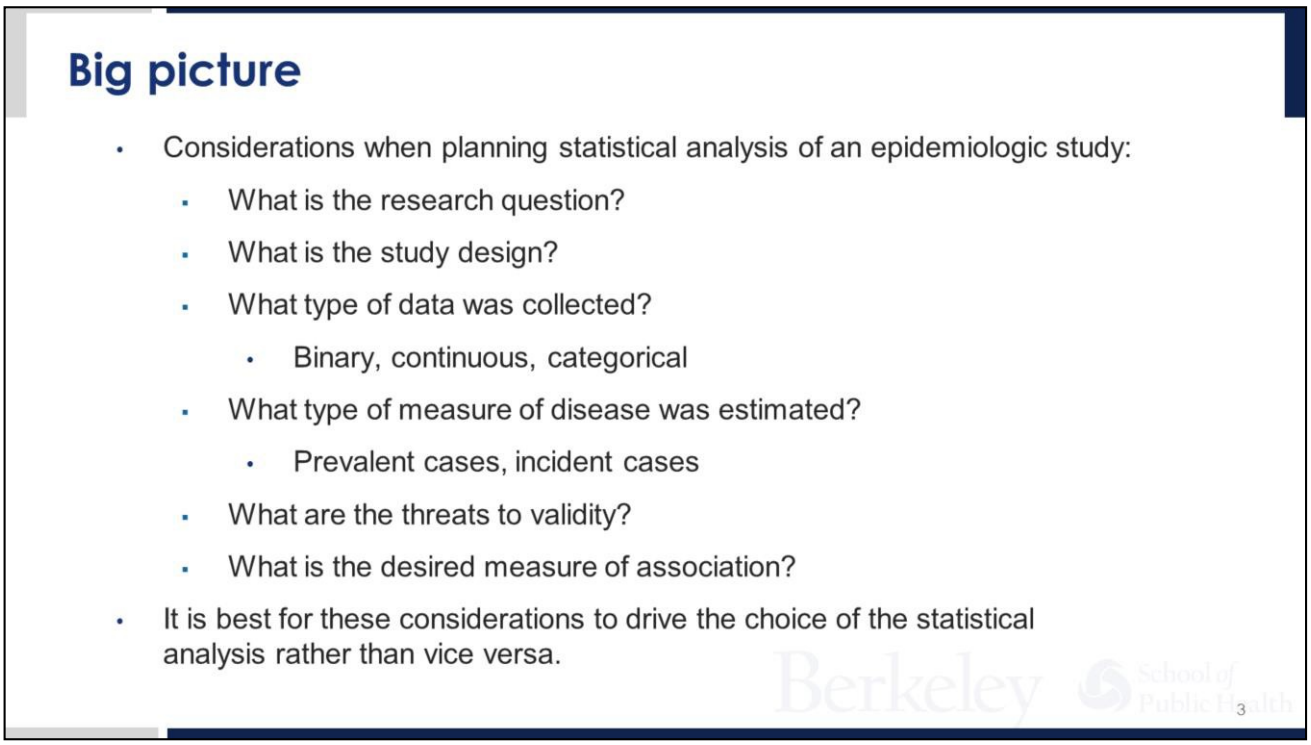
We're going to cover a number of analysis topics this week-- as you'll see on this slide. But in this video, we're going to focus first, with a broad overview of epianalysis in general. The types of variables used in these analysis. And then, in particular, univariable and bivariable analysis.



JACK COLFORD: In this lesson, we're going to discuss the analysis of univariable and bivariable epidemiologic data.



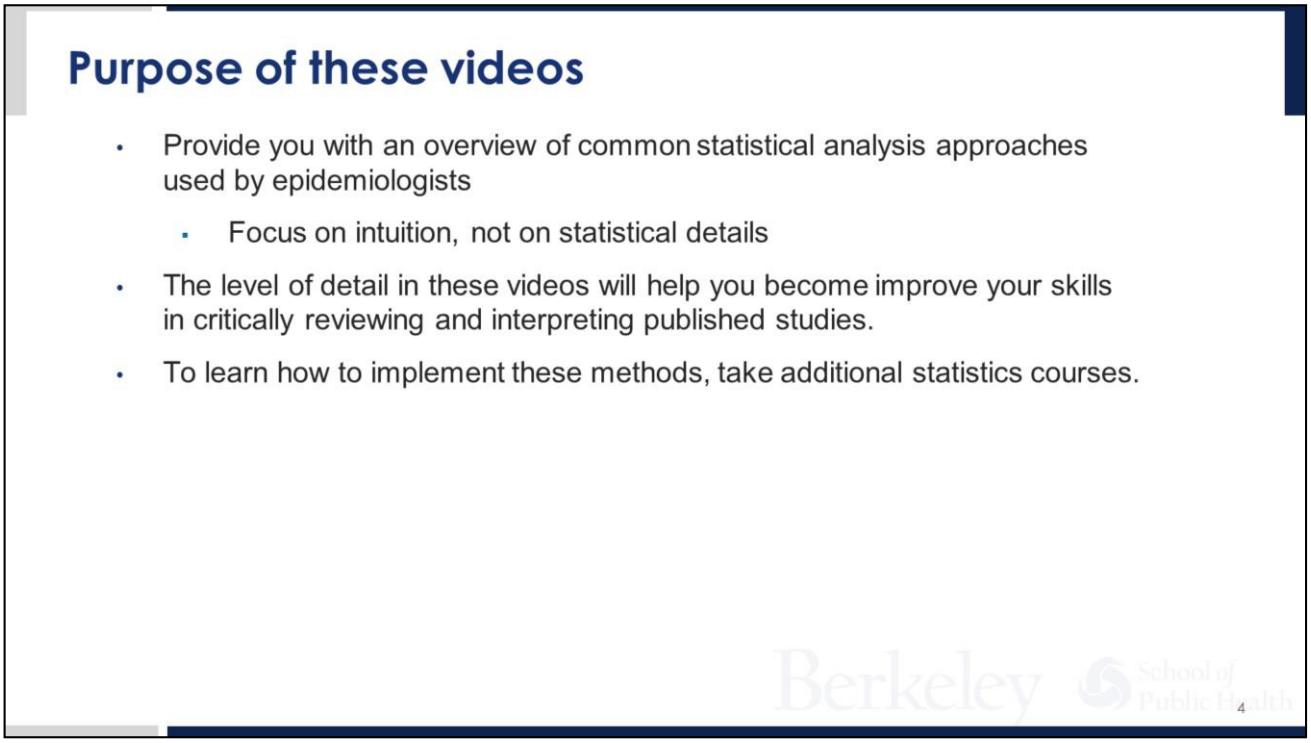
Next week, we will go forward and talk about modern analytic approaches from the causal inference literature. And these will include techniques like propensity score matching. Inverse probability of treatment weighting-- you'll often see that abbreviated IPTW. G-computation. Double robust estimation. And then, finally, machine learning.



Let's step back and take a look at the big picture and think about some of the considerations needed when planning a statistical analysis of epidata. So very basic approaches and questions need to start our analysis. And they are things like, what is the research question we're actually trying to answer? What was the study design?

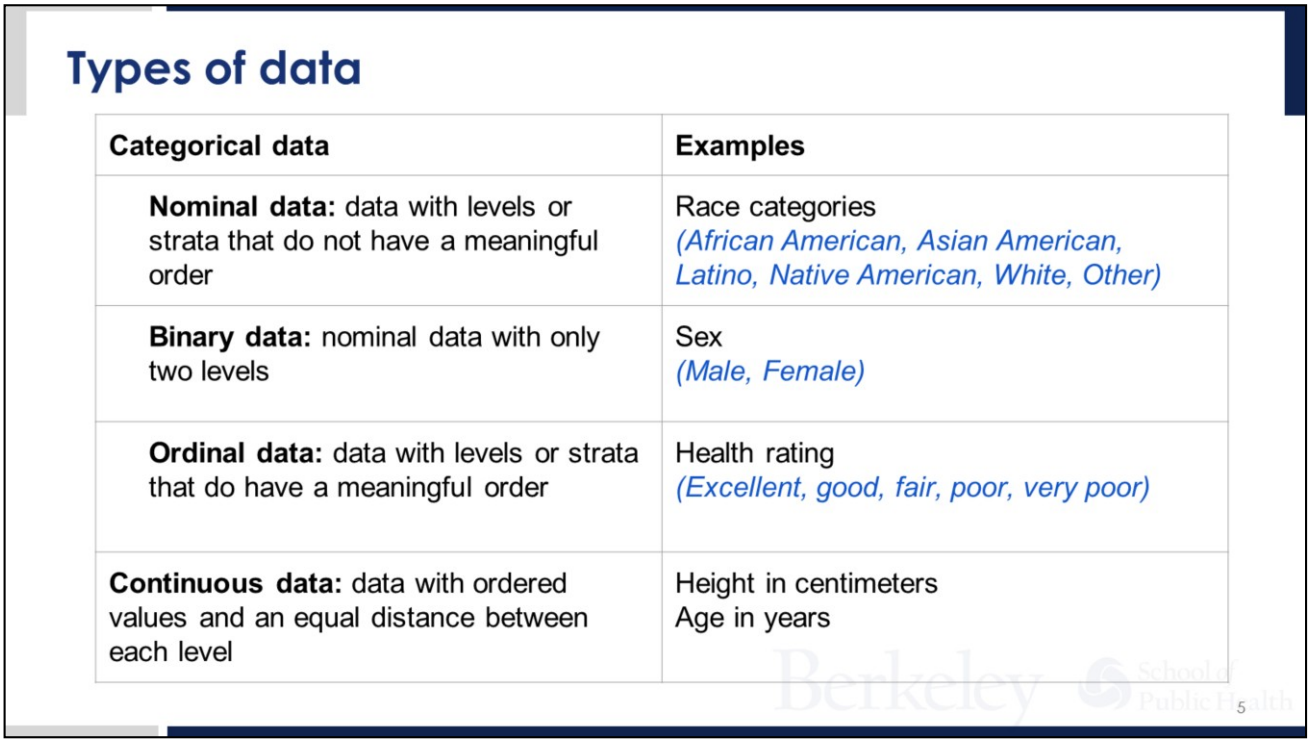
Next, what type of data were collected? Were they binary, continuous, categorical? What type of measure of disease was estimated? Did the authors estimate prevalence cases or incidence cases?

Are there threats to validity? We need to be aware of these. What is the desired measure of association? And it's best for these considerations to drive the choice of the statistical analysis rather than vise versa.



In these videos, we hope to provide you with an overview of common statistical analysis approaches used by epidemiologists that you'll often see in articles that you read or talks that you attend. Your focus should be on intuition, not on the statistical details. Those would come in later classes, if you chose to pursue more advanced work.

The level of detail in these videos will help improve your skills in critically reviewing and interpreting published studies. To learn how to implement these methods, take additional statistics courses.

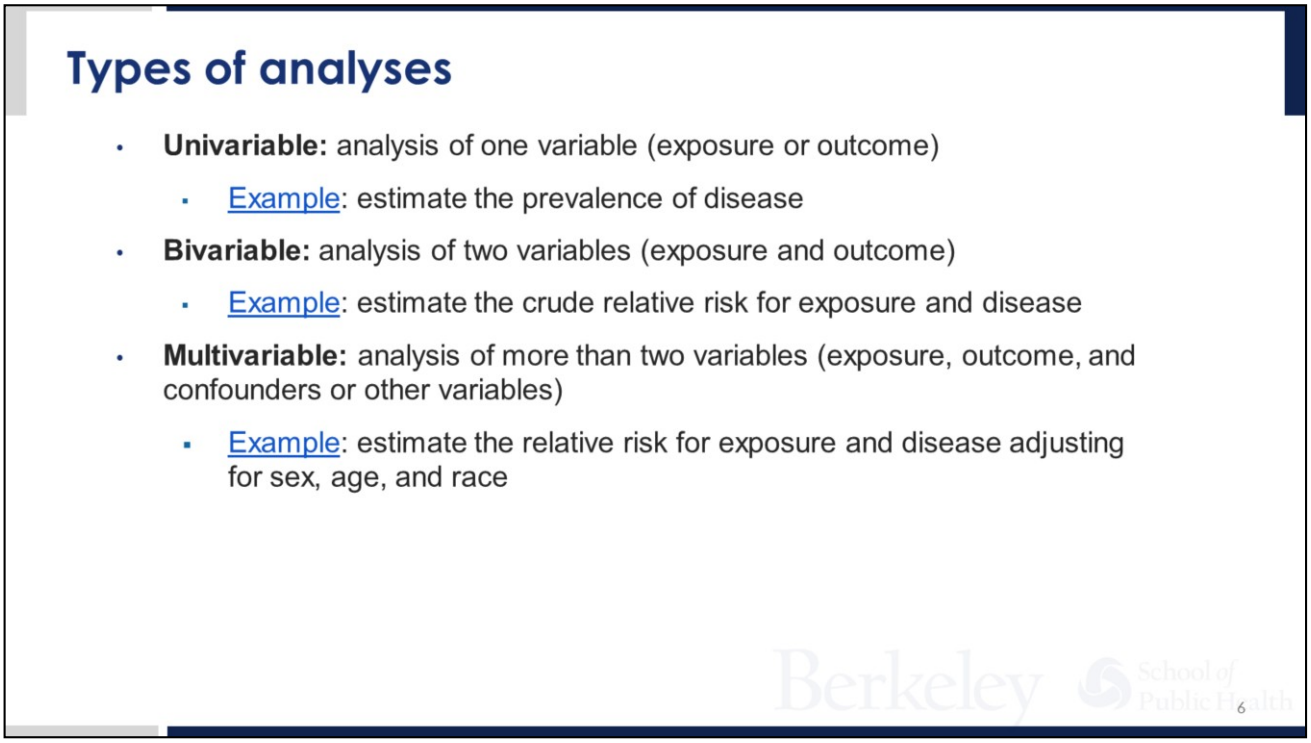


Let's first talk about the various types of data that an epidemiologist might encounter when analyzing data. The first example is of categorical data and we can divide categorical data into three different types. The first is nominal data-- that is data with levels or strata that do not have a meaningful order. So racial categories are an example of categorical nominal data-- such as African-American, Asian-American, Latino, Native American, white, and other. These have no particular order to them, they're just called nominal.

Binary data is a type of categorical data with only two levels. For instance, sex is often used as a binary variable with only two levels-- male and female. Even though we know that doesn't capture the full range of biologic possibilities.

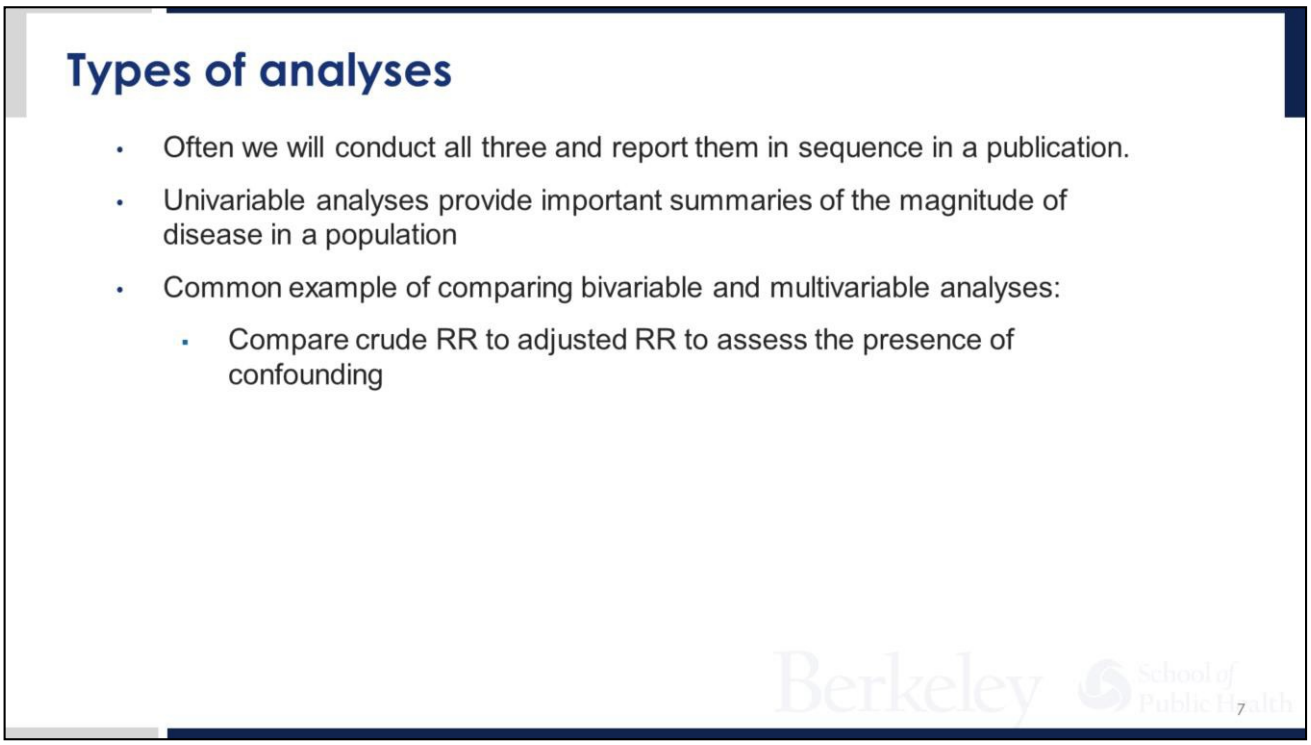
And finally, nominal data that's ordinal are data that have an order. They have levels or strata with a meaningful order. So for example, a health rating of excellent, good, fair, poor, or very poor has an order to it. It's a type of ordinal categorical data.

Next, a different type of data would be continuous data. These are data with ordered values and an equal distance between each level. So height in centimeters would be an example of continuous data. Age in years would be another example of continuous data.

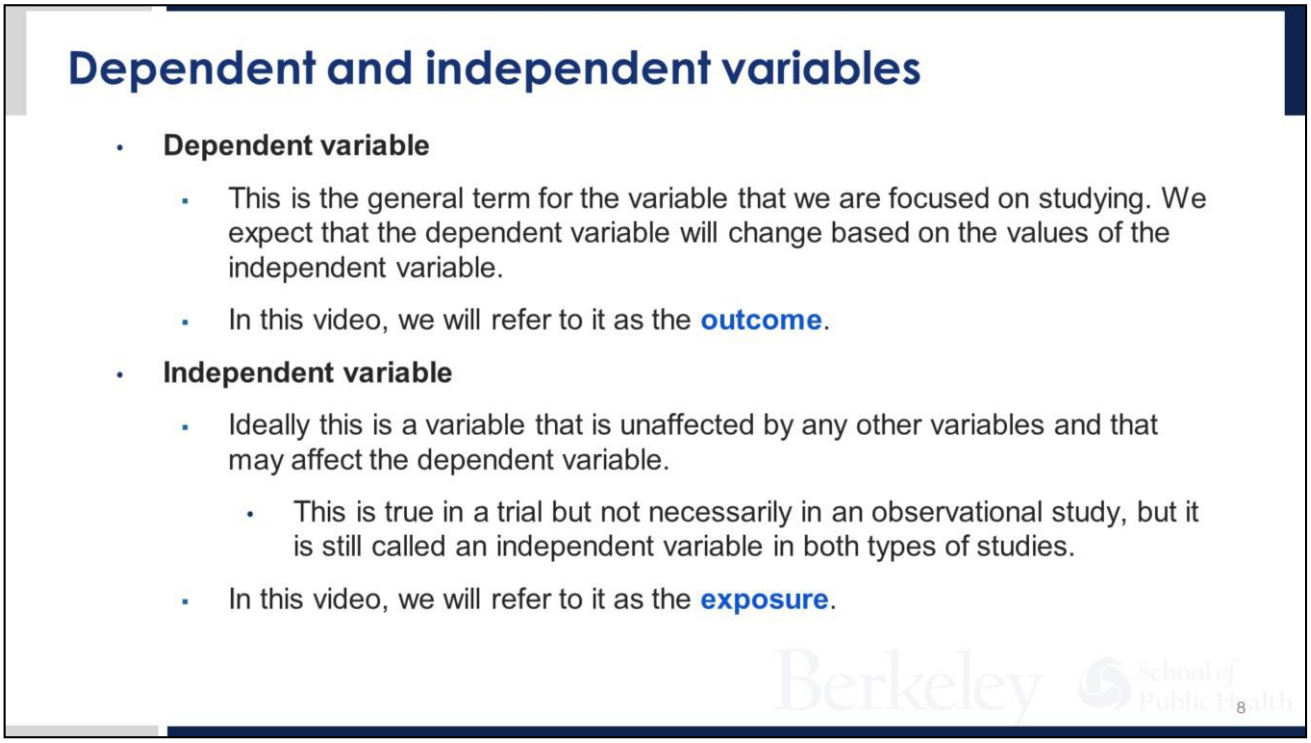


There are three broad types of analysis we might perform on our data. The first is univariable analysis, where-- as the name implies-- we analyze only one variable, such as exposure or outcome. For example-- if we estimate the prevalence of disease, that is a univariable analysis.

Bivariable analysis involves the analysis of two variables, such as exposure and outcome. An example here might be estimation of the crude relative risk for exposure and disease. Multivariable analysis would be the analysis of more than two variables, such as exposure, outcome, and confounders or other variables. An example of this might be to estimate the relative risk for exposure and disease adjusting for sex, age, and race.



All three of these types of analysis will be conducted and then reported in sequence in a publication, each one a little more complicated than the prior. Variable analysis provide important summaries of the magnitude of disease in a population. A common example of comparing bivariable and multivariable analysis might be to compare the crude relative risk to the adjusted relative risk, to assess the presence of confounding. Remember-- when the adjusted relative risk is different in a meaningful way from the crude relative risk, we say that confounding is present.



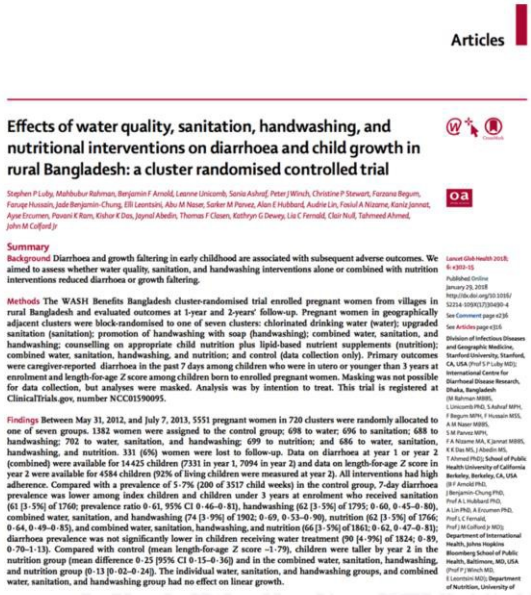
Another way to describe variables is whether they are dependent or independent. Dependent variable is the term, in general, for a variable that we are focused on studying. We expect that the dependent variable will change based on the values of the independent variable. So we often refer in epistudies to the dependent variable as the outcome variable.

An independent variable is ideally, a variable that's unaffected by any other variables. And it may affect the dependent variable. This is true in a trial, but not necessarily an observational study. But it's still called an independent variable in both types of studies.

In this video, we'll refer to it as the exposure. So generally, we think of the exposure as the independent variable affecting the dependent variable or outcome. But of course, the exposure itself might actually be affected by other variables. But for the purposes of our discussion here, let's just refer to the main exposure variable as our independent variable.

Case study in this video: WASH Benefits

- We will examine examples of univariable and bivariable analyses in this trial.
- The research question drove the study design and choice of outcome measures, which drove the choice of statistical analyses.
- The publication reported univariable, bivariable, and multivariable analyses.



Articles

Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural Bangladesh: a cluster randomised controlled trial

oa

Stephen P Luby, Mahbubur Rahman, Benjamin F Arnold, Leanne Unicomb, Sania Ashraf, Peter J Winch, Christine P Stewart, Farzana Begum, Faruqe Hussain, Jade Benjamin-Chung, Elli Leontsini, Abu M Naser, Sarker M Parvez, Alan E Hubbard, Audrie Lin, Faisul A Nizame, Kaniz Jannat, Ayse Ercumen, Pavani K Ram, Kishor K Das, Jaynal Abedin, Thomas F Clasen, Kathryn G Dewey, Lia C Fernald, Clair Null, Tahmeed Ahmed, John M Colford Jr

Summary

Background Diarrhoea and growth faltering in early childhood are associated with subsequent adverse outcomes. We aimed to assess whether water quality, sanitation, and handwashing interventions alone or combined with nutrition interventions reduced diarrhoea or growth faltering.

Methods The WASH Benefits Bangladesh cluster-randomised trial enrolled pregnant women from villages in rural Bangladesh and evaluated outcomes at 1-year and 2-years' follow-up. Pregnant women in geographically adjacent clusters were block-randomised to one of seven clusters: chlorinated drinking water (water); upgraded sanitation (sanitation); promotion of handwashing with soap (handwashing); combined water, sanitation, and handwashing; counselling on appropriate child nutrition plus lipid-based nutrient supplements (nutrition); combined water, sanitation, handwashing, and nutrition; and control (data collection only). Primary outcomes were caregiver-reported diarrhoea in the past 7 days among children who were in utero or younger than 3 years at enrolment and length-for-age *Z* score among children born to enrolled pregnant women. Masking was not possible for data collection, but analyses were masked. Analysis was by intention to treat. This trial is registered at ClinicalTrials.gov, number NCC01590095.

Findings Between May 31, 2012, and July 7, 2013, 5551 pregnant women in 720 clusters were randomly allocated to one of seven groups. 1382 women were assigned to the control group; 698 to water; 696 to sanitation; 688 to handwashing; 702 to water, sanitation, and handwashing; 699 to nutrition; and 686 to water, sanitation, handwashing, and nutrition. 331 (6%) women were lost to follow-up. Data on diarrhoea at year 1 or year 2 (combined) were available for 14 425 children (7331 in year 1, 7094 in year 2) and data on length-for-age *Z* score in year 2 were available for 4584 children (92% of living children were measured at year 2). All interventions had high adherence. Compared with a prevalence of 5·7% (200 of 3517 child weeks) in the control group, 7-day diarrhoea prevalence was lower among index children and children under 3 years at enrolment who received sanitation (61 [3·5%] of 1760; prevalence ratio 0·61, 95% CI 0·46–0·81), handwashing (62 [3·5%] of 1795; 0·60, 0·45–0·80), combined water, sanitation, and handwashing (74 [3·9%] of 1902; 0·69, 0·53–0·90), nutrition (62 [3·5%] of 1766; 0·64, 0·49–0·85), and combined water, sanitation, handwashing, and nutrition (66 [3·5%] of 1861; 0·62, 0·47–0·81); diarrhoea prevalence was not significantly lower in children receiving water treatment (90 [4·9%] of 1824; 0·89, 0·70–1·13). Compared with control (mean length-for-age *Z* score −1·79), children were taller by year 2 in the nutrition group (mean difference 0·25 [95% CI 0·15–0·36]) and in the combined water, sanitation, handwashing, and nutrition group (0·13 [0·02–0·24]). The individual water, sanitation, and handwashing groups, and combined water, sanitation, and handwashing group had no effect on linear growth.

Lancet Glob Health 2018; 6: e302–15

Published Online January 29, 2018 http://dx.doi.org/10.1016/S2214-109X(17)30490-4

See Comment page e236

See Articles page e316

Division of Infectious Diseases and Geographic Medicine, Stanford University, Stanford, CA, USA (Prof S P Luby MD); International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh (M Rahman MBBS, L Unicomb PhD, S Ashraf MPH, F Begum MPH, F Hussain MSS, A M Naser MBBS, S M Parvez MPH, F A Nizame MA, K Jannat MBBS, K K Das MS, J Abedin MS, T Ahmed PhD); School of Public Health University of California Berkeley, Berkeley, CA, USA (B F Arnold PhD, J Benjamin-Chung PhD, Prof A L Hubbard PhD, A Lin PhD, A Ercumen PhD, Prof L C Fernald, Prof J M Colford Jr MD); Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA (Prof P J Winch MD, E Leontsini MD); Department of Nutrition, University of

9

Here's a case study from the WASH Benefit study-- a study we've discussed earlier in the course. And in this specific example, we'll examine examples of univariable and bivariable analysis in the trial. The research question of this study is, what drove the design and choice of outcome measures? And that drove the choice of statistical analysis. The publication reported univariable, bivariable. and multivariable analysis.

Univariable analyses

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Continuous	0	N/A	Univariable	Mean, median
Categorical	0	N/A	Univariable	Absolute number, percentage, prevalence, incidence

- **Example for continuous outcome:** height-for-age Z-score
  - Mean height-for-age Z-score at specific ages
- **Example for categorical outcome:** diarrhea status (yes / no)
  - Prevalence of diarrhea at a specific age

Berkeley School of Public Health

Here's some examples of univariable analysis in the WASH Benefits trial. There were continuous outcome measures or a continuous outcome measure for the height-for-age Z-score. There were many other continuous outcome measures, of course, as well, but the principle one was the height-for-age Z-score. This is specifically the mean height for age expressed as a Z-score at specific ages of the child.

An example of categorical outcome variables in this study was diarrhea status, which was a yes/no variable. And here, this is telling us about the prevalence of diarrhea at a specific age. Is diarrhea present in a child or not at a specific age? And then across all the children at a specific age, how many of them had diarrhea?

So the two types of variables represented here are continuous for the height-for-age Z-score and categorical for the prevalence of diarrhea at a specific age. I just don't want you to be confused that the categorical variable is a yes/no variable, but when you add it up across a number of children, then you can-- for example-- have a prevalence from it.

So the analysis method for continuous variables in this study were to look at the mean and the median. And the categorical variable analysis methods were to look at the absolute number, the percentage, the prevalence, the incidence of diarrhea.

Example of univariable analysis in WASH Benefits

	N	Mean* prevalence
Control *vs* intervention		
Control	3517	5·7%
Water	1824	4·9%
Sanitation	1760	3·5%
Handwashing	1795	3·5%
Water, sanitation, and handwashing	1902	3·9%
Nutrition	1766	3·5%
Water, sanitation, handwashing, and nutrition	1861	3·5%

Table 4: **Diarrhoea prevalence 1 and 2 years (combined) after intervention**

Berkeley School of Public Health

11

Here's an example of univariable analysis in WASH Benefits. So this is looking at the mean prevalence of diarrhea for the children at age one and two years. And combining all the data for one and two years to have a really large dataset. So you see that in the control group, the prevalence of diarrhea for the one and two year analysis combined was 5.7%. That means 5.7% of the children at age one and two years had diarrhea.

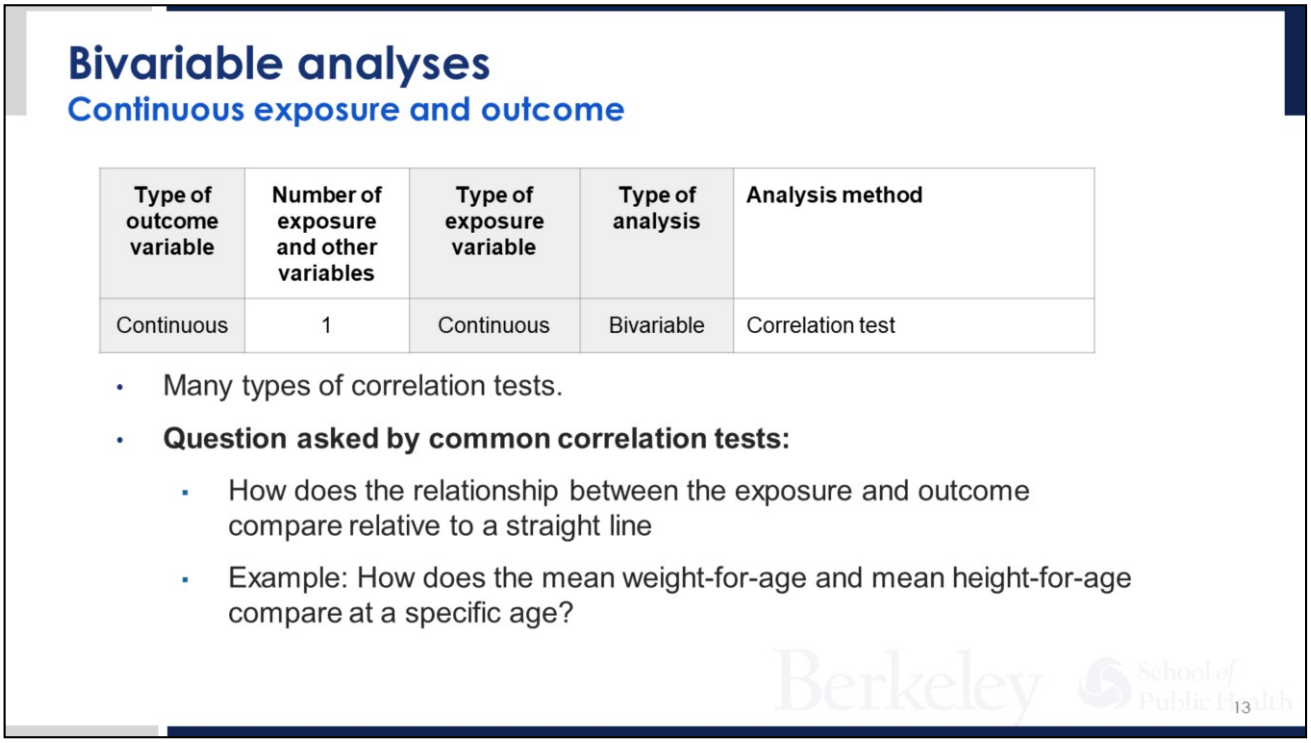
And then you see for each of the intervention arms-- six other intervention arms-- the prevalence of diarrhea at those points. So what we're going to need to be doing eventually is comparing each of those other values against the 5.7%. The control serves as our reference group.

Bivariable analyses

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Continuous	1	Continuous	Bivariable	Correlation test
Categorical	1	Categorical	Bivariable	Chi-square test
Continuous	1	Binary	Bivariable	T-test
Continuous	1	Categorical	Bivariable	Analysis of Variance (ANOVA)
Categorical	1	Continuous	Bivariable	Simple logistic or log-linear regression

Berkeley School of Public Health

Now, a number of bivariable analysis are conducted using different analysis methods to look at both continuous and categorical outcome variables. And you see here, a table showing the number of exposure and other variables in each of these. The type of exposure variable. And the type of analysis that's needed-- all of these are bivariable analysis. And the specific analysis method we'll go through, which include the correlation test, the chi-square test, the t-test, and analysis of variance or ANOVA. And then simple logistic or log linear regression.



Let's focus for a moment on a bivariable analysis looking at both the continuous exposure and a continuous outcome. So here, we're looking at one exposure variable that's continuous and one outcome variable that's continuous. And we're going to use a correlation test.

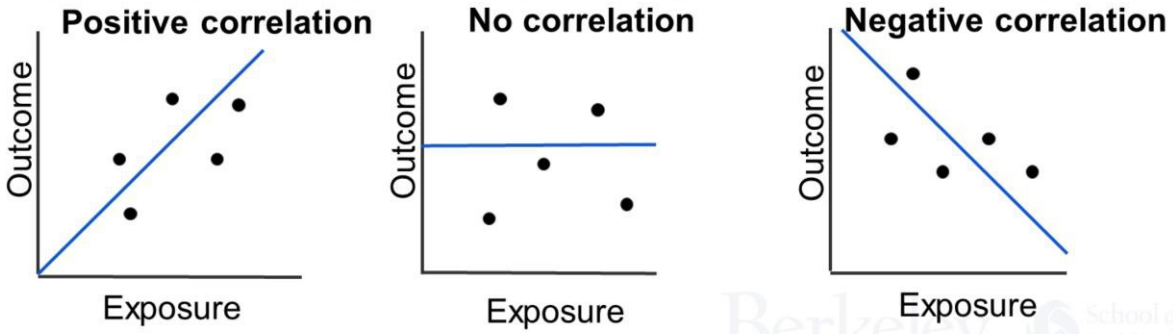
There are many types of correlation test. And the question that they commonly ask is, how does the relationship between the exposure and outcome compare relatively to a straight line? So if there's no correlation or relationship, we would plot their relationship with a straight line. If the relationship is not a straight line, then it implies some sort of correlation.

So a specific question for a bivariable analysis here might be-- how does the mean weight-for-age and the mean height-for-age compare at a specific age?

Bivariable analyses

Continuous exposure and outcome

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Continuous	1	Continuous	Bivariable	Correlation test



Berkeley School of Public Health

14

Here's three plots that show possible relationships.

So as I said-- there's a correlation. We have a non-straight line so we could have a positive or negative correlation. Positive correlation implies as the exposure gets higher, the outcome gets higher. Negative correlation applies as the exposure gets higher, the outcome value gets lower. And no correlation means that the best line plotting our actual data shows a straight line-- there's no relationship.

Bivariable analyses

Categorical exposure and outcome

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Categorical	1	Categorical	Bivariable	Chi-square test

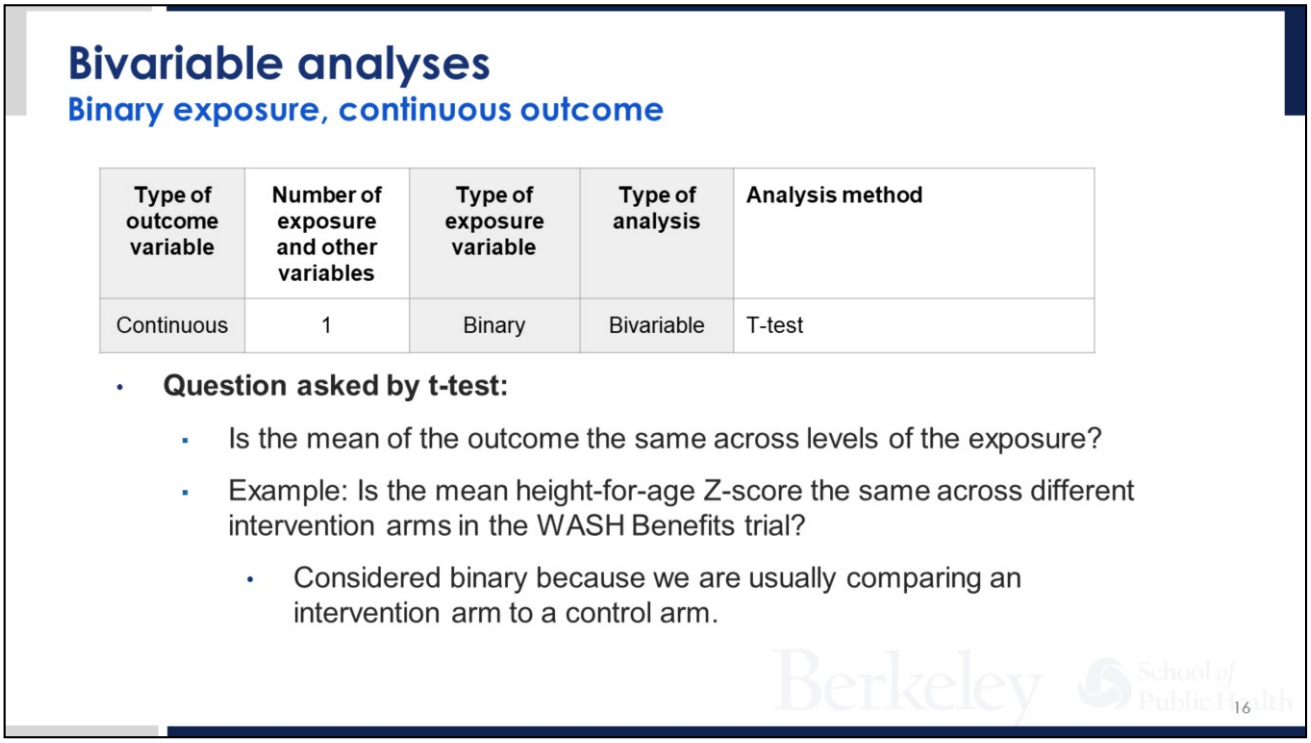
- **Question asked by Chi-square test for independence:**
  - Is the proportion of the outcome the same across levels of the exposure?
  - Example: Is the prevalence of diarrhea the same across different intervention arms in the WASH Benefits trial?
- Note: the chi-square test for independence is different from the chi-square test for homogeneity (used to assess the presence of effect modification)

Berkeley School of Public Health

15

Let's look now at a categorical exposure and outcome. And the question being asked here is going to be addressed with a statistical test called a chi-square test for independence. So what we're asking is, is the proportion of the outcome the same across different levels of the exposure?

And an example is, is the prevalence of diarrhea the same across different intervention arms in the WASH Benefits trial? So one note is that the chi-square test for independence is different from the chi-square test for homogeneity-- which we used earlier in the course to assess the presence of effect modification.



What about a situation where we have a binary exposure and a continuous outcome? In this situation, we use a statistical test called the t-test. And the t-test is asking, is the mean of the outcome the same across the level of the exposure?

In this study-- for example-- a specific use of the t-test would be, is the mean height-for-age Z-score the same across different intervention arms in the WASH Benefits trial. This is considered binary because we are usually comparing an intervention to a control arm-- those are two arms, that's why it's binary.

Bivariable analyses

Categorical exposure, continuous outcome

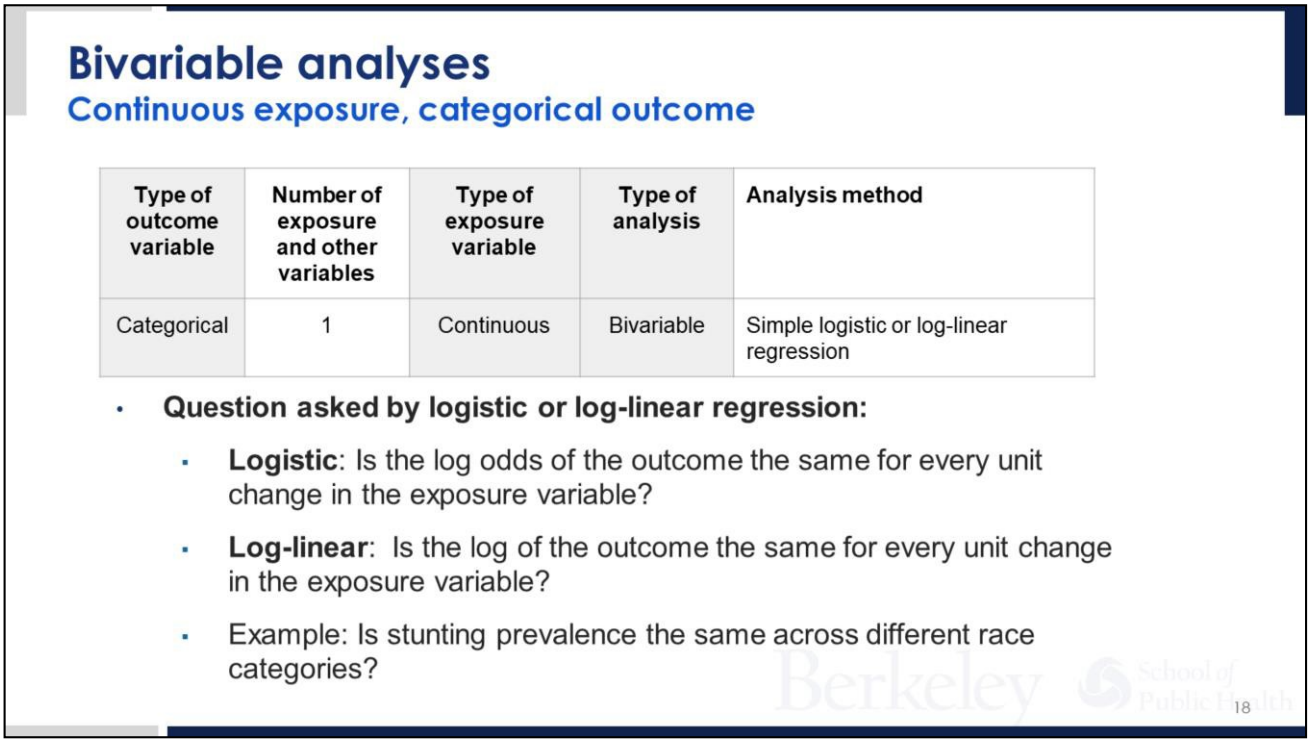
Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Continuous	1	Categorical	Bivariable	Analysis of Variance (ANOVA)

- **Question asked by ANOVA:**
  - Is the mean of the outcome the same across levels of the exposure?
  - Example: Is the mean height-for-age Z-score the same across different race categories?

Berkeley School of Public Health

The next example looks at a categorical exposure and a continuous outcome. And uses a statistical technique called analysis of variance, most often referred to as ANOVA-- an acronym for analysis of variance. And the question being asked by the ANOVA analysis is whether the mean of the outcome is the same across all the levels of the exposure.

And a specific example to realize this question is, is the mean height-for-age Z-score the same across different race categories? So if we look at the height-for-age Z-score in different racial categories or whatever different groups we choose to categorize the study by-- is the mean height-for-age Z-score the same across those different race categories?



Let's look at another bivariable analysis looking at continuous exposure and categorical outcome. So for this type of question, there are two analysis methods that we'll learn about now. One's called simple logistic and one's called log linear regression. And they both ask the question, is the log of the odds of the outcome the same for every unit change in the exposure variable?

So an example of a question where this might be applied is, is the mean height-for-age Z-score the same across different racial categories? So if I have five different races and they each have a mean height-for-age Z-score, these are two techniques I could use, again, to compare those means across the different racial categories.

Example of bivariable analyses in WASH Benefits

	N	Mean* prevalence	Unadjusted† prevalence difference (95% CI)
Control vs intervention			
Control	3517	5·7%	..
Water	1824	4·9%	−0·6 (−1·9 to 0·6)
Sanitation	1760	3·5%	−2·2 (−3·4 to −1·0)
Handwashing	1795	3·5%	−2·3 (−3·4 to −1·1)
Water, sanitation, and handwashing	1902	3·9%	−1·7 (−2·9 to −0·6)
Nutrition	1766	3·5%	−2·0 (−3·1 to −0·8)
Water, sanitation, handwashing, and nutrition	1861	3·5%	−2·2 (−3·3 to −1·0)

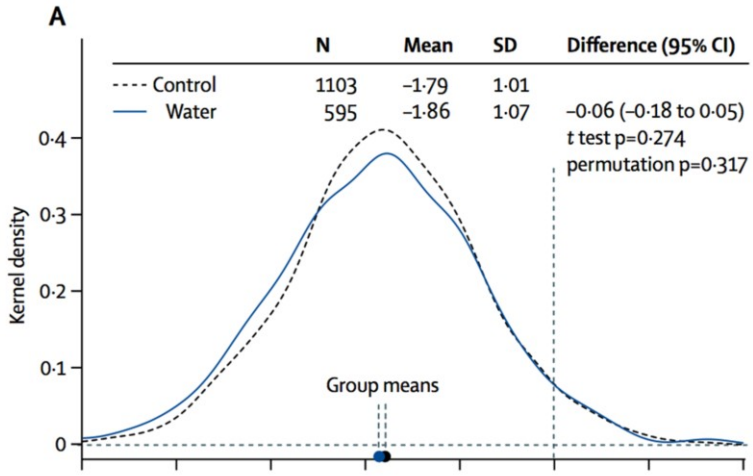
Table 4: **Diarrhoea prevalence 1 and 2 years (combined) after intervention**

Berkeley School of Public Health

Here's an example of a bivariable analysis in WASH Benefits. So we're looking at the mean values of diarrhea prevalence in each of the different intervention arms, including the control arm. And what we're showing is the unadjusted prevalence difference. So for example-- in the water arm, 5.7 minus 4.9 is a difference of minus 0.6. And then you see a confidence interval that's been calculated for each of these differences, as well.

Example of bivariable analyses in WASH Benefits

Figure 3: Intervention effects on length-for-age Z scores in 4584 children after 2 years of intervention



Berkeley School of Public Health

Here's another example of a different way to present these data for the intervention effect comparing the control arm to the water only arm. And you see the length-for-age Z-score here so this is a continuous analysis, here. Continuous variables being analyzed and we're comparing the mean value in the water group to the mean value in the control group.

And you see here, a difference of minus 0.06. And there are statistical tests associated with this difference. And the t-test here gives a p-value of 0.274. Permutation test gives a p-value of 0.317.

Summary of univariable and bivariable analyses

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Continuous	0	N/A	Univariable	Mean, median
Categorical	0	N/A	Univariable	Absolute number, percentage
Continuous	1	Continuous	Bivariable	Correlation test
Categorical	1	Categorical	Bivariable	Chi-square test
Continuous	1	Binary	Bivariable	T-test
Continuous	1	Categorical	Bivariable	Analysis of Variance (ANOVA)
Categorical	1	Continuous	Bivariable	Simple logistic or log-linear regression

Berkeley School of Public Health

21

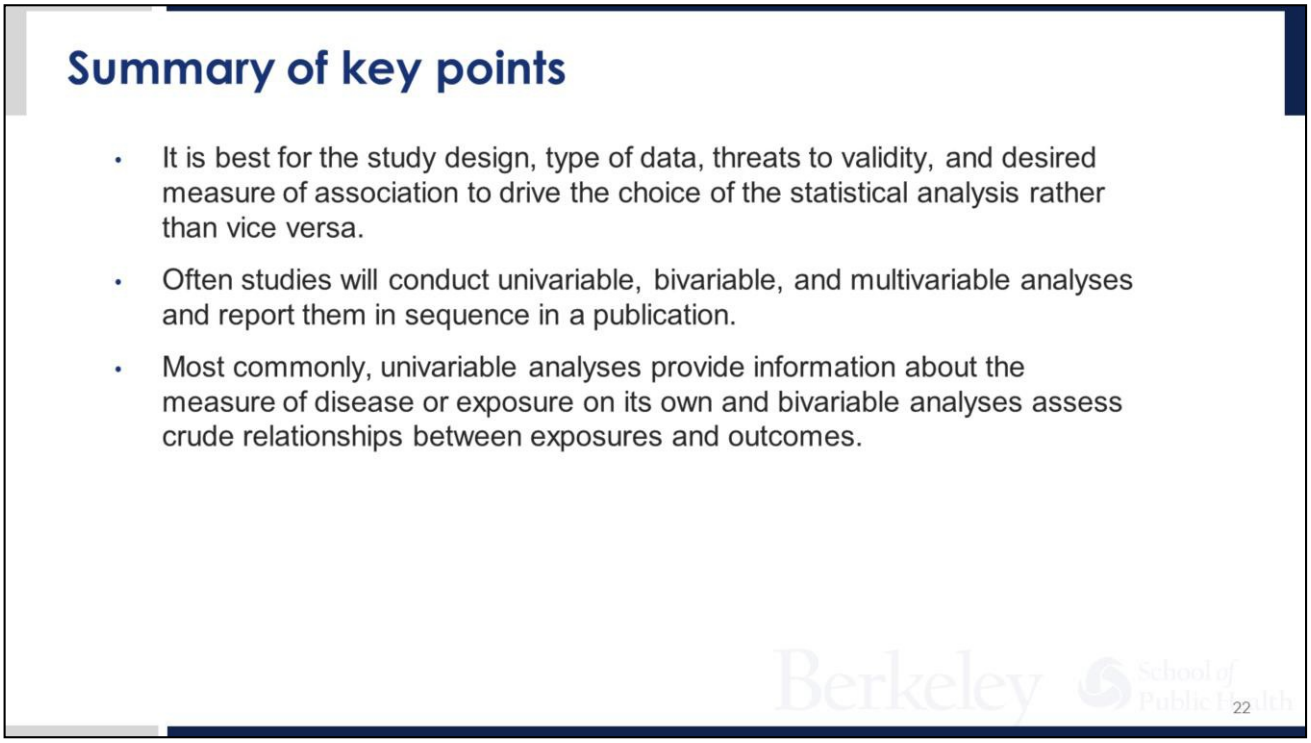
So to summarize the univariable and bivariable analysis we've discussed, we think about what our exposure and outcome variables have as their structure. So if the outcome variable is continuous and we're only looking at outcome variable, we're going to analyze that with techniques that look at the mean or median. If the outcome variable is categorical and we don't have an exposure variable-- it's just a univariable analysis, again-- then we look at the absolute number or the percentage.

If the outcome variable is continuous and the exposure variable is also continuous, that's a bivariable analysis and we use a correlation test. If the outcome variable is categorical and the exposure variable is also categorical, that's, again, bivariable analysis. And for that situation, we use a chi-square test.

The outcome variable's continuous, the exposure variable is binary. Then we use a t-test as our analysis method.

The outcome variable is continuous, the exposure variable is categorical-- we would use a technique called analysis of variance or ANOVA.

And finally, if the actual outcome variable is categorical and the exposure variable is continuous, we have two techniques to use. One is called simple logistic regression or log linear regression.



So to summarize the key points in this module, it's best when the study design, the type of data, reduction in threats to validity, and desired measure of association are what drive the choice of the statistical analysis that you do rather than vise versa. Sometimes studies will conduct univariable, bivariable, and multivariable analysis and report them in sequence in a publication. And that is often recommended.

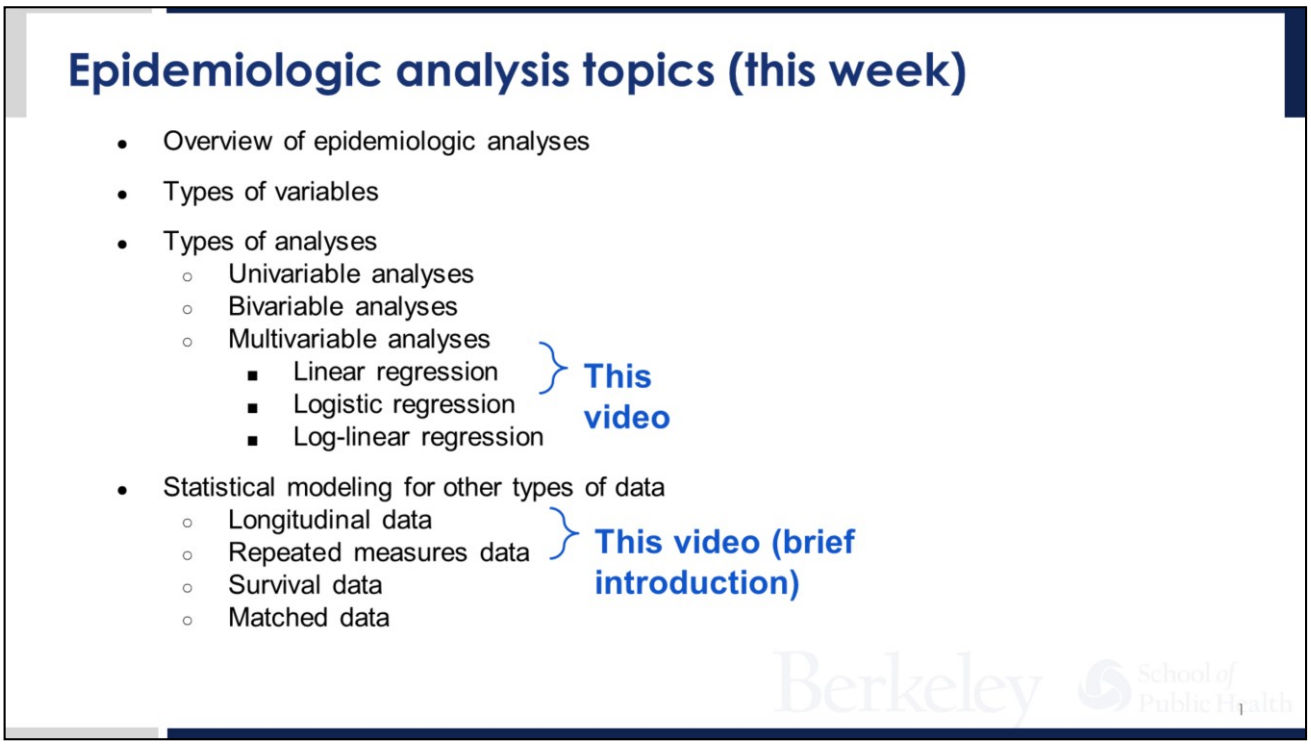
And most commonly, univariable analysis provide information about the measure of disease or exposure on their own. And bivariable analysis assess crude relationships between exposures and outcomes.

Berkeley School of Public Health

Multivariable linear regression models

PHW250 B – Andrew Mertens

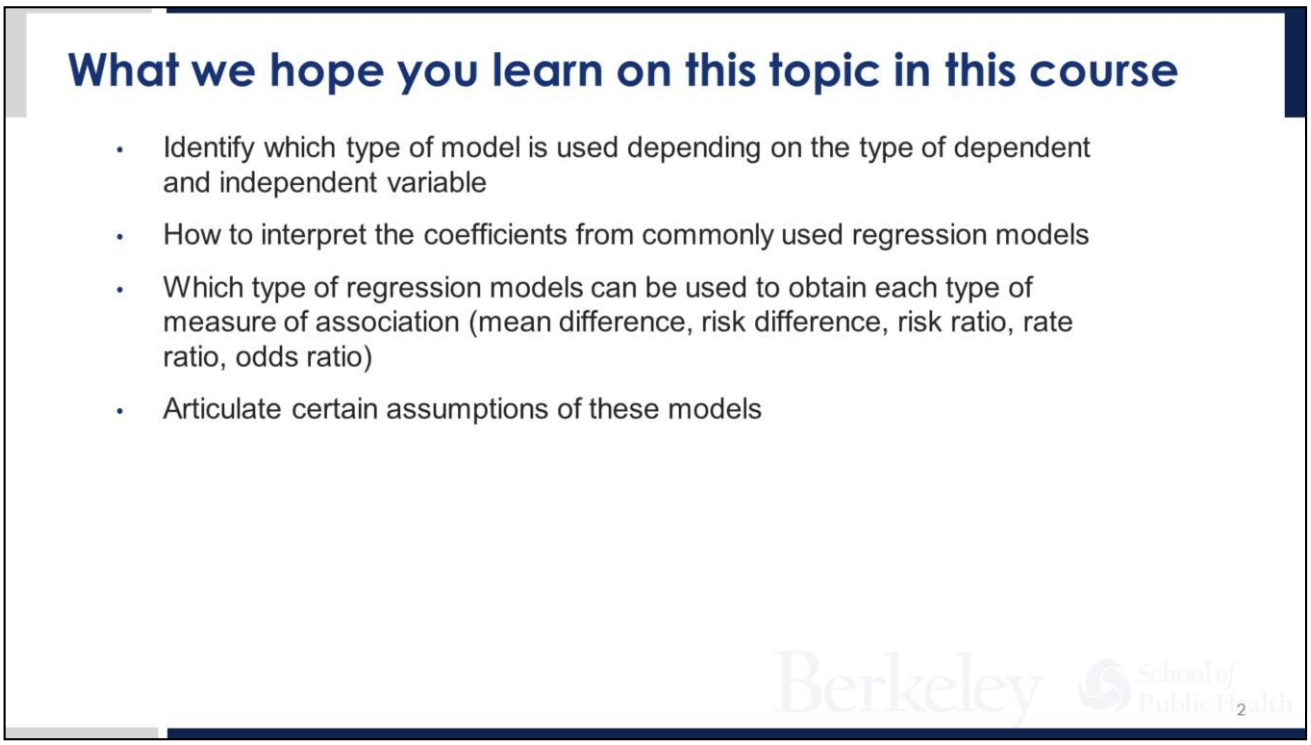
Andrew Mertens: This video covers multivariable linear regression models.



Here's the list of epidemiologic analysis topics that we're covering this week. You've already seen an overview of epidemiologic analysis. We've reviewed types of variables used in the analysis. And the prior video discussed univariable and bivariable analysis that had one outcome variable or one exposure and one outcome.

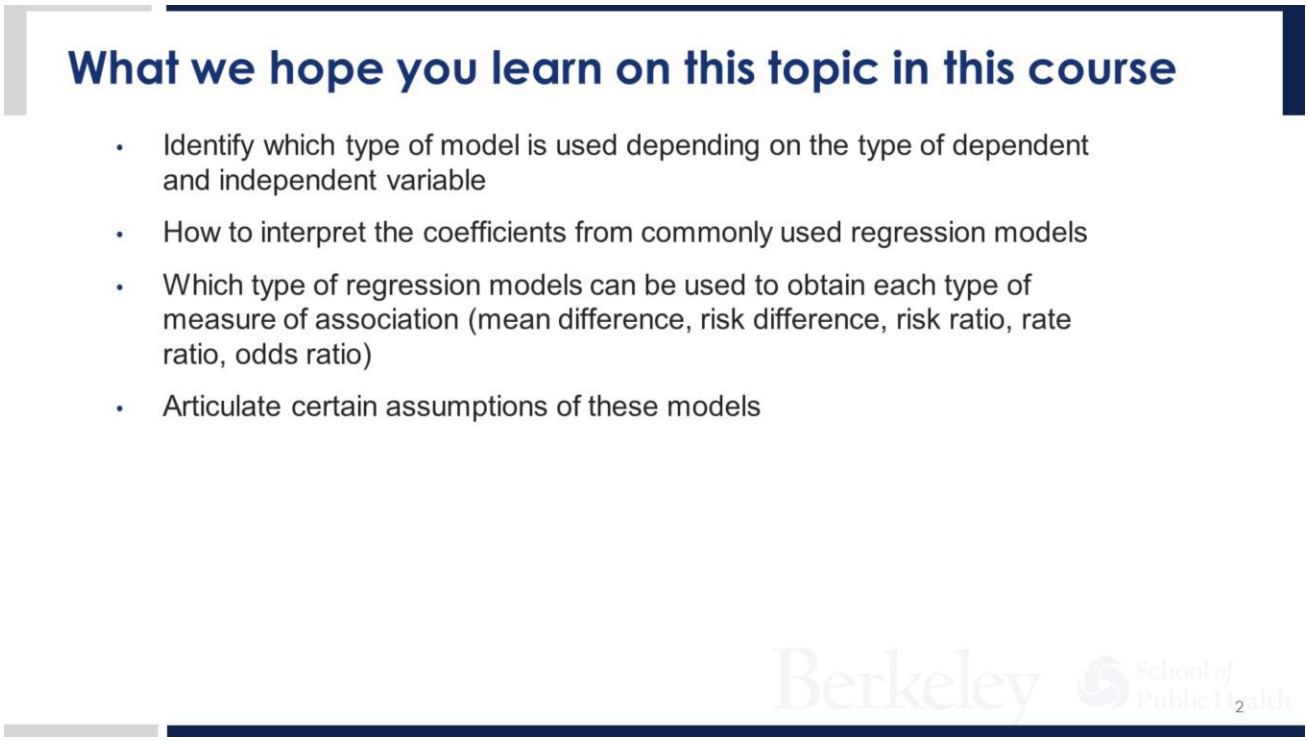
Next, we're turning to multivariable analysis. And this video focuses on a certain subset of these analysis called linear regression or multivariable linear regression. So that's the main emphasis of this video.

And then, we're also going to briefly introduce modeling approaches for longitudinal data and repeated measures data. And there is entire courses on that topic. Our goal today is just to briefly introduce you to what you need to know when thinking about which models to choose. And then if this is something that you need to do in your work, definitely recommend that you take a course on that subject.



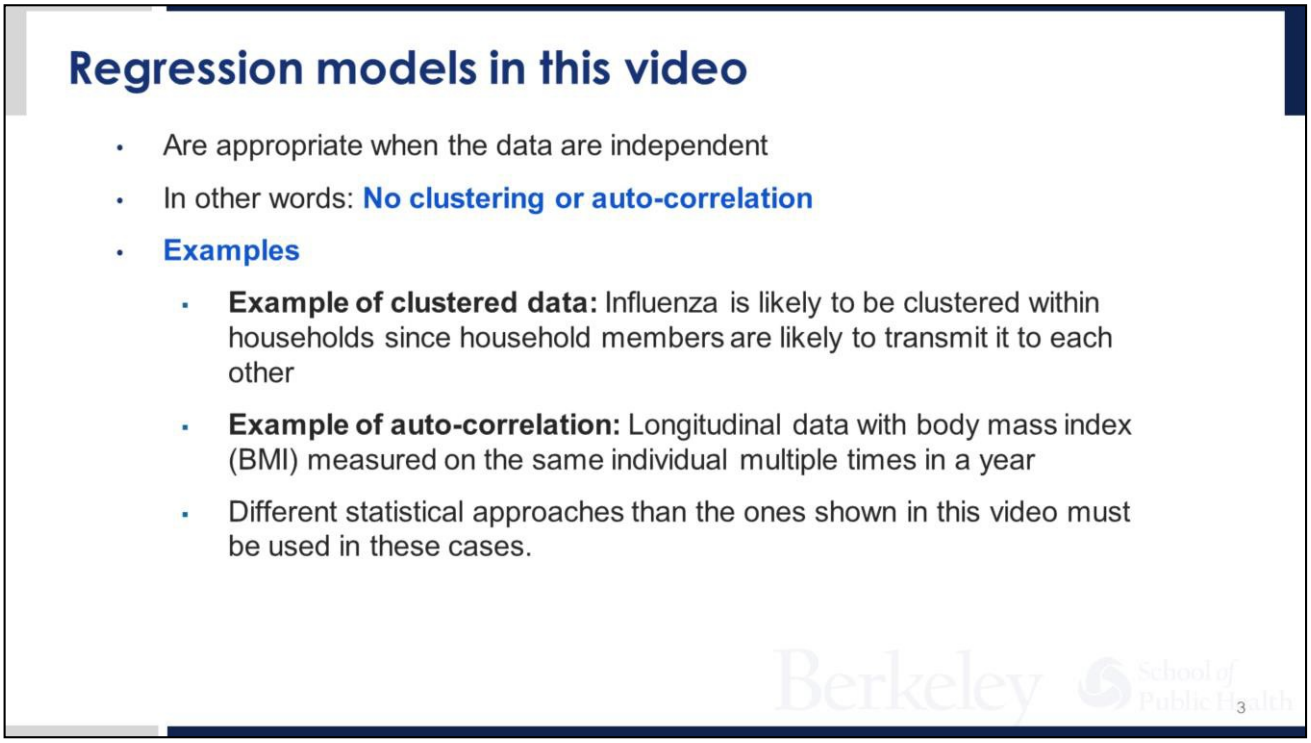
So some of these slides are going to feel quite technical. I do you want to just briefly mention that in this course, we really aren't teaching you linear regression or any kind of regression from a technically statistical standpoint. However, it may still feel quite notation heavy compared to other parts of this course. And so what we'd really like you to take away is the ability to identify which type of model is used depending on the type of dependent or independent variable your study has measured.

How to interpret coefficients from these commonly used regression models. You're going to learn what a coefficient is in this video. How to choose which type of regression model can be used to obtain a certain type of measure of association, such as a mean difference or risk difference, et cetera.



And then, we'd like you to be able to articulate key assumptions of these models. These models make many more assumptions than we're going to cover in these videos. And that's because some of these assumptions get into pretty technical statistical territory. Our emphasis, in terms of teaching you assumptions, is merely to help you understand when these models are just completely not appropriate for certain types of data.

So again, this is what we're hoping you take away from these videos in this series. And you may feel that at times, it's more technical than this and that's really to help you understand where some of these formulas are coming from. But we don't necessarily expect you to be able to derive things-- for example-- in an exam setting in this course. However, we do feel that it's helpful for you to have seen it before so that's our general approach to teaching this topic.

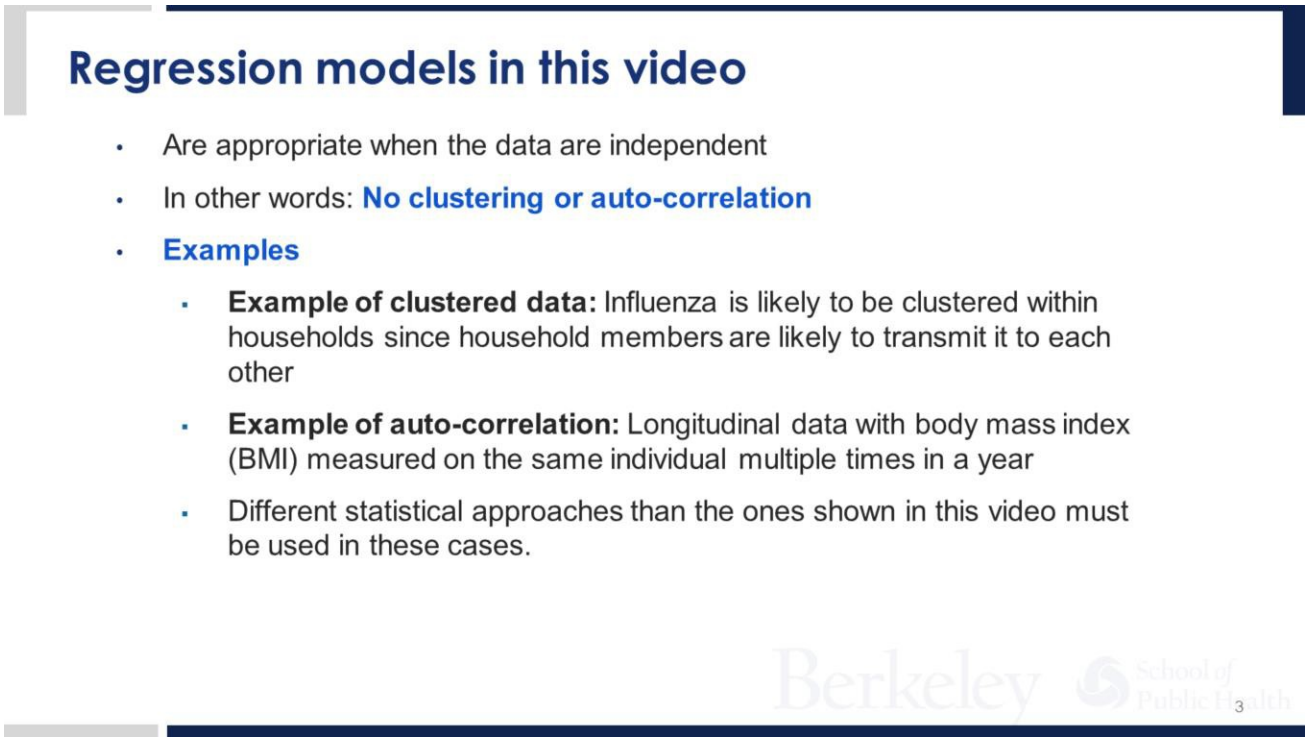


In this video, the regression models I'm going to present to you make a key assumption. They assume that data are independent. In other words, that there is no clustering or auto-correlation. So what do we mean by those words?

Well, clustered data is data that is collected among groups of individuals who share certain characteristics. And as a result, outcomes of people in the same group are more likely to be similar to each other than to outcomes of people in different groups. And so an example would be influenza. Influenza is highly clustered within households.

If you've ever had it yourself-- I personally had a few Christmases ago when I was with family. One person got sick first and because we were spending a lot of time together indoors in the winter, our transmission rate was quite high within our household. So the risk of influenza is well-known to be clustered within households because of this. Because it happens to be circulated in the winter in the US. And influenza is transmitted through a respiratory pathway.

And so that's an example of clustered data. If we had data on individual information related to influenza illness from groups of households, we would expect the people in the same household to have a similar risk of influenza. And that risk to differ more when we consider different households than when we consider different people within the same household.



The next assumption is related to auto-correlation. And this comes up when we think about longitudinal data. The idea is that certain data structures are correlated within themselves over time, most frequently in epidemiology.

And so an example of this would be longitudinal data that measures body mass index-- BMI. And if we measure this on the same person multiple times in a year-- let's say, monthly-- the BMI of a person in January is likely to be correlated with their BMI in February. And to their BMI in February is likely to be correlated with their BMI in March. And the correlation is strongest in the measurements closest to each other. And that's because just from a biologic standpoint, people can't change their body mass index that quickly so we wouldn't expect to see massive random fluctuations in body mass index.

This type of data can't be considered independent, nor can clustered data. If we are using this type of data in our analysis-- which is actually quite common in epidemiology-- we need to use different statistical approaches than the ones shown in this video. So that's really the take home message here is that you should always think, are my data independent? Do I have any clustering? Do I have any time series or longitudinal data?

If the answer is yes, the models that I'm going to show you in this video are not necessarily appropriate. And so you should proceed with caution.

Summary of multivariable analyses

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method	Measure of association estimated
Continuous	>1	Continuous or binary	Multivariable	Linear regression	Mean difference
Binary or count	>1	Continuous or binary	Multivariable	Log-linear regression	Risk ratio or rate ratio
Binary	>1	Continuous or binary	Multivariable	Logistic regression	Odds ratio

Note: multivariable regression is also referred to as multiple regression

Berkeley School of Public Health

So we're building this table. We had a similar looking table for our univariable and bivariable analysis and now, we're extending it for multivariable analysis. And we're actually adding even more than these three rows this week to this table.

So if we look at the left hand column, we have our type of outcome variable for a continuous binary or count and binary outcomes. Because these are all multivariable analysis, the second column for the number of exposure and other variables is greater than one because we're considering an exposure plus potential confounders or other variables.

For all of the analysis we're going to talk about in this unit, the exposure variable could be continuous or binary. It also, technically, could be categorical, and we'll come to how to code that briefly, in a few moments.

Summary of multivariable analyses

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method	Measure of association estimated
Continuous	>1	Continuous or binary	Multivariable	Linear regression	Mean difference
Binary or count	>1	Continuous or binary	Multivariable	Log-linear regression	Risk ratio or rate ratio
Binary	>1	Continuous or binary	Multivariable	Logistic regression	Odds ratio

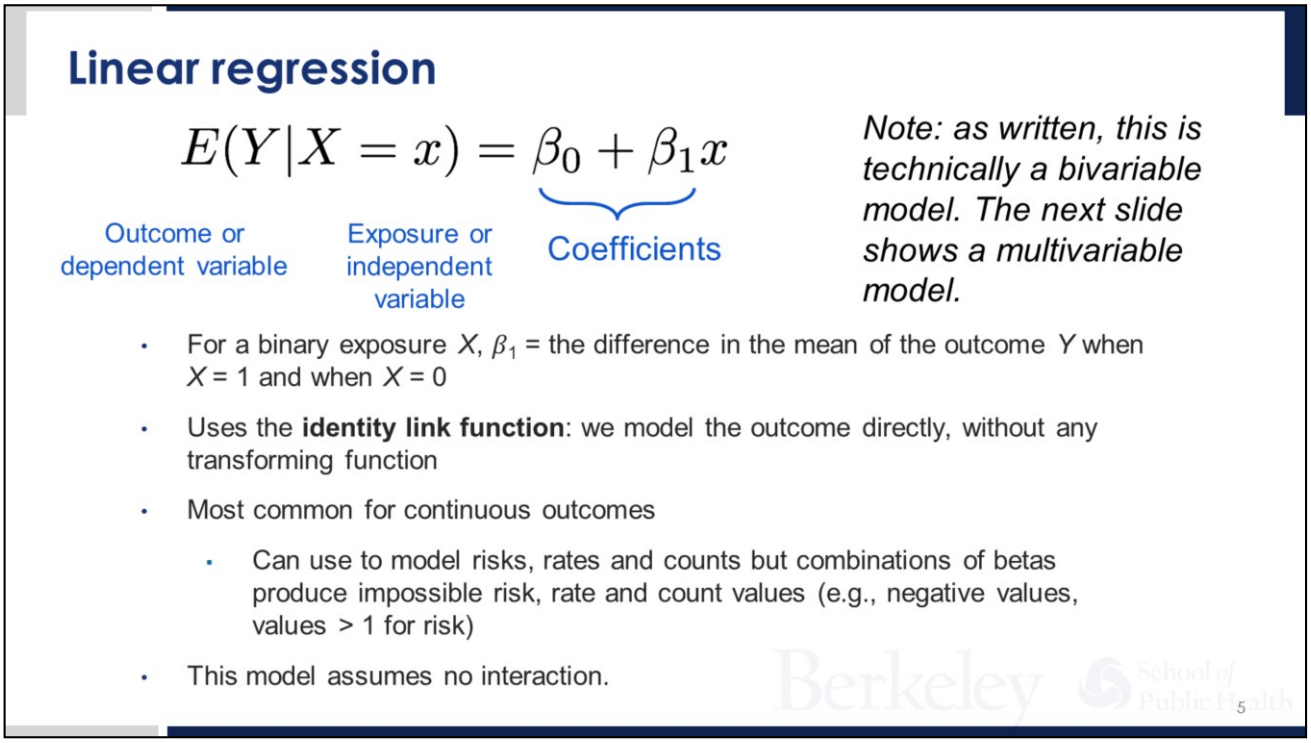
Note: multivariable regression is also referred to as multiple regression

Berkeley School of Public Health

And then, there's these three different types of analysis methods-- linear regression, log linear regression, and logistic regression. And these estimate different measures of association and they're linked to different types of outcome variables. So this video is going to focus on linear regression because it's the most straightforward, easy to understand.

We actually don't talk very much about continuous outcomes in this course. When we covered measures of association, we didn't really talk about mean differences, but they are an important measure of association for continuous outcomes. And that's really when we want to use a linear regression model.

So we'll go into quite a bit of depth on this topic. And then the next video-- or two videos from now, rather, we'll focus on log linear regression and logistic regression, which are appropriate for binary or count outcomes. And there's a lot of similarities between these models, but also, a few key differences. So we'll go into the most depth on linear models.



Let me walk you through this notation. Here, this capital E stands for expectation. It's sort of similar to thinking about a mean. And the Y stands for our outcome of interest. The big X is our exposure variable. And when we use a capital letter, it means it's technically a random variable, which means that it could take on different values. We're not specifying if the value is zero or one when we use a capital.

And the lower case means that we are specifying a specific value. But when we use a lowercase x, we haven't said exactly what it could be. And so this first part here-- E of Y conditional on X equals x-- this is basically saying that we're modeling the mean outcome conditional on our covariate x.

And remember that in previous parts of the course, we've used the word condition as a synonym for stratification. We can think of regression models as an extension of stratification methods that we've learned about earlier in this course. But they actually go further than that, and I'll show you what that looks like in a few slides. So that's the first part of this equation.

The second part of the equation here, we have beta sub 0 plus beta sub 1 little x. The betas here are our regression coefficients. And I'm going to show you how this formula works in a few slides, but for now, what you need to know is that for a binary exposure x-- so that's an exposure that is equal to 0 or 1, such as smoker, x equals 1. Non-smoker, x equals 0.

Linear regression

$$E(Y|X = x) = \underbrace{\beta_0 + \beta_1 x}_{\text{Coefficients}}$$

Outcome or dependent variable

Exposure or independent variable

Note: as written, this is technically a bivariable model. The next slide shows a multivariable model.

- For a binary exposure X, β_1 = the difference in the mean of the outcome Y when $X = 1$ and when $X = 0$
- Uses the **identity link function**: we model the outcome directly, without any transforming function
- Most common for continuous outcomes
  - Can use to model risks, rates and counts but combinations of betas produce impossible risk, rate and count values (e.g., negative values, values > 1 for risk)
- This model assumes no interaction.

Berkeley School of Public Health

Beta 1 is the difference in the mean outcome Y when x equals 1 and when x equals 0. So you're basically saying take the mean of Y when x is 1. And then take the mean of Y when x is 0. And take the difference if those two means. So our measure of association here is a mean difference.

So we've defined beta 1. But what dies beta 0 mean? So if x is equal to 0-- and in that example I just gave, I said smoking is indicated by x equals 1. And non-smoking is indicated by x equals 0. So if someone's a non-smoker, what happens is their value of x is equal to 0. And when we multiply 0 times beta 1, that term is no longer present because it's equal to 0.

And so when we have a non-smoker, our Beta-1 is equal to 0. And so this beta 0 here is called the intercept. I'm going to show you how this is graphed little bitlater in this video. And beta 0 is the mean outcome among those who are unexposedor among those with x equals to 0.

Now, to introduce some terminology to you-- this type of model uses an identity link function. It means that we model the outcome directly and we don't transform it. Now, that's going to seem kind of vague to you and the purpose of the link function will become much clearer when we talk about log linear and logistic models. So just keep that phrase in your mind-- identity link function-- for now and I think it will become clearer in a little bit.

Linear regression

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

Outcome or dependent variable

Exposure or independent variable

Coefficients

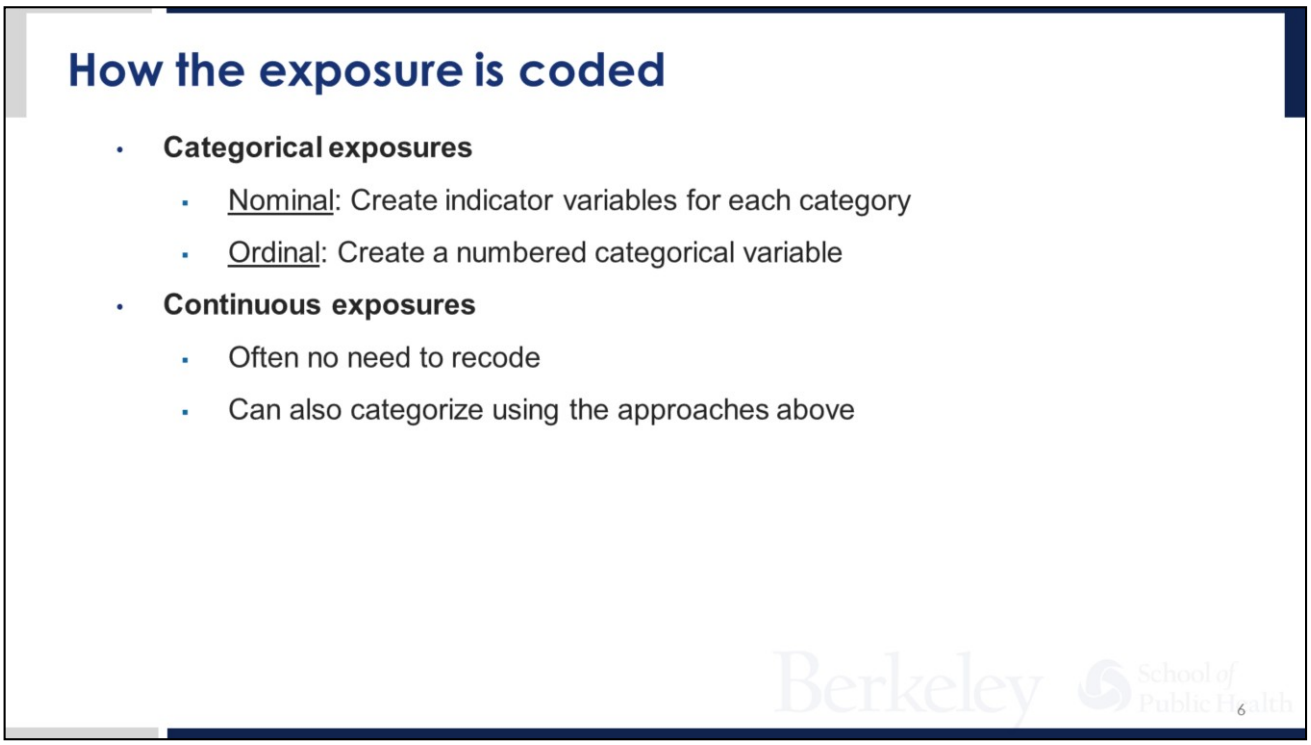
Note: as written, this is technically a bivariable model. The next slide shows a multivariable model.

- For a binary exposure X, β_1 = the difference in the mean of the outcome Y when $X = 1$ and when $X = 0$
- Uses the **identity link function**: we model the outcome directly, without any transforming function
- Most common for continuous outcomes
  - Can use to model risks, rates and counts but combinations of betas produce impossible risk, rate and count values (e.g., negative values, values > 1 for risk)
- This model assumes no interaction.

Berkeley School of Public Health

So we most commonly use linear regression models for continuous outcomes, but we can also use these models to estimate associations for risks, and rates, and counts. And so that would give us a difference in risk or a difference in rates, instead of a difference in the mean. Unfortunately, this doesn't always work from a statistical standpoint because certain combinations of the betas or the coefficients can produce values that are impossible for risks or rates. So risk is bounded from 0 to 1 and linear regression models don't force the estimate of the risk difference to actually be bounded in a plausible way.

So for example-- we could end up with an estimated risk value that exceeds one, which doesn't make any sense. As a result, we typically use different modeling approaches for risks, rates, and counts. But sometimes, it's still OK to use a linear regression model. In this course, we're really going to just emphasize the use of these models for continuous outcomes.
It's also important to mention that the way this model is written at the top of the slide assumes no interaction. We have a whole separate video on what these models look like when we think that interaction may have occurred. But if we have a model like this, then as it's currently written, it's making an assumption that there was no interaction between different variables in our dataset.
I want to also just briefly point out here that this, as written, is technically a bivariable model because it only has X and Y. In a few slides, we'll show you a multivariable model.



Before I go into detail about how these models work, I want to talk about how exposures are coded. If we have a categorical exposure that's nominal-- meaning the different levels don't have a particular ordering to them, such as race categories-- we can create something called indicator variables for each category.

If we have an ordinal categorical variable that does have some numerical ordering, we can create a numbered categorical variable. If we have a continuous exposure we actually don't need to recode it, but we can also choose to categorize it using these approaches defined above. So it's sort of up to our choice when we're designing our model.

How indicator variables are created

- **Categorical exposure example:**
  - City of residence

	X1	X2	X3	X4	X5
Original value					
Berkeley (reference)	0	0	0	0	0
Oakland	1	0	0	0	0
Albany	0	1	0	0	0
Piedmont	0	0	1	0	0
Emeryville	0	0	0	1	0
Kensington	0	0	0	0	1

Berkeley School of Public Health

And let's go over an example of how a nominal categorical variable might be coded. So city of residence-- perhaps you want to control for city of residence in our model. And what we need to do is choose one of the levels of the category as our reference group. So this is kind of like thinking about our unexposed group.

Essentially, what we're going to do is we're going to have pair-wise comparisons. So we'll have Oakland compared to Berkeley, if Berkeley is our reference. And then Albany compared to Berkeley. And then Piedmont compared to Berkeley. And so on. In other words, we're taking the mean in Oakland compared to the mean in Berkeley and we'll take the difference between them.

So when we choose our reference, we just want to pick the level that it makes the most sense to compare everything to. And often with a categorical nominal variable, there just isn't an obvious level so we just have to choose one.

How indicator variables are created

- **Categorical exposure example:**
  - City of residence

	X1	X2	X3	X4	X5
Original value					
Berkeley (reference)	0	0	0	0	0
Oakland	1	0	0	0	0
Albany	0	1	0	0	0
Piedmont	0	0	1	0	0
Emeryville	0	0	0	1	0
Kensington	0	0	0	0	1

Berkeley School of Public Health

And then, here is how the coding would work. So take a look at these columns X1 through X5. These will each become a variable or a column in the dataset that we feed into our regression model. And for X1, we would code a one for Oakland and a zero for everything else. For X2, we'd code a one for Albany and a zero for everything else and so on.

And so as you'll see here, the reference row is always equal to zero. So Berkeley never gets coded as a one and that's how essentially, the regression model knows to treat it as the reference level. And if all the x's are 0, the mean outcome for Berkeley is then the Beta-0, the intercept. When we say an indicator variable, this recoding process is what we mean.

And the reason we have to do this is that a regression model wouldn't understand the words Berkeley, Oakland, Albany. So all the data that we feed into a model, it needs to be numerical. And so this allows us to basically translate this categorical nominal information into numeric variables.

How indicator variables are created

Ordinal exposure example:

- Highest education level measured with three values:

Original value	Value in variable used in regression model
Less than high school	1
Completed high school	2
More than high school	3

Berkeley School of Public Health

If we have an ordinal exposure that's categorical- such as the highest education level a person has achieved-- we could code it like this. So we could have a single column that is coded as a one, if the participant completed less than high school. Coded as two, if they completed high school. And coded as three, if they completed more than high school.

Obtaining the mean difference from linear regression coefficients

1. Start with the linear model specification

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

Berkeley School of Public Health

Now that we've talked about coding, let me explain how we obtain this mean difference using a combination of different linear regression coefficients. And so the first step is we start with our linear model specification that I showed you a few slides back. So we have the mean of our outcome or the expectation of our outcome Y, conditional on covariate X and that's equal to beta 0 plus beta 1 little x. So our goal is to estimate the mean difference when comparing X equals 1 to X equals 0

Obtaining the mean difference from linear regression coefficients

1. Start with the linear model specification

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

2. We use the notation from the model to write the formula for the mean difference.

$$E(Y|X = 1) - E(Y|X = 0)$$

Berkeley School of Public Health

And so the next step is to use the notation from above and to fill in the values X equals 1 an X equals 0. So what we want is the mean outcome of Y conditional on X equals 1 minus the mean out of Y conditional on X equals 0.

Obtaining the mean difference from linear regression coefficients

1. Start with the linear model specification

$$E(Y|X=x) = \beta_0 + \beta_1 x$$

2. We use the notation from the model to write the formula for the mean difference.

$$E(Y|X=1) - E(Y|X=0)$$

3. Next we fill in the coefficients for each term in the difference.

$$E(Y|X=1) - E(Y|X=0) = \big(\beta_0 + \beta_1 \cdot (1)\big) - \big(\beta_0 + \beta_1 \cdot (0)\big)$$

Berkeley School of Public Health

So in step 3, we're going to fill in these portions of the equation using the coefficients at the top of the slide. So for the first term-- the mean of Y conditional on X equals 1, we take beta 0 plus beta 1 and then we fill in 1 for X here. And then we do the same for the second term. So we say minus beta 0 plus beta 1 times and then we fill in 0 for the value of X.

Obtaining the mean difference from linear regression coefficients

1. Start with the linear model specification

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

2. We use the notation from the model to write the formula for the mean difference.

$$E(Y|X = 1) - E(Y|X = 0)$$

3. Next we fill in the coefficients for each term in the difference.

$$E(Y|X = 1) - E(Y|X = 0) = \big(\beta_0 + \beta_1 \cdot (1)\big) - \big(\beta_0 + \beta_1 \cdot (0)\big)$$

4. After simplifying, we see that β_1 is equal to the mean difference.

$$\beta_1 = E(Y|X = 1) - E(Y|X = 0)$$

Berkeley School of Public Health

So after we simplify, what ends up happening is that beta 1 times 0 term drops out so that's no longer present. And then the beta 0s also are canceled. And that leaves us with beta 1 is equal to the mean of Y conditional on X equals 1 and the mean of Y conditional on X equals 0. And so that's the way that we obtain our mean difference.

What this means in practice is we're going to run a statistical model. And we usually do that in a statistical software program, such as R or Stata. And those programs are going to give us back our beta coefficients. And so the beta that's attached to the variable X is equal to the mean difference when X equals 1 compared to when X equals 0. And now, that's if we have a binary exposure. If X is a continuous exposure, then beta 1 is the mean difference for a one unit change in the value of X.

Example of multivariable linear regression

- **Outcome**: "height-for-age" (continuous)
  - Z-score with values ranging from -6 to 6
- **Exposure**: "sanitation" (binary)
  - Indicator variable with the values:
    - 0 = unimproved sanitation
    - 1 = improved sanitation
- **Confounder**: "wealth" (binary)
  - Indicator variable with the values:
    - 0 = below median household wealth
    - 1 = above median household wealth

Berkeley School of Public Health

13

So let's go over an example. Our outcome variable-- in this example-- is the height-for-age score and this is a continuous value that ranges typically, from negative six to six. And it's essentially, taking a child's height or length at a specific age-- usually when the children are quite young-- and then standardizing their height against what's considered healthy for their exact age and sex.

And that gives us a Z-score. And if the Z-score is below negative two, the child is considered to be stunted, which is short for their age.

Then our exposure is sanitation. So let's say we define this as a binary variable where one is equal to improve sanitation and zero is equal to unimproved sanitation. And so, this might mean an improved latrine versus an unimproved latrine-- for example.

And then, let's say we have one confounder we want to adjust for-- his wealth. Because the household's wealth is likely to affect their quality of sanitation and it's also likely to affect the child's growth. And we code this as a binary variable where we have some continuous measure of wealth. It could actually be their household income or it could be some other index of their assets.

And we code this variable as zero, if the household has a wealth level below the median. And one, if it's above the median.

Example of multivariable linear regression

$$E(\text{Height-for-age}|\text{Sanitation, Wealth}) = \beta_0 + \beta_1 \text{Sanitation} + \beta_2 \text{Wealth}$$

- The independent variables in this model are Sanitation and Wealth.
- These are often referred to as **covariates**, a term which includes the exposure / intervention variable and potential confounders and other variables that are adjusted for.

Berkeley School of Public Health

Here's how we can specify the model in this example. We want to estimate the mean height-for-age Z-score conditional on sanitation and wealth. And this model is equal to beta 0 plus beta 1 times an indicator for improved sanitation plus beta 2 times an indicator for above median wealth. In this model we call sanitation and wealth independent variables and height-for-age the score is the dependent variable.

Covariates Is another term that people use to describe the variables sanitation and wealth. Covariates can include both an exposure variable, as well as potential confounders. Or other types of variables you might choose to adjust for. So it's a handy term because it basically means all the variables on the right hand side of the equation.

Example of multivariable linear regression

$$E(\text{Height-for-age}|\text{Sanitation, Wealth}) = \beta_0 + \beta_1 \text{Sanitation} + \beta_2 \text{Wealth}$$

- β_0 = Mean height-for-age z-score among those without improved sanitation and with below median household income

$$E(\text{Height-for-age}|\text{Sanitation} = 0, \text{Wealth} = 0) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0) = \beta_0$$

Berkeley School of Public Health

Now, we're going to work through the formula to understand what the coefficients mean in this example. So beta 0 is the mean height-for-age Z-score among those without improved sanitation and below median household income. to Show this, what we do is we plug-in the values for sanitation and wealth into this formula. We have our mean height-for-age Z-score conditional on sanitation equaling 0 and wealth equaling 0.

So it is beta 0 plus beta 1 times 0 since sanitation equals 0. Plus beta 2 times 0 since wealth equals 0. And the beta 1 beta 2 times cancel out and that leaves us with beta 0-- the mean difference in height-for-age Z-score among those without improved sanitation and with below median income. Since these are reference groups for these binary covariates, we can think of this as our baseline level of height-for-age Z-score.

Example of multivariable linear regression

$$E(\text{Height-for-age}|\text{Sanitation, Wealth}) = \beta_0 + \beta_1 \text{Sanitation} + \beta_2 \text{Wealth}$$

- β_0 = Mean height-for-age z-score among those without improved sanitation and with below median household income

$$E(\text{Height-for-age}|\text{Sanitation} = 0, \text{Wealth} = 0) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0) = \beta_0$$

- β_1 = Difference in mean height-for-age z-score among people with vs. without improved sanitation with below median household income

$$E(\text{Height-for-age}|\text{Sanitation} = 1, \text{Wealth} = 0) - E(\text{Height-for-age}|\text{Sanitation} = 0, \text{Wealth} = 0) =$$
$$\big(\beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (0)\big) - \big(\beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0)\big) = \beta_1$$

Berkeley School of Public Health

And then, beta 1 is the difference in the mean height-for-age Z-score among people with versus without improved sanitation and with below median household income. So we can do the same thing where we plug-in these values into our formula.

This is how we can write this. The mean height-for-age Z-score conditional on sanitation equaling 1 and wealth equaling 0 minus the mean height-for-age Z-score if sanitation is 0 and wealth is 0.

If we first plug-in the values for this half of the formula here, that's beta 0 plus beta 1 times 1 since sanitation equals 1. Plus beta 2 times 0 since wealth equals 0. Minus the sum of beta 0 plus beta 1 times 0 since sanitation equals 0 in the second half. Plus beta 2 time 0 since wealth equals 0 in the second half.

Now, the beta 0s cancel. Beta 2s are equal to 0 since we multiply each of them time 0 so those cancel. And then the second beta 1 times 0 also cancels, and that leaves us with just beta 1. That's our difference in the mean height-for-age Z-score among people with versus without an improved latrine with below median household income.

Example of multivariable linear regression

$$E(\text{Height-for-age}|\text{Sanitation, Wealth}) = \beta_0 + \beta_1 \text{Sanitation} + \beta_2 \text{Wealth}$$

- β_0 = Mean height-for-age z-score among those without improved sanitation and with below median household income

$$E(\text{Height-for-age}|\text{Sanitation} = 0, \text{Wealth} = 0) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0) = \beta_0$$

- β_1 = Difference in mean height-for-age z-score among people with vs. without improved sanitation with below median household income

$$E(\text{Height-for-age}|\text{Sanitation} = 1, \text{Wealth} = 0) - E(\text{Height-for-age}|\text{Sanitation} = 0, \text{Wealth} = 0) =$$
$$\big(\beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (0)\big) - \big(\beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0)\big) = \beta_1$$

- β_2 = Difference in mean height-for-age z-score among people in below vs. above median household income categories among those without improved sanitation

$$E(\text{Height-for-age}|\text{Sanitation} = 0, \text{Wealth} = 1) - E(\text{Height-for-age}|\text{Sanitation} = 0, \text{Wealth} = 0) =$$
$$\big(\beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (1)\big) - \big(\beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0)\big) = \beta_2$$

Berkeley School of Public Health

17

And beta 2 is very similar to this, but with our exposure variables flipped. So beta 2 is the difference in mean height-for-age Z-score among people in below versus above median household income categories among those without improved sanitation.

And so here, we have our two parts of the formula-- the mean height-for-age Z-score conditional on sanitation equaling 0 and wealth equaling 1. Minus the mean height-for-age Z-score if sanitation and wealth are both equal to 0. We plug-in the values for sanitation and wealth-- as we just did in the beta 1 example.

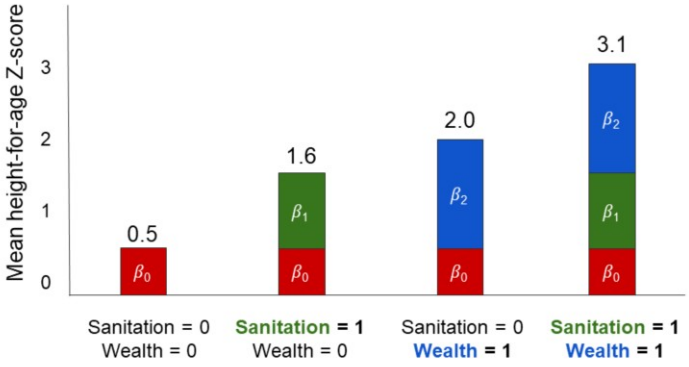
And here's the formula that we get. What we see is that, again, the beta 0s cancel. Beta 1s are both multiplied by 0 so those cancel. And then the second beta 2 is multiplied by 0 and so that leaves us with just beta 2. The difference in the mean height-for-age Z-score below and above median household income categories and without improved sanitation.

Example of multivariable linear regression

$$E(\text{Height-for-age}|\text{Sanitation, Wealth}) = \beta_0 + \beta_1 \text{Sanitation} + \beta_2 \text{Wealth}$$

$$E(\text{Height-for-age}|\text{Sanitation, Wealth}) = 0.5 + 1.1 \cdot \text{Sanitation} + 1.5 \cdot \text{Household wealth}$$

- Mean difference for **sanitation** alone: 1.6 - 0.5 = 1.1
- Mean difference for above median **wealth** alone: 2.0 - 0.5 = 1.5
- Mean difference for **sanitation** + above median **wealth**: (1.5 + 1.1 + 0.5) - 0.5 = 2.6



Berkeley School of Public Health

18

Now, let's look at what this looks like when we actually have some estimates for our betas-- our coefficients. And again, you would usually obtain these by using statistical software, but here, I'm just giving you the values for the purpose of teaching you these concepts.

So here's our same model. And now, beta 0 is equal to 5. Beta 1 is equal to 1.1. And beta 2 is equal to 1.5. This graph is meant to illustrate how these different betas fit together when we're thinking about the mean height-for-age Z-score under different combinations of covariates.

So in the y-axis, we have the mean height-for-age Z-score. Our first bar here is the mean HAZ-- that's shorthand for a height-for-age Z-score-- when sanitation is equal to 0 and wealth is equal to 0. And so if we quickly just plug this into the equation on the top, we plug-in 0 for sanitation and 0 for household wealth-- that knocks out those last two terms. And we see that the mean is 0.5,

So beta 0 is equal to 0.5-- as shown in this red bar.

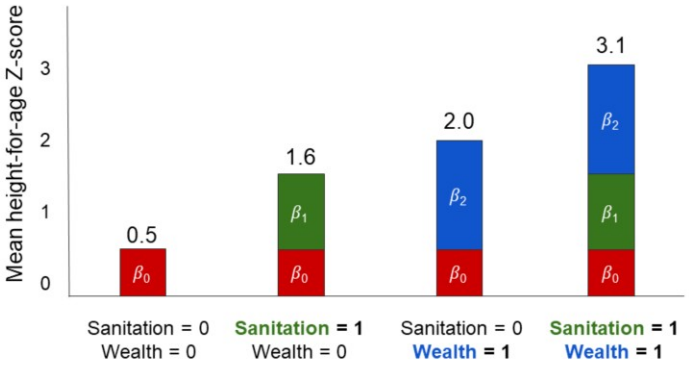
Now, what about the mean HAZ when sanitation is equal to 1 and wealth is equal to 0? So if we plug-in these values into our formula above, we see that the wealth term cancels out. And we have a sum of 0.5 plus 1.1 and that gives us 1.6. So that's the mean difference for sanitation on its own among people with below median household wealth.

Example of multivariable linear regression

$$E(\text{Height-for-age}|\text{Sanitation, Wealth}) = \beta_0 + \beta_1 \text{Sanitation} + \beta_2 \text{Wealth}$$

$$E(\text{Height-for-age}|\text{Sanitation, Wealth}) = 0.5 + 1.1 \cdot \text{Sanitation} + 1.5 \cdot \text{Household wealth}$$

- Mean difference for **sanitation** alone: 1.6 - 0.5 = 1.1
- Mean difference for above median **wealth** alone: 2.0 - 0.5 = 1.5
- Mean difference for **sanitation** + above median **wealth**: (1.5 + 1.1 + 0.5) - 0.5 = 2.6



Berkeley School of Public Health

And then, the bars here are stacked to show you that 1.6 is the sum of the coefficients B0 plus B1-- beta 0 plus beta 1.

For those with unimproved sanitation and household wealth above median, we sum beta 0 plus beta 2. Because when we plug-in sanitation as 0 and wealth equals 1, the sanitation term drops out. And so as these bars show here-- the sum of beta 0 plus beta 2 is 2.0.

And then if you want to know the mean HAZ, for those with both improve sanitation and above median wealth-- we can sum up all three betas together. Beta 0 plus beta 1 plus beta 2 and that gives us a mean height-for-age Z-score of 3.1.

So the bar graph is showing us how these coefficients can stack or sum together in order to obtain the mean height-for-age Z-score. And if we want to obtain mean differences, we can use the group who has unimproved sanitation and below median household wealth as our reference group or our baseline group.

And so for the mean difference for sanitation alone-- as shown in this bullet point here on the left-- that's 1.6-- as we see in this red and green stacked bar. Minus 0.5 from our reference level of mean HAZ. If you want to get the mean difference for above median wealth alone, we take 2.0-- as shown by the red and blue bar-- minus 0.5 from our baseline bar, and that gives us 1.5.

And if we want the mean difference for both sanitation and wealth compared to without those two, we take the sum of these three bars-- red, green, and blue-- which is 3.1 and we subtract the red bar, 0.5, and that gives us a difference of 2.6.

Now, all of these sanitation and wealth covariates have been binary so far. How does this work if we have a continuous exposure or a continuous covariate?

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- Continuous independent variables are interpreted as the association of the dependent variable (the outcome) with a one-unit change in the independent variable.
- **Example:** What is the difference in mean height-for-age (HAZ) Z-score among those without improved sanitation (Sanitation = 0) **for a one-kilogram change in weight from 0 kg to 1 kg?**
  - Mean difference in HAZ = (-2.5 + 1.1 (0) + 0.05 (1)) - (-2.5 + 1.1 (0) + 0.05 (0)) = 0.05

Berkeley School of Public Health

So I've changed the model here. Now, we're not looking at sanitation and wealth. We're looking at sanitation and weight. So we have the mean height-for-age Z-score conditional on sanitation and weight. And it's equal to minus 2.5 plus 1.1 times sanitation plus 0.05 times weight.

So if we want to estimate a mean difference with a continuous covariate, such as weight-- let's say, this is measured in kilograms-- it gives us a one unit change in this independent variable. So let's say we were interested in the mean difference in height-for-age Z-score among those without improved sanitation-- so that's sanitation equals 0-- for a 1 kg change in weight from 0 to 1 kg. And so we can plug that in here and that's minus 2.5 plus 1.1 times 0-- since sanitation equals 0-- plus 0.005 times 1 for 1 kg minus the sum of minus 2.5 plus 1.1 times 0 since sanitation is still 0. Plus 0.05 times 0. And so everything cancels out and we get 0.05 and that's the beta coefficient on weight.

Now, let's pause for a second and think about what we're doing here. It's easy to get caught up in the numbers, but does this make any sense? No, it doesn't because a person's weight couldn't be equal to 0 kilograms. So you have to be a little careful about this with these models. It's easy to forget about what is biologically plausible or what's even meaningful.

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- Continuous independent variables are interpreted as the association of the dependent variable (the outcome) with a one-unit change in the independent variable.
- **Example:** What is the difference in mean height-for-age (HAZ) Z-score among those without improved sanitation (Sanitation = 0) **for a one-kilogram change in weight from 0 kg to 1 kg?**
  - Mean difference in HAZ = (-2.5 + 1.1 (0) + 0.05 (1)) - (-2.5 + 1.1 (0) + 0.05 (0)) = 0.05
- … **for a one-kilogram change in weight from 10 kg to 11 kg?**
  - Mean difference in HAZ = (-2.5 + 1.1 (0) + 0.05 (11)) - (-2.5 + 1.1 (0) + 0.05 (10)) = 0.05

Berkeley School of Public Health

Let's take a look at this in a different way. What's the answer if we're looking at a 1 kg change in weight from 10 kilograms to 11 kilograms? Well, we basically take this exact same formula, but we swap in 11 instead of 1 right here. And then 10 instead of 0 right here. And so this is definitely more plausible. At least for a small child, it would be feasible that they could gain weight starting at 10 kilograms and gain weight and get 11 kilograms in weight.

And so, we get the same answer-- interestingly. When we go from 10 to 11 and from 0 to 1, that mean difference is 0.05. This underscores an important point, which is that these linear models are assuming a linear relationship between each covariate and the outcome over the full range of the covariates values.

So whether the person weighs 0 or 1 kilogram or 100 or 200 kilograms, that 1 kilogram change will be assumed to be the same by the model. That doesn't mean that this is true. So just stepping back from a scientific standpoint, we might hypothesize that sanitation and weight could affect a child's height-for-age Z-score differently depending on the person's starting weight or starting age.

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- Continuous independent variables are interpreted as the association of the dependent variable (the outcome) with a one-unit change in the independent variable.
- **Example:** What is the difference in mean height-for-age (HAZ) Z-score among those without improved sanitation (Sanitation = 0) **for a one-kilogram change in weight from 0 kg to 1 kg?**
  - Mean difference in HAZ = (-2.5 + 1.1 (0) + 0.05 (1)) - (-2.5 + 1.1 (0) + 0.05 (0)) = 0.05
- … **for a one-kilogram change in weight from 10 kg to 11 kg?**
  - Mean difference in HAZ = (-2.5 + 1.1 (0) + 0.05 (11)) - (-2.5 + 1.1 (0) + 0.05 (10)) = 0.05

Berkeley School of Public Health

So this is a limitation of a linear regression model. It doesn't always make sense to make this assumption of linearity. But often, we do because it really simplifies things. And there are models who can specify that don't make this assumption, but they go well beyond the scope of this class.

But it's just important for you to walk away realizing that we need to think carefully about what values we're plugging into our model. In other words, in this example, we don't really recommend you think about a 0 to 1 kilogram shift, but rather something more realistic, like a 10 to 11 kilogram shift. It'll give you the same answer, but it's still good to be thoughtful about what these values mean.

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- The coefficients for some continuous independent variables may not be easily interpretable, so it may be more appropriate to estimate, for example, a 1 SD increase in the continuous variable.
- **Example:** What is the difference in mean height-for-age (HAZ) Z-score among those without improved sanitation (Sanitation = 0) **for a one SD (20 kg) change in weight?**

Berkeley School of Public Health

Another way that this is often handled-- instead of looking at a one unit increase which can often just be so small that it's not particularly meaningful, you could look at a 1 standard deviation increase in that continuous variable. So let's say 1 standard deviation in weight is equal to 20 kg

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- The coefficients for some continuous independent variables may not be easily interpretable, so it may be more appropriate to estimate, for example, a 1 SD increase in the continuous variable.
- **Example:** What is the difference in mean height-for-age (HAZ) Z-score among those without improved sanitation (Sanitation = 0) **for a one SD (20 kg) change in weight?**
  - If we are holding everything else constant (e.g., Sanitation) then we can just multiply the coefficient on weight times the SD.
  - Mean difference in HAZ = (-2.5 + 1.1 (0) + 0.05 (30)) - (-2.5 + 1.1 (0) + 0.05 (10)) = 1
  - More simply: 0.05 * 20 = 1

Berkeley School of Public Health

And we want to estimate the mean difference holding everything else constant. So holding sanitation at 0.

All we need to do is multiply our coefficient of 0.05 times 20. But to show you the math, we can do this with some specific values. So we could say, what's the shift from 10 to 30? And if I repeated this for any other shift of 20, we'd get the exact same answer. So it's minus 2.5 plus 1.1 time 0-- since sanitation equals 0-- plus 0.05 times 30 minus the sum of minus 2.5 plus 1.1 times 0-- again, because sanitation equals 0-- plus 0.05 times 10. And that gives us a value of 1.

Well, where did this 1 come from? If we just take the beta coefficient 0.05 and multiply it times our standard deviation of 20, that gives us 1. The mean difference in HAZ is a unit of one.

This is much more useful than what we looked at on the previous slide-- just going back there for one second. When we look at a change in 0.05 Z-score units, that's just so small that it's not terribly meaningful. But if we're talking about a change in one standard deviation, this translates to a one unit change in the z-score, which is something we can understand because a shift from minus 2 height-for-age Z-score to minus 3 height-for-age Z-score is actually something that we would consider.

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- The coefficients for some continuous independent variables may not be easily interpretable, so it may be more appropriate to estimate, for example, a 1 SD increase in the continuous variable.
- **Example:** What is the difference in mean height-for-age (HAZ) Z-score among those without improved sanitation (Sanitation = 0) **for a one SD (20 kg) change in weight?**
  - If we are holding everything else constant (e.g., Sanitation) then we can just multiply the coefficient on weight times the SD.
  - Mean difference in HAZ = (-2.5 + 1.1 (0) + 0.05 (30)) - (-2.5 + 1.1 (0) + 0.05 (10)) = 1
  - More simply: 0.05 * 20 = 1

Berkeley School of Public Health

A shift from a child whose height classifies them as stunted versus severe stunted-- that's the difference between height-for-age Z-score of minus 2 and height-for-age Z-score of minus 3.

So usually, we will actually want to consider carefully what shift is most relevant and most helpful to our research question. And often, a one unit shift is not what we really find to be most useful. Standard deviation is a great option to consider when you're thinking about this.

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- This illustrates one of the assumptions of linear regression models:
  - There is a **linear relationship** between independent and dependent variables

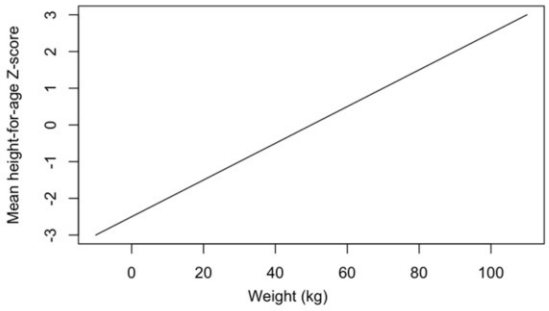
Berkeley School of Public Health

Now, let's look at this linear model, here-- our model for height-for-age Z-score conditioning on sanitation and weight. And let's go back to basic algebra and think about how we can interpret these coefficients within the context of a simple linear model.

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- This illustrates one of the assumptions of linear regression models:
  - There is a **linear relationship** between independent and dependent variables
- Plot of the mean height-for-age Z-score when Sanitation = 0
- y = mx + b
  - y = dependent variable
  - m = slope = 0.05
  - x = independent variable
  - b = intercept = -2.5



Berkeley School of Public Health

24

So we're going to look at the linear relationship between independent and dependent variables. And we're going to focus on a plot of the mean height-for-age Z-score when sanitation equals 0.

So if we plug 0 up at the top of the formula here in for sanitation, it means this middle term is canceled. And so we have minus 2.5 plus 0.05 times weight.

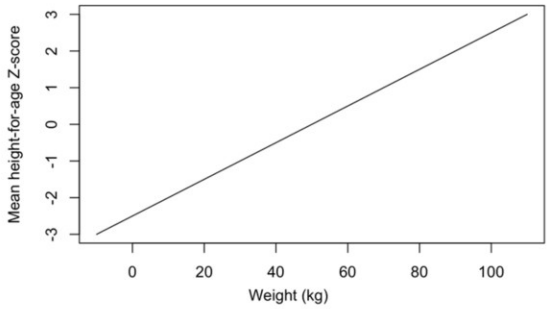
So recall from algebra-- a long time ago-- that our formula for a straight line is y equals mx plus b. y is our dependent variable. x is our independent variable. m is the slope of the line. And b is the intercept. Our formula, once we cancel out the sanitation term, resembles this y equals mx plus b formula, where the intercept is minus 2.5, the slope is 0.05, and x is weight-- our independent variable.

So here on the right is a plot showing us this straight line. So if we take a line and we mark the point where 0 kilograms hits this line and then we draw a horizontal line over to the y-axis, what we see is that when the weight is equal to 0, the y value equals minus 2.5. So notice this line doesn't cross the y-axis formally on this plot.

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- This illustrates one of the assumptions of linear regression models:
  - There is a **linear relationship** between independent and dependent variables
- Plot of the mean height-for-age Z-score when Sanitation = 0
- y = mx + b
  - y = dependent variable
  - m = slope = 0.05
  - x = independent variable
  - b = intercept = -2.5



Berkeley School of Public Health

24

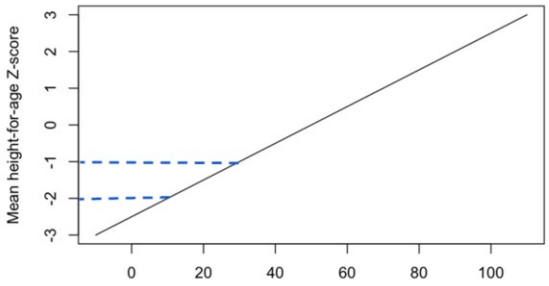
And notice that that's because the weight would have to actually be less than 0 kilograms for this line to intersect the y-axis and that just doesn't make sense. In fact, this portion of the plot here where the line goes beyond 0 to values smaller than 0-- is a classic example of how these kinds of models extrapolate beyond reasonable values in the data. Something we have to really careful about.

So just for illustration purposes, what I'm showing you is that when we plot this line using this formula without really thinking about what the plausible range of the values of the data would be-- when the weight is equal to 0, if we plug 0 in up here at the top, that second term cancels and we have minus 2.5. And so that's where this line meets the y-axis in this plot. And this is why the beta 0 coefficient is also known as the intercept term.

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- This illustrates one of the assumptions of linear regression models:
  - There is a **linear relationship** between independent and dependent variables
- Plot of the mean height-for-age Z-score when Sanitation = 0
- y = mx + b
  - y = dependent variable
  - m = slope = 0.05
  - x = independent variable
  - b = intercept = -2.5



- From last slide: Mean difference in HAZ = (-2.5 + 1.1 (0) + 0.05 (30)) - (-2.5 + 1.1 (0) + 0.05 (10)) = 1
  - = -1
  - = -2

Berkeley School of Public Health

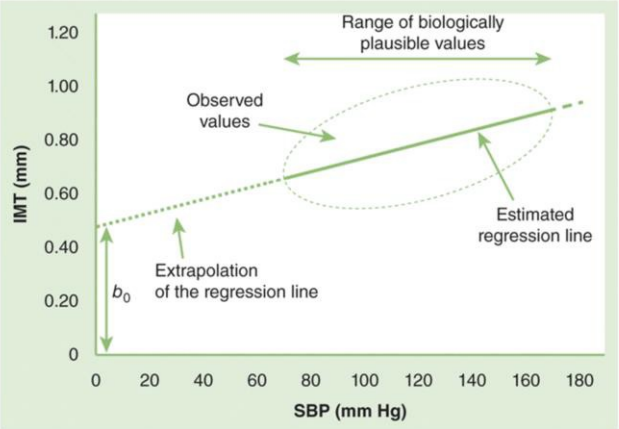
25

Now, to link this back to the previous slide, let's do an example where we look at 20 kilogram shift and we get the main difference in HAZ. So if we take our formula at the top and we plug-in our different terms, we have minus 2.5 plus 1.1 times 0-- since we're looking at this when sanitation equals 0-- plus 0.05 times 30 minus the sum of minus 2.5 plus 1.1 times 0 plus 0.05 times 10. And that equals 1. The first portion of this term is equal to minus 1 and the second portion of this term is equal to minus 2.

Now, let's look at our plot here. So I've drawn in these dashed blue lines to show you that when we plug-in 10 from our second term here-- so 10 kilograms for weight-- this corresponds to a mean height-for-age Z-score value of minus 2. And when I plug-in a value of 30, this corresponds to a mean height-for-age Z-score of minus 1.

Because our slope is constant, it doesn't matter what part of the plot I make this comparison. No matter what the starting kilogram value is, if I make a 20 kilogram change-- whether it be from 10 to 30, or 20 to 40, or 80 to 100 kilograms-- because the slope is still 0.05, we'll always get a difference of 1. So this is a key assumption of a linear model.

Extrapolation in regression models



- Linear regression models assume a linear relationship between X and Y.
- The intercept is an extrapolation of the regression line to the y-axis.
- In some cases, this is beyond the plausible range of values.
- It is best to be very cautious when interpreting regression results extrapolated beyond the range of the observed data.

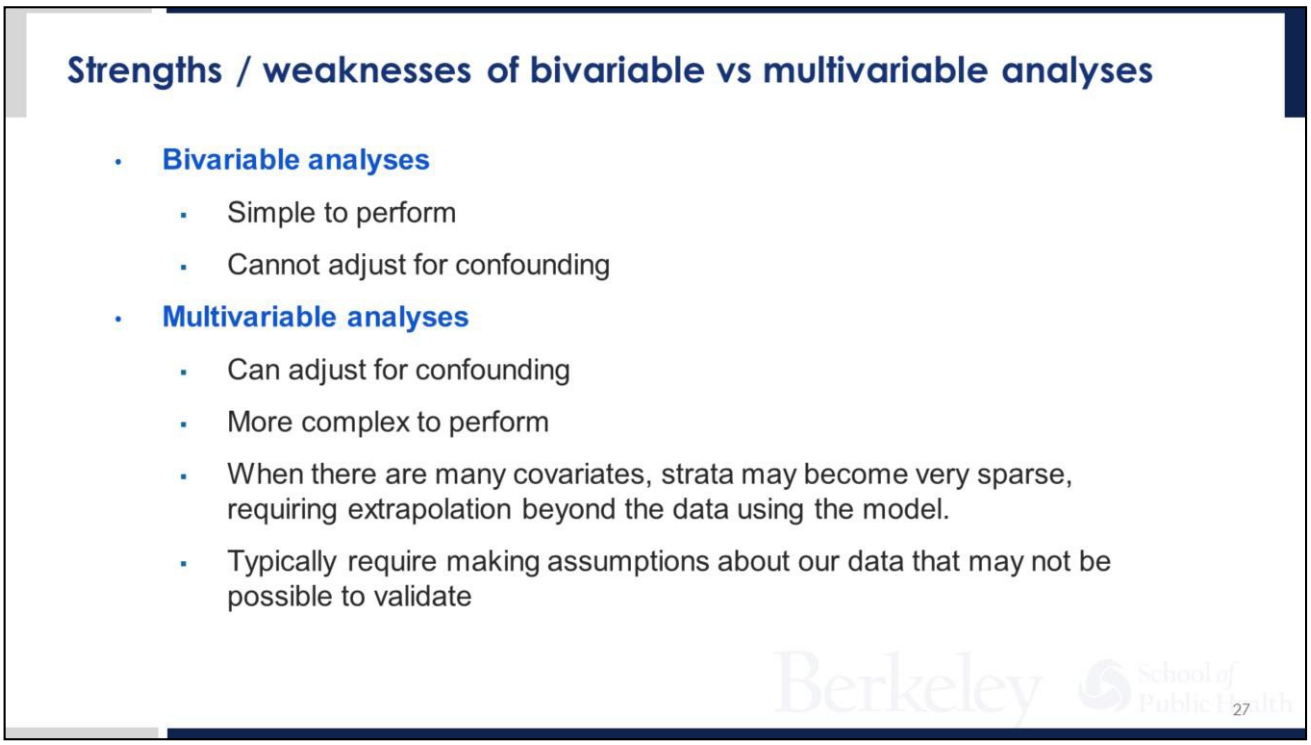
Berkeley School of Public Health

And now, to talk a little more about this issue of extrapolation-- this is a different plot. This is from the Szklo textbook. And what it's showing you with this oval here, is where we would expect to see observed values of the data.

Let's not worry for now what the particular values of the x and y-axis are, but let's just take it as a given that we would only ever expect to see x values between 80 and 160. It's just not biologically plausible to see values below that. However, regression models are just thinking about simple algebra and so they are willing to draw a line that goes through your observed values-- and that's shown here with this darker bolded estimated regression line.

The line will extend beyond the observed values and that's where the line turns dashed. And so, this is the area where the model is extrapolating beyond the observed data. And it will give you an answer so this means that even if you didn't have values of your x-axis variable SVP below 80, you could estimate contrast for the change in y from 20 to 40-- for example.

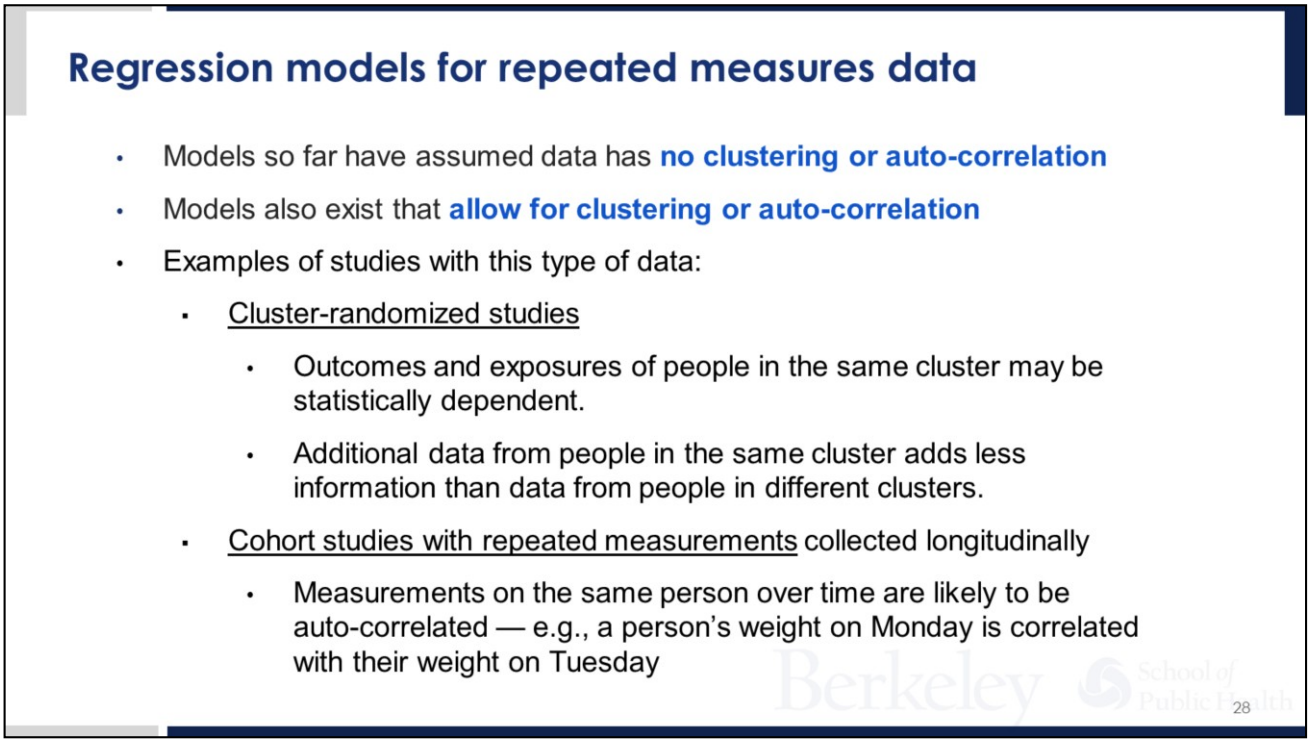
But we wouldn't want to do that because we have no data on that so we don't know if the data actually follows that same pattern. And as a result, any interpretation we make outside of the range of the observed data requires us to fully rely on extrapolation from the regression model. And so, it's best to be very cautious when doing so. In fact, I would advise you to just avoid making any extrapolation beyond the range of the observed data. It just requires us to take a leap of faith that often, is not reasonable to make in the types of research we do in epidemiology.



Now that we've talked in more detail about linear multivariable models, let's talk about the strengths and weaknesses of bivariable versus multivariable analysis. So in the last video, you learned about by bivariable analysis and hopefully, you concluded that they're relatively simple to perform.

But a major limitation of them is that they can't adjust for confounding because you're just looking at the exposure and the outcome together. Multivariable models allow us to adjust for confounding, but generally speaking, they're more complex to perform. When there are many covariates, strata can become very sparse. And what we mean by this is that if we stratify our data based on all the different levels of all the different covariates.

So for example, if we're looking at race, sex, age-- we have to have data within each level of race, age, and sex. And if we don't, our model will extrapolate beyond the data to estimate our measure. And so this is something we want to be very careful about with multivariable analysis. Because if we get into a situation where we're extrapolating beyond the data, we are required to make some assumptions that are really not possible to validate.



Now, so far in this video, we've been talking about a situation where our data has no clustering or auto-correlation. And I just want to briefly introduce you to the names of the models we would use if we do have clustering or auto-correlation.

Here are two examples of types of studies that would have this type of data-- cluster randomized trials inherently involve the collection of clustered data. And outcomes and exposures of people who live in the same cluster are likely to be statistically dependent. That could be because of shared genetics, shared risk factors, shared transmission of disease.

And what this means from a statistical standpoint is that additional data collected from the people in the same cluster will add less new statistical information than additional data from people in different clusters. This really impacts the variance and what it means is that in order to correctly calculate the variance, standard error, and confidence intervals, and p-values, we need to use special methods.

Cohort studies that involve repeated measurements and longitudinal data collection typically have auto-correlation in their data. And I talked earlier about an example with BMI. To talk about someone's weight as another example-- a person's weight on Monday is likely to be strongly correlated with their weight on Tuesday, just because biologically, people's weight can't fluctuate too much in a short period of time. And so for this type of data, we also need to use special statistical models.

Data in a repeated measures study

Data structure for single measurement

id	weight
1	60
2	80
3	55
4	62
5	81

Data structure for repeated measurement

id	time	weight
1	1	60
1	2	62
2	1	80
2	2	76
3	1	55
3	2	54

Berkeley School of Public Health

29

And here's just a brief glimpse at what this data structure looks like. So we can use the term repeated measures study to capture both data that's clustered and also, auto-correlated. And on the left, we have a table with a data structure for single measurements. We can see is that there is a column for ID and a column for a weight measurement. And each ID only appears once in the table.

And then in the table on the right, we have a data structure for repeated measurement. And each ID appears twice so we have time one and time two for each ID. And the weight for each measurement. And what we can see is that the weight isn't too different at the two time points for each individual.

Regression models for repeated measures data

- As is true for data with single measurements per individual, with repeated measures data,the type of outcome and desired measure of association determines whether we use a linear, logistic, or log-linear model.
- Whenever there are repeated measures on individuals, the data has to be analyzed accounting for correlation among observations within individuals
- More advanced courses cover the specific models appropriate for these types of data:
  - Generalized linear models with robust standard errors
  - Mixed models (or "random effects" models)
  - Generalized estimating equations

Berkeley School of Public Health

As is the same for models that look at single measurements that have independence-- with repeated measures data, the type of outcome and the measure of association that we want to estimate is what affects our choice of linear logistic or log linear models. Again, we're coming to logistic and log linear models in the next video.

And whenever we have repeated measures on individuals, we need to use special analytic methods to account for this. Here are the three types of models that it's good for you to just have heard of. We're not going to go into detail, but just so you know their names-- generalized linear models with robust standard errors. Mixed models, which can also be called random effects models. And generalized estimating equations.

So now you've heard those terms. When you see them in the future, it's just good for you to know that these models are useful when you have repeated measures data.

Summary of key points

- **Identify which type of model is used depending on the type of dependent and independent variable**
  - **Continuous outcome**: linear regression
- **How to interpret the coefficients from this model**
  - $E(Y|X = x) = \beta_0 + \beta_1 x$
  - Mean difference = β_1
- **Key assumptions of linear models:** no clustering or autocorrelation

Berkeley School of Public Health

To summarize, the focus of this video was linear multivariable models. And here are three of the key take home points we're hoping you'll walk away with. The first is that when we have a continuous type outcome, a linear regression model is going to be appropriate. And when we think about how to interpret the coefficients from this model, here's our model specification. And the mean difference is the measure of association estimated and it's equal to beta 1.

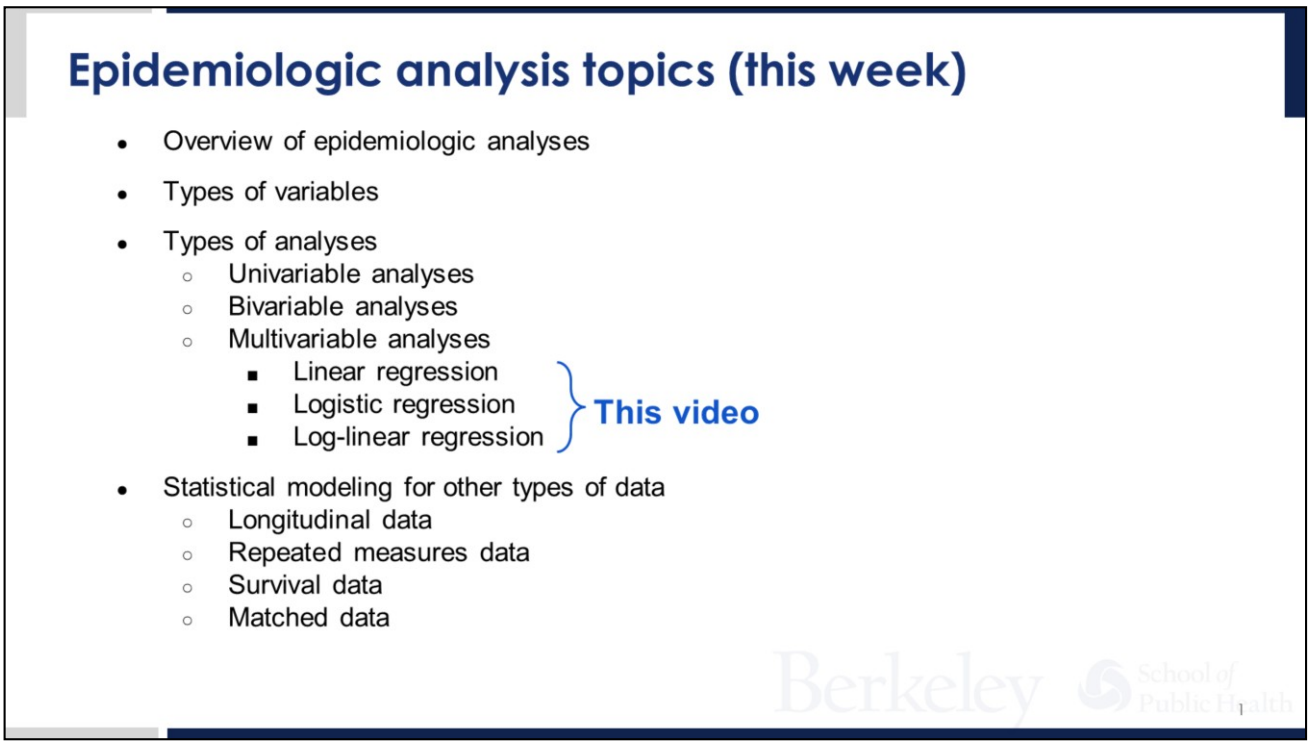
And then, the key assumptions of linear models are that there's no clustering or auto-correlation in the data. And if that assumption is not met, we have to think about some of the other types of statistical models that were introduced at the very end of this video.

Berkeley School of Public Health

Interaction in multivariable analyses of epidemiologic data

PHW250 B – Andrew Mertens

Andrew Mertens: This video will introduce you to interaction in multivariableanalysis of epidemiologic data.



Here's a list of the topics we've covered or plan to cover related to epidemiologic analysis. We didn't specifically list interaction in this list, but it's worth mentioning that the concepts we cover in this video are related to all three types of models here, linear, logistic, and log-linear models

We're going to illustrate them to you with linear models because they're the most straightforward, but all the concepts will also apply to logistic and log-linear models as well.

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- So far the models we have looked at have assumed there is no interaction.
- How do we specify potential interaction in our model if we suspect that it exists and want to investigate it?
- To assess possible interaction using a regression model, we include a term in the model that is the product of two other covariates.
- In the model above, β_3 assess the interaction between x_1 and x_2.

Berkeley School of Public Health

So far the models we've looked at have assumed that there's no interaction, but what if we want to specify a potential interaction in our model because we suspect that there may be an interaction between two variables and we want to investigate it?

We can do this by including a term in the model that's a product of two covariates. So in the model at the top we have variables x1 and x2, and the last term beta 3 times x1 times x2 is our interaction term.

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

	x_1	x_2	$x_1 * x_2$
x_1 present, x_2 absent	1	0	0
x_1 absent, x_2 present	0	1	0
x_1 present, x_2 present	1	1	1
x_1 absent, x_2 absent	0	0	0

- β_0 = mean of Y when x_1 and x_2 are absent
- $\beta_0 + \beta_1$ = mean of Y when x_1 is present and x_2 is absent
- $\beta_0 + \beta_2$ =mean of Y when x_1 is absent and x_2 is present
- $\beta_0 + \beta_1 + \beta_2 + \beta_3$ = mean of Y when x_1 and x_2 are both present

Berkeley School of Public Health

Let's take a look at how multiplying two indicator variables x1 and x2 works. Now, again, this is assuming that x1 is a binary variable. There are 1 and x2 is a binary variable, 0, 1 as well.

So when x1 is present and x2 is absent, x1 equals 1 and x2 equals 0, and if we multiply them it's equal to 0. When x1 is absent and x2 is present x1 is 0, x2 is 1, and when we multiply them it still equals 0. When both are present they're both equal to 1, and when we multiply them the product is equal to 1. And then when they're both absent they're both equal to 0 and the product equals 0.

So essentially when we multiply them what we're doing is we're creating an indicator that's equal to 1 when both are present and it's equal to 0 otherwise. This means that beta 0 is equal to the mean of y when x1 and x2 are absent. So we can figure that out by plugging in 0 to x1 and x2 and seeing that the last three terms of the model cancel, leaving us with beta 0.

Beta 0 plus beta 1 is the mean of y when x1 is present and x2 is absent. So if we plug a 0 in for x2 above beta 2x2 and then beta 3x1 times x2 drop out, leaving us with beta 0 plus beta 1. Beta 0 plus beta 2 is the mean of y when x1 is absent and x2 is present. So if we plug in a 0 for x1, beta 1x1 drops out and beta 3x1x2 drop out, leaving us with beta 0 plus beta 2, and then the sum of beta 0 through beta 3 is the mean of y when both are present. And that's because when we plug a 1 in for x1 and a 1 in for x2 nothing cancels out.

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

	x_1	x_2	$x_1 * x_2$
x_1 present, x_2 absent	1	0	0
x_1 absent, x_2 present	0	1	0
x_1 present, x_2 present	1	1	1
x_1 absent, x_2 absent	0	0	0

- β_1 = Mean difference Y between those with $x_1 = 0$ and $x_1 = 1$ among those with $x_2 = 0$
- β_2 = Mean difference Y between those with $x_2 = 0$ and $x_2 = 1$ among those with $x_1 = 0$
- β_3 = Mean additional difference in the value of Y beyond the effect of x_1 among those with $x_2 = 0$ and the effect of x_2 among those with $x_1 = 0$

Berkeley School of Public Health

So the last slide was talking about means. Now let's talk about mean differences. Beta 1 is the mean difference in y between those with x1 as 0 and x1 as 1 among those with x2 equal to 0. So this is the association with x1 when x2 is absent. Beta 2 is the mean difference of y between those with x2 as 0 and x2 as 1, among those with x1 and 0. So this is the association between y and x2 when x1 is absent.

Beta 3 is the mean additional difference in the value of y beyond the effect of x1 among those with x2 and 0 and the effect of x2 among those with x1 equals 0. Now that's quite a mouthful. Basically what that last sentence is saying is that beta 3 measures the additional joint effective x1 and x2 beyond the sum of the individual effects of x1 and the individual effect of x2.

So in our interaction week we talked about how we can assess the presence of interaction by comparing the observed association when both factors are present to the expected. And that difference is beta 3. We're going to come back to that momentarily.

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_1 = Mean difference Y between those with $x_1 = 0$ and $x_1 = 1$ among those with $x_2 = 0$

1. Plug in values $x_1 = 1$ and $x_2 = 0$ into the model

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (0) + \beta_3 \cdot (1) \cdot (0)$$
$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$

Berkeley School of Public Health

So now I'm going to walk you through how we got that mean difference for beta 1. I'm not going to do the same for beta 2 because it's quite analogous and I encourage you to write that out on your own to see how that works. And then after that I'll walk you through how we got the result for beta 3. So beta 1 is the mean difference in y between those with x1 is 0 and x1 is 1, among those with x2 is 0.

And our first step is to plug in the values x1 equals 1 and x2 equals 0 into our model above. And so we have the mean of y conditional in x1 equals 1, x2 equals 0. And that's beta 0 plus beta 1 times 1 plus beta 2 times 0 plus beta 3 times 1 times 0. And so we did times 0 plus beta 3 times 1 times 0 drop out. This simplifies to beta 0 plus beta 1 times 1 because the other terms cancel out.

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_1 = Mean difference Y between those with $x_1 = 0$ and $x_1 = 1$ among those with $x_2 = 0$

1. Plug in values $x_1 = 1$ and $x_2 = 0$ into the model

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (0) + \beta_3 \cdot (1) \cdot (0)$$
$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$

2. Plug in values $x_1 = 0$ and $x_2 = 0$ into the model

$$E(Y|x_1 = 0, x_2 = 0) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0) + \beta_3 \cdot (0) \cdot (0)$$
$$E(Y|x_1 = 0, x_2 = 0) = \beta_0$$

Berkeley School of Public Health

So what we did in step one was we got our mean of y when x1 is 1 and x2 is 0 and that's the first term in our difference. But the second term is the mean of y when both are 0. And so that's what we do next. You plug in the values x1 is 0, x2 is 0 into the formula and every term cancels except beta 0.

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_1 = Mean difference Y between those with $x_1 = 0$ and $x_1 = 1$ among those with $x_2 = 0$

1. Plug in values $x_1 = 1$ and $x_2 = 0$ into the model

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (0) + \beta_3 \cdot (1) \cdot (0)$$
$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$

2. Plug in values $x_1 = 0$ and $x_2 = 0$ into the model

$$E(Y|x_1 = 0, x_2 = 0) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0) + \beta_3 \cdot (0) \cdot (0)$$
$$E(Y|x_1 = 0, x_2 = 0) = \beta_0$$

3. Take the difference in betas from Step 1 and 2

$$E(Y|x_1 = 1, x_2 = 0) - E(Y|x_1 = 0, x_2 = 0) = \beta_0 + \beta_1 - \beta_0$$

Berkeley School of Public Health

Our third step is to take the difference in the betas from steps one and two. So the mean of y when x1 is 1 and x2 is 0 minus the mean of y when x1 is 0 and x2 is 0 equals beta 0 plus beta 1 from right here in step one minus beta 0 from right here in step two. So beta 0 and beta 0 drop out…

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_1 = Mean difference Y between those with $x_1 = 0$ and $x_1 = 1$ among those with $x_2 = 0$

1. Plug in values $x_1 = 1$ and $x_2 = 0$ into the model

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (0) + \beta_3 \cdot (1) \cdot (0)$$
$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$

2. Plug in values $x_1 = 0$ and $x_2 = 0$ into the model

$$E(Y|x_1 = 0, x_2 = 0) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0) + \beta_3 \cdot (0) \cdot (0)$$
$$E(Y|x_1 = 0, x_2 = 0) = \beta_0$$

3. Take the difference in betas from Step 1 and 2

$$E(Y|x_1 = 1, x_2 = 0) - E(Y|x_1 = 0, x_2 = 0) = \beta_0 + \beta_1 - \beta_0$$

This equation is equivalent to the text above.

$$E(Y|x_1 = 1, x_2 = 0) - E(Y|x_1 = 0, x_2 = 0) = \beta_1$$

Berkeley School of Public Health

8

…and that gives us an answer of beta 1. So that's equivalent to what we said at the top of the slide that beta 1 is equal to the mean difference in y among those with x1 is 0 and x1 is 1 when x2 is equal to 0.

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_2 = Mean difference Y between those with $x_2 = 0$ and $x_2 = 1$ among those with $x_1 = 0$

Practice this on your own!

Berkeley School of Public Health

So I'm not going to show you this step by step for beta 2 because it's very analogous to what I just showed you and I'm hoping that you take a few moments to pause the video and work through this on your own. I think it's a really helpful exercise.

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_3 = Mean additional difference in the value of Y beyond the effect of x_1 among those with $x_2 = 0$ and the effect of x_2 among those with $x_1 = 0$

1. From previous slide, the mean difference for $x_1 = 1$ and $x_2 = 0$

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$
$$\text{Mean Diff}_{10} = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

Berkeley School of Public Health

Now let's talk about why beta 3 is equal to the mean additional difference in the value of y beyond the effect of x1 among those without x2 and the effect of x2 among those without x1.

First let's just get the mean when x1 is 1 and when x2 is 0 and then we'll subtract the mean when both are 0. And we can just always remember that when both x1 and x2 are 0, the mean of y is beta 0. So our mean when x1 is 1 and x2 is 0 equals beta 0 plus beta 1 times 1. The beta 2 beta 3 terms cancel out because x2 is 0.

And so the mean difference-- and we're denoting this with this notation sub 1 0, which means when x1 is 1 and x2 is 0, the mean difference is equal to beta 0 plus beta 1 from above minus beta 0. And that's the mean when both are 0 and that equals beta 1.

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_3 = Mean additional difference in the value of Y beyond the effect of x_1 among those with $x_2 = 0$ and the effect of x_2 among those with $x_1 = 0$

1. From previous slide, the mean difference for $x_1 = 1$ and $x_2 = 0$

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$
$$\text{Mean Diff}_{10} = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

2. Calculate the mean difference for $x_1 = 0$ and $x_2 = 1$

$$E(Y|x_1 = 0, x_2 = 1) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (1) + \beta_3 \cdot (0) \cdot (1)$$
$$\text{Mean Diff}_{01} = \beta_0 + \beta_2 - \beta_0 = \beta_2$$

Berkeley School of Public Health

11

Now let's calculate a difference for x2 when x1 is 0. So it's basically what we did in step one, just flipping the roles of x1 and x2. So first let's get the mean of y when x1 is 0 and x2 is 1. So it's plugging this in. It's beta 0 plus beta 1 times 0 plus beta 2 times 1 plus beta 3 times 0 times 1. The second and fourth terms cancel and the mean difference when x1 is absent for x2 is beta 0 plus beta 2 minus beta 0 the mean difference when both are 0. And that gives us beta 2.

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_3 = Mean additional difference in the value of Y beyond the effect of x_1 among those with $x_2 = 0$ and the effect of x_2 among those with $x_1 = 0$

1. From previous slide, the mean difference for $x_1 = 1$ and $x_2 = 0$

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$
$$\text{Mean Diff}_{10} = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

2. Calculate the mean difference for $x_1 = 0$ and $x_2 = 1$

$$E(Y|x_1 = 0, x_2 = 1) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (1) + \beta_3 \cdot (0) \cdot (1)$$
$$\text{Mean Diff}_{01} = \beta_0 + \beta_2 - \beta_0 = \beta_2$$

3. Calculate the mean difference for $x_1 = 1$ and $x_2 = 1$

$$E(Y|x_1 = 1, x_2 = 1) = \beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (1) + \beta_3 \cdot (1) \cdot (1)$$
$$\text{Mean Diff}_{11} = \beta_0 + \beta_1 + \beta_2 + \beta_3 - \beta_0 = \beta_1 + \beta_2 + \beta_3$$

Berkeley School of Public Health

So the first two steps gave us the mean difference for each covariate on its own. Now let's calculate the mean difference for x1 and x2 together. So let's first get the mean of y when x1 is 1 and x2 is 1 and we plug in ones for x1 and x2 and that gives us beta 0 plus beta 1 plus beta 2 plus beta 3 beta 0 cancels and the mean difference then subtracting the mean when both are 0 is beta 1 plus beta 2 plus beta 3.

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_3 = Mean additional difference in the value of Y beyond the effect of x_1 among those with $x_2 = 0$ and the effect of x_2 among those with $x_1 = 0$

1. From previous slide, the mean difference for $x_1 = 1$ and $x_2 = 0$

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$
$$\text{Mean Diff}_{10} = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

2. Calculate the mean difference for $x_1 = 0$ and $x_2 = 1$

$$E(Y|x_1 = 0, x_2 = 1) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (1) + \beta_3 \cdot (0) \cdot (1)$$
$$\text{Mean Diff}_{01} = \beta_0 + \beta_2 - \beta_0 = \beta_2$$

3. Calculate the mean difference for $x_1 = 1$ and $x_2 = 1$

$$E(Y|x_1 = 1, x_2 = 1) = \beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (1) + \beta_3 \cdot (1) \cdot (1)$$
$$\text{Mean Diff}_{11} = \beta_0 + \beta_1 + \beta_2 + \beta_3 - \beta_0 = \beta_1 + \beta_2 + \beta_3$$

Take the difference between equations 1+2 and equation 3.

$$\text{Mean Diff}_{11} - \text{Mean Diff}_{10} - \text{Mean Diff}_{01} =$$
$$= (\beta_1 + \beta_2 + \beta_3) - \beta_1 - \beta_2 = \beta_3$$

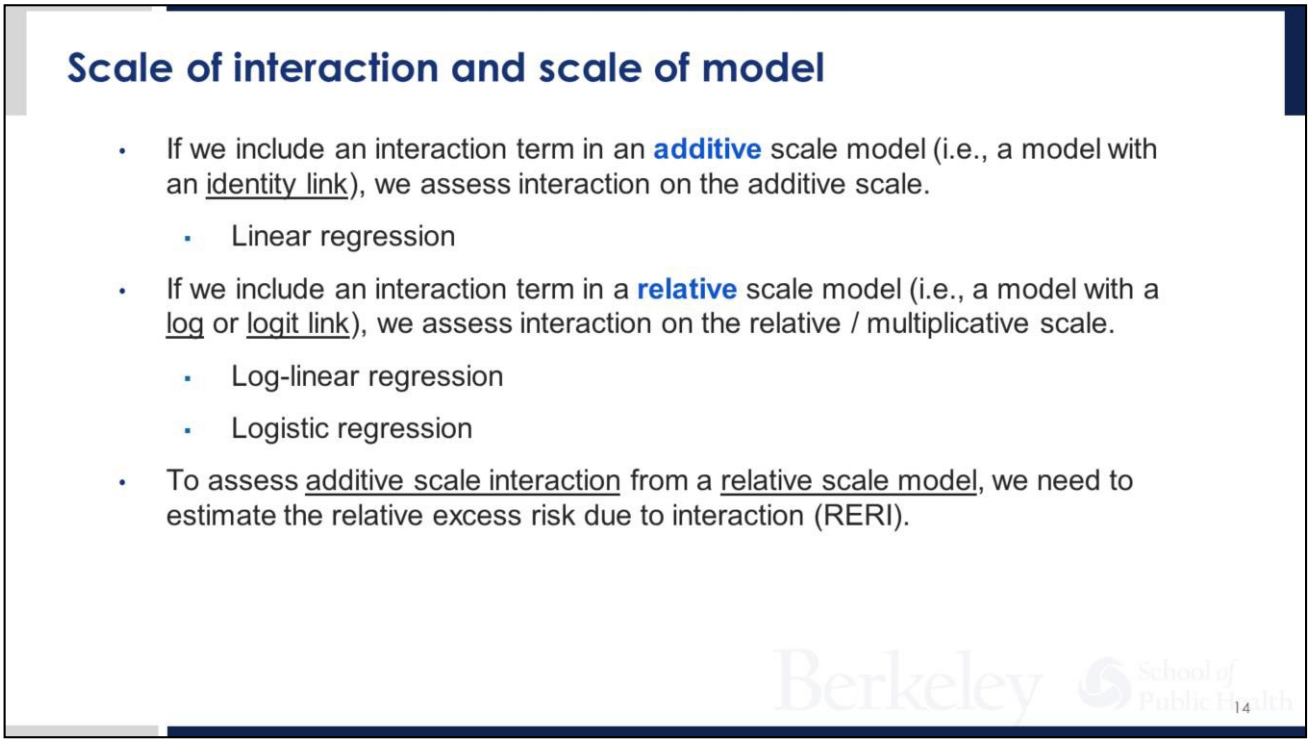
Berkeley School of Public Health

13

Now if we take the mean difference for both minus the mean difference for x1 minus the mean difference for x0, what do we get? Beta 1 plus beta 2 plus beta 3, so that's our mean difference for 1, 1, minus beta 1, which is our mean difference for x1 minus beta 2, which is our mean difference for x2. The beta 1 beta 2 cancel, leaving us with beta 3.

So this beta 3 and this formula here with our mean difference comparison is getting at this exact same concept we've covered a couple of different ways in this course. When we're interested in assessing interaction we want to know whether the expected joint interaction is different from the observed joint interaction. And this is exactly what beta 3 tells us. It tells us this difference.

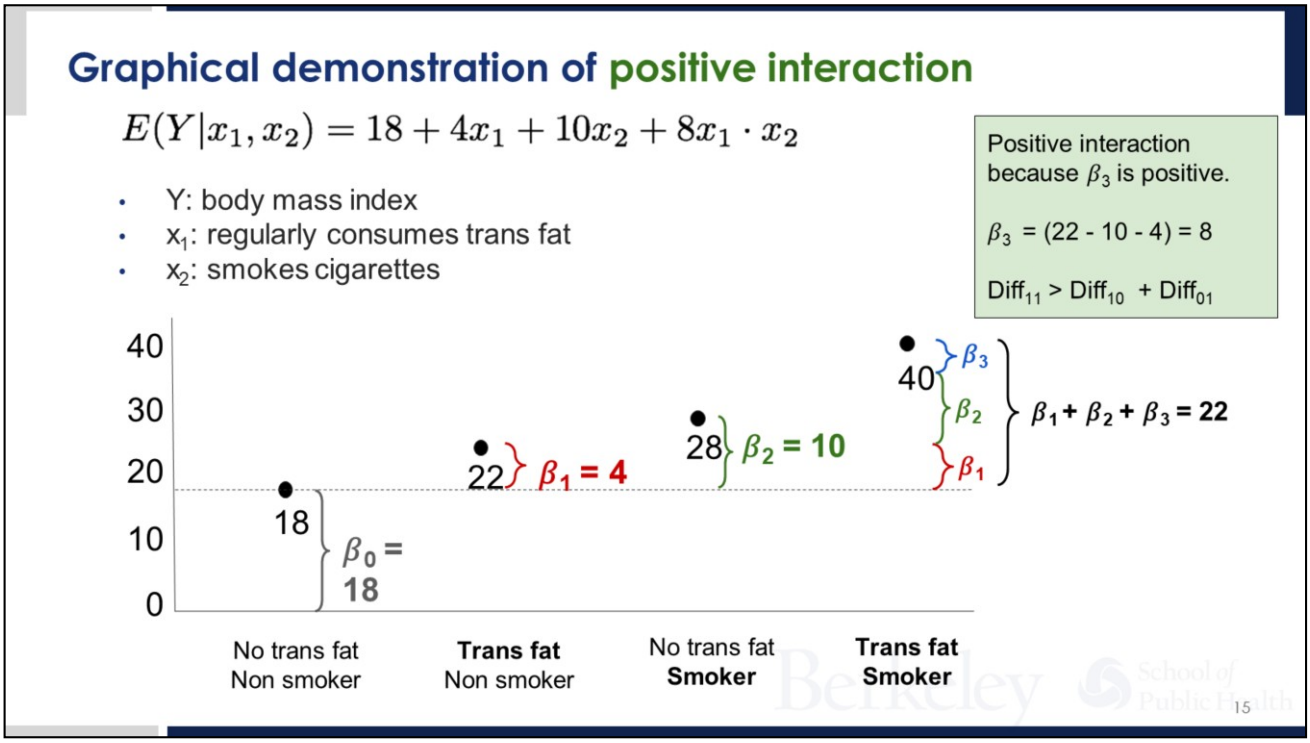
I'm going to show you graphically how this works in a few slides.



Now remember when we talked about the scale of interaction in our interaction module? So we've been focusing on additive scale interaction so far and that's because we're using a linear model. So this is a model with an identity link.

We're coming in the next video to models that use log or logit links and those estimate log-linear or logistic regression models. And those are models on the multiplicative or the relative scale. A relative scale model with a product term assesses interaction on the relative scale. A linear model with a product term assesses interaction on the additive scale.

If we only have a relative scale model, but we want to assess additive scale interaction,that's when we need to estimate the relative excess risk due to interaction.



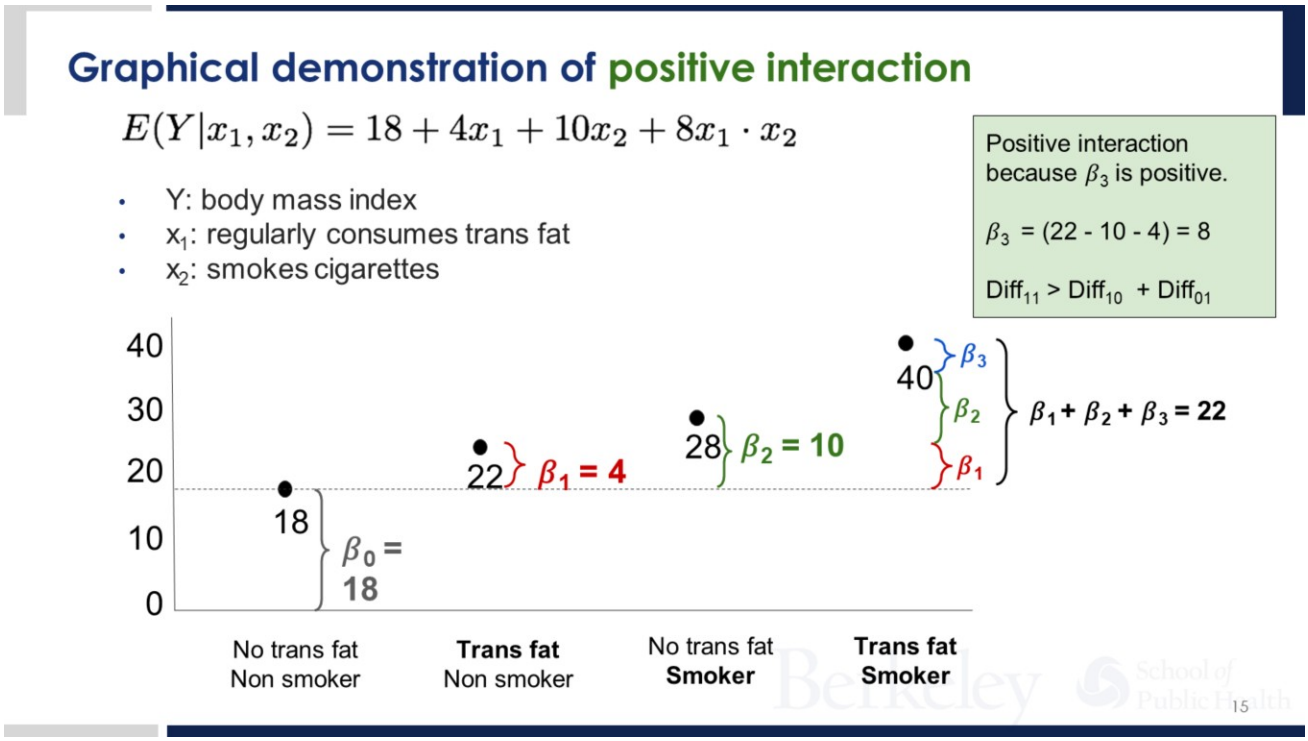
Let's link these betas to a graphical demonstration to help us better understand what beta 3 really means.

So here our example we have a y, our outcome of body mass index, x1 is an indicator 0, 1 for whether or not someone regularly consumes trans fat, x2 indicates whether they smoke cigarettes, and our BMI is on the y-axis here.

So these are our different combinations of beta coefficients when we have someone who is not a regular trans fat consumer and is a non-smoker, beta 0 is 18. OK. So when we plug in zeros for x1 and x2 what we're left with is the mean of y is 18. So that's our reference level.

Now let's get the mean for someone who consumes trans fat and is not a smoker. So that's x1 equals 1 and x2 equals 0. So what we get is 18 plus 4 because the 10 and the eight terms cancel out. And that's equal to 22. So 22 is coming from the sum of beta 0 plus beta 1. So this is our mean when someone consumes trans fat that doesn't smoke.

Now what's our mean when someone doesn't consume trans fat but they do smoke? So that means x1 is 0 and x2 is 1. If plug this into our formula at the top we get 18 plus 4 times 0 plus 10 times 1 plus 8 times 1, 8 times 0 times 1. And that gives us 18 plus 10, which is 28. And that's equal to beta 0 plus beta 2.



So what about when both are present trans fat consumption and smoking? x1 is 1, x2 is 2. We plug that in and we get 18 plus 4 plus 10 plus 8, and that gives us a BMI of 40. And then some of those is equal to 22.

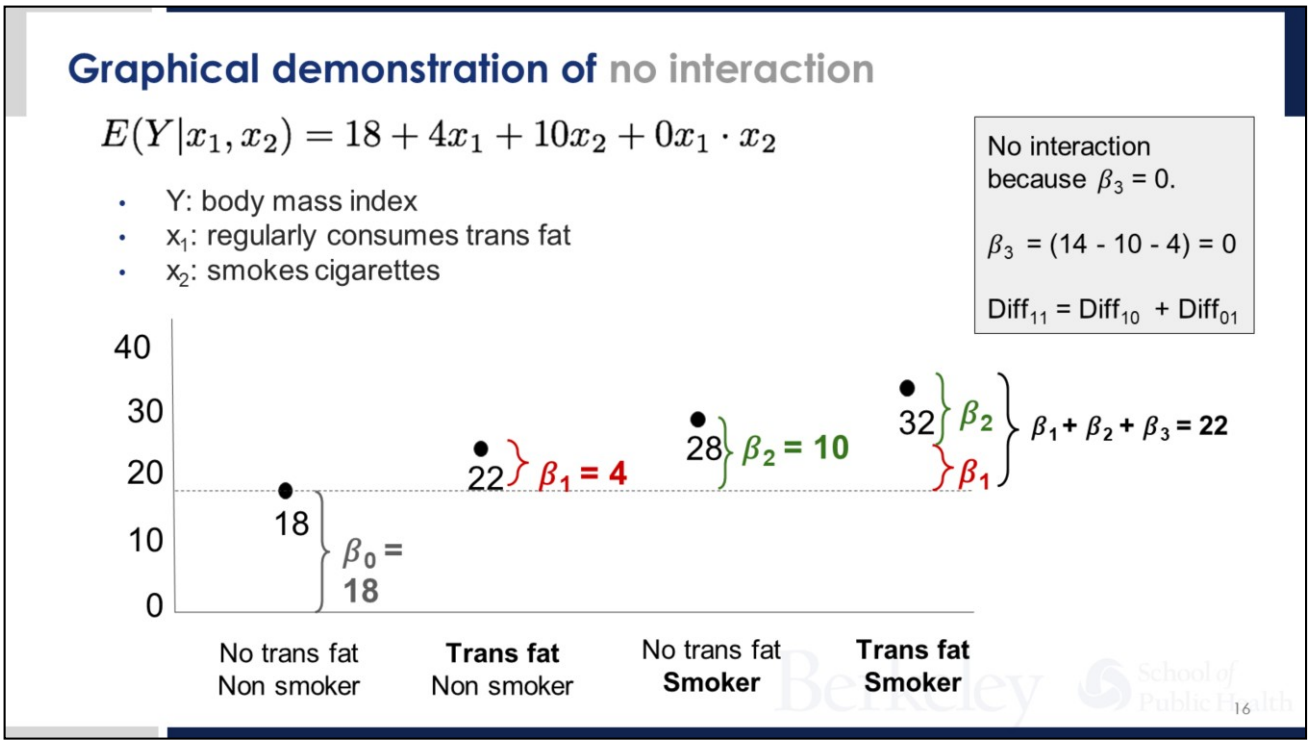
So now getting back to what beta 3 means, if we take the mean difference for each of these what we need to do is subtract out the mean BMI when there's no trans fat consumption and the person is a non-smoker. And that's equal to a BMI of 18.

So up here in our green box beta 3 is equal to 22 minus 10 minus 4. What does that come from? Well, beta 1 plus beta 2 plus beta 3 is equal to 22. And that is the difference in mean for someone who is a trans fat consumer and smoker minus the difference when neither are present.

The 10 comes from beta 2 plus beta 0 minus beta 0. So that's our mean difference for people who smoke and do not consume trans fat. And the 4 comes from beta 1 plus beta 0 minus beta 0. And that's for people who consume trans fat but don't smoke.

And so what we've done with beta 3 is we've said let's take the observed joint difference of 22 and subtract the sum of the individual differences, 10 and 4. Because beta 3 is positive and we have a linear model, this indicates that we have positive interaction on the additive scale. In other words, the observed joint effect of trans fat consumption and smoking is greater than what we would expect when we just sum up the individual effects of each exposure on its own.

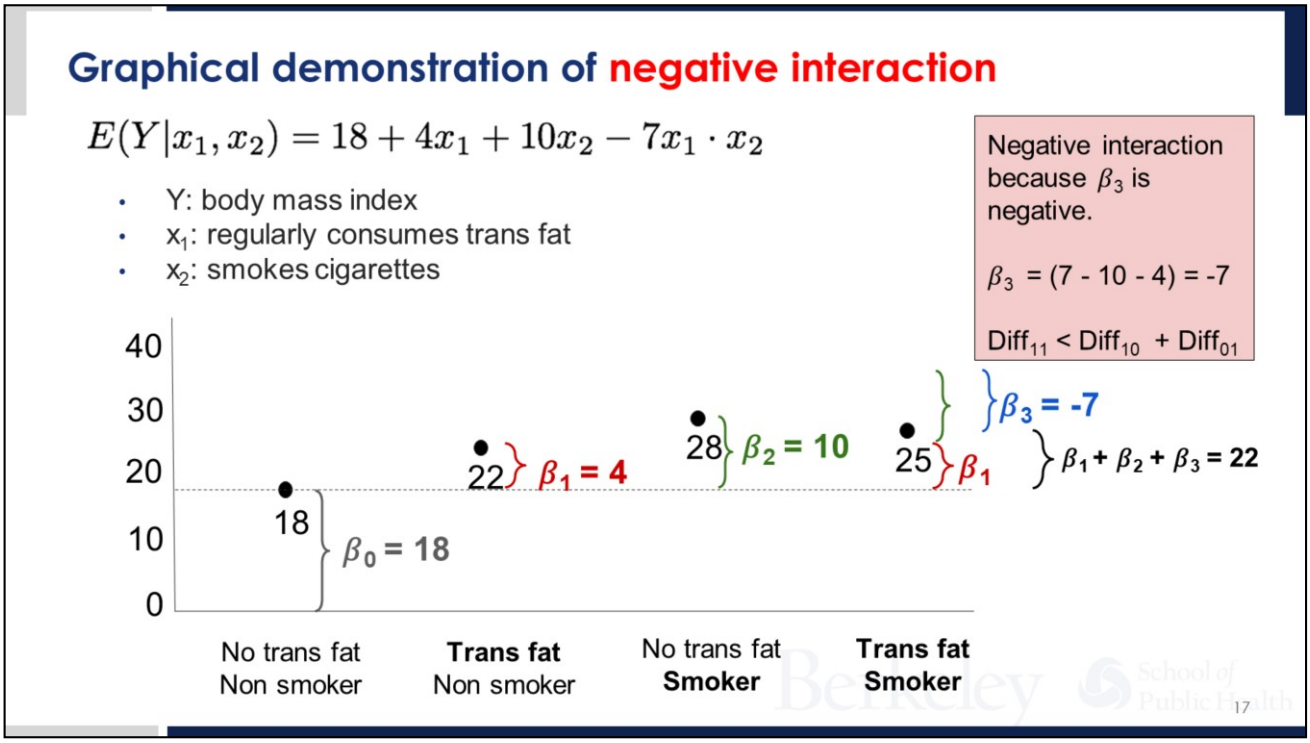
So the individual effect would be beta 1 and beta 2. So the sum of those is 14, but the joint effect is 22. So this is again just showing that beta 3 links back exactly to our concept for assessing the presence of interaction.



Here's an example of no interaction. So in this example, most things have stayed the same, but the difference now is in this last fourth column over here. So now we see that the 0, we have a beta equal to 0 on our interaction term. And as a result, the mean body mass index for people who consume trans fat and smoke is 32.

So our mean difference for both trans fat and smoking together is equal to beta 1 plus beta 2 plus beta 3 plus beta 0 minus beta 0. OK. So that's 18 plus 4 plus 10 minus 18, which is 14. And then our mean difference for each on their own, again, are still equal to beta 1 and beta 2, so that's 4 and 10. So the sum of these individual associations is 14 and that's equal to the observed joint association so beta 3 is 14 minus 10 minus 4 and it equals 0.

So this is an example of how beta 3 when beta 3 equals 0, it indicates an absence of interaction between these two variables, x1 and x2.

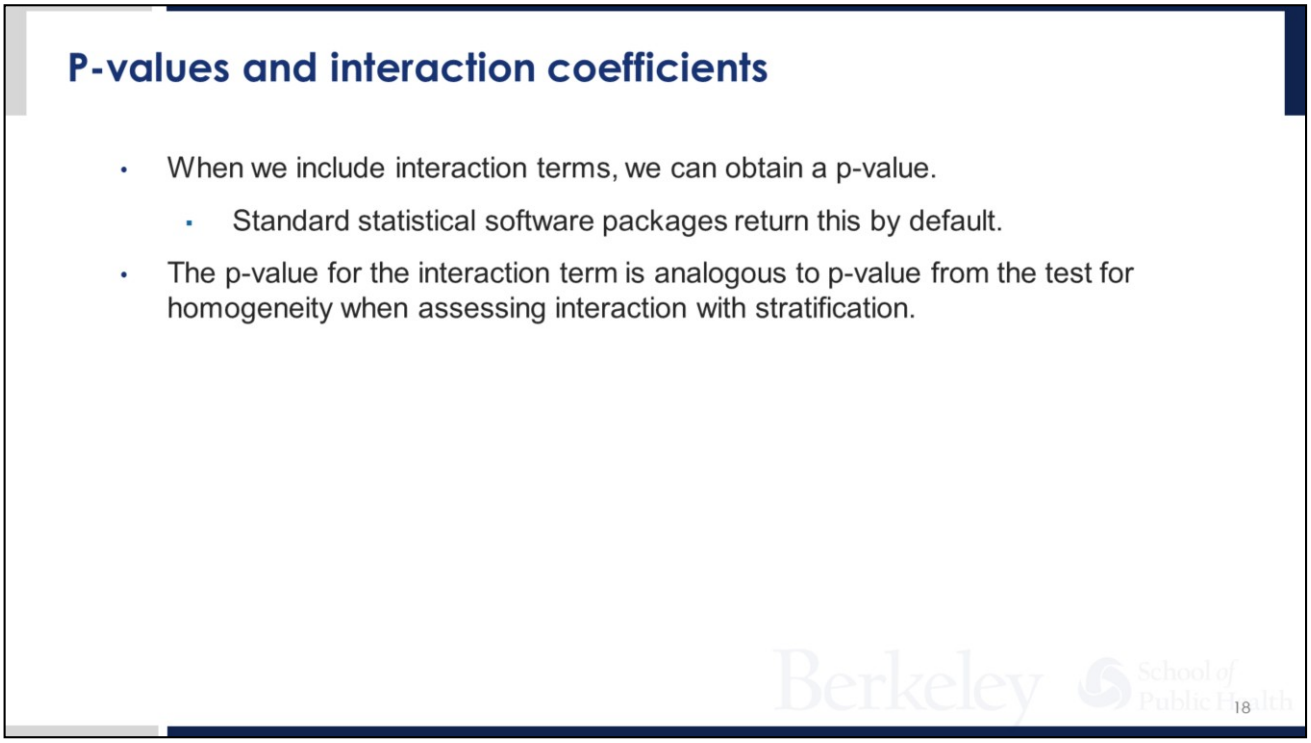


Now let's look at an example of negative interaction. So here we have a negative term on our interaction coefficient. It's negative 7. And again, everything on the first three columns of this plot has stayed the same. It's this fourth column that's different.

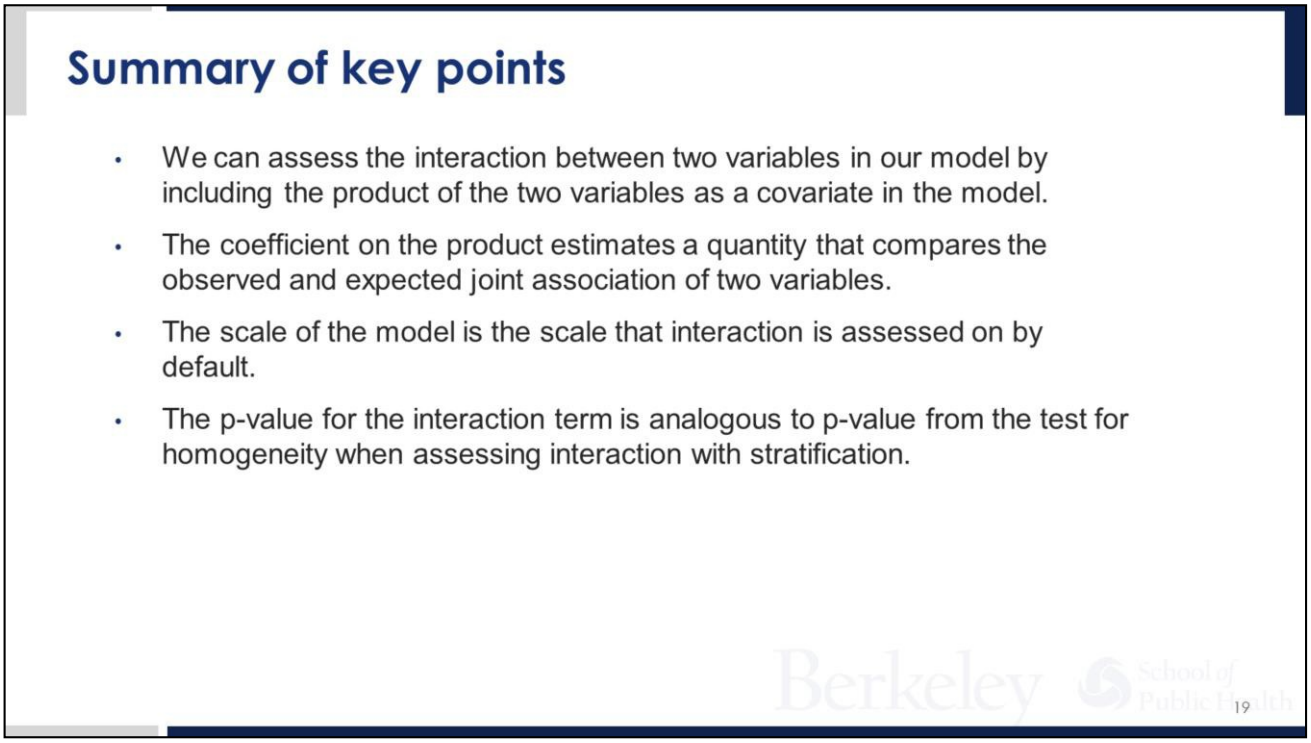
So now we have the sort of interesting situation where we sum up the effective of trans fat that's beta 1 and the effect of smoking, beta 2, but then beta 3 is negative. And so what we do is we take the sum of beta 1 and beta 2, so that's 14, but beta 3 is negative 7, so we subtract 7. And that means the sum of beta one beta 2 and beta 3 equals 7.

So that's our mean difference for people who consume trans fat and smoke compared to those who do not. And when we take our joint observed mean difference and compare it to the sum of our individual mean differences, the sum is greater. In other words, the joint observed effect is smaller than the expected joint effect. And as a result, we say we have negative interaction.

So this was a long way of saying that the direction of the sign on our interaction coefficient tells us the direction of the interaction. So positive coefficient indicates positive interaction, a 0 coefficient indicates no interaction, and a negative coefficient indicates negative interaction.



I just wanted to briefly mention that when we use statistical software to obtain models they're going to give us a p-value with our coefficients on the interaction term. And the p-value for the interaction term is analogous to p-value from the test of homogeneity that we learned about earlier in this course. So that's just something worth briefly mentioning if you want to assess the significance of an potential effect modifier using regression models in your public health career.



To summarize, we can assess the interaction between two variables in a statistical model by including a product term of those two variables as a covariate in the model. Now this is true when we have binary covariates. It's much more complicated when we have continuous covariates, and so we're not going to go into that situation in this course.

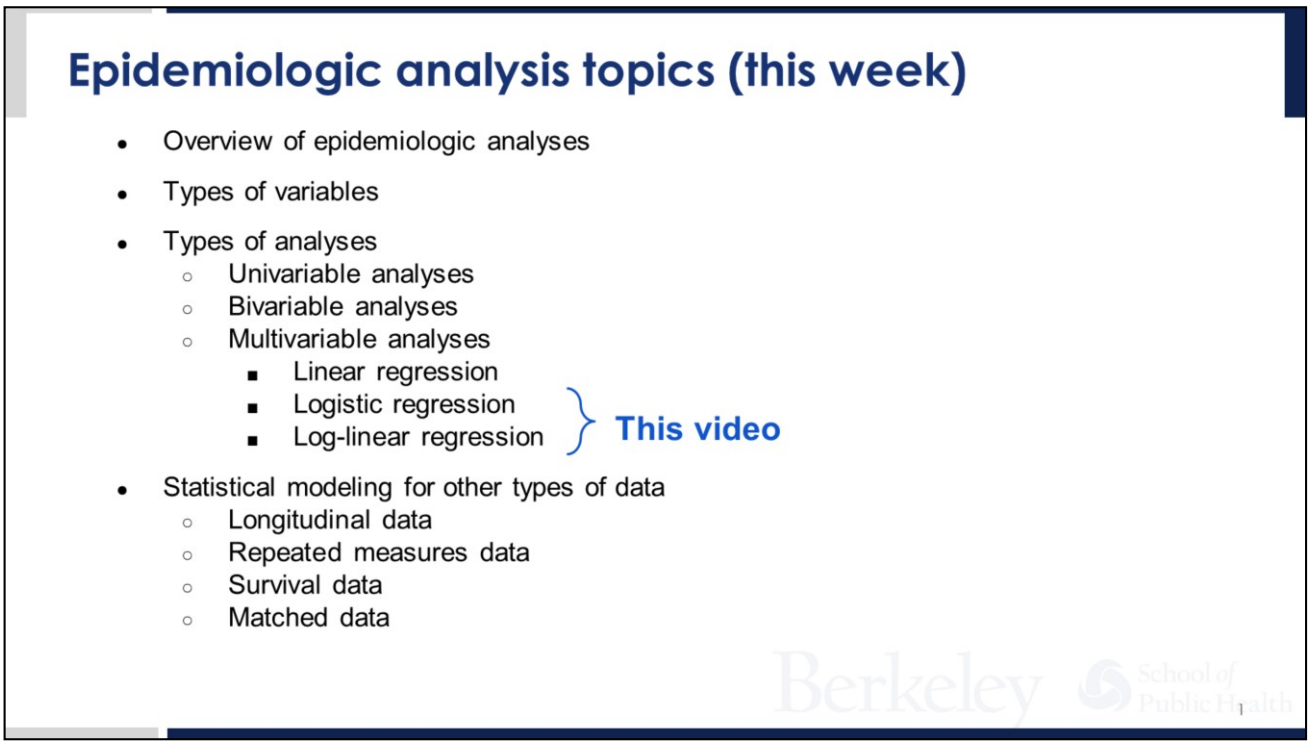
The coefficient on the product estimates a quantity that compares the observed and expected joint association of these two variables and indicates whether interaction is present or absent and the direction of the interaction. The scale of the model is the scale that the interaction is assessed on by default. And again, the p-value for the interaction term is analogous to the p-value from the test for homogeneity when assessing interaction with stratification.

Berkeley School of Public Health

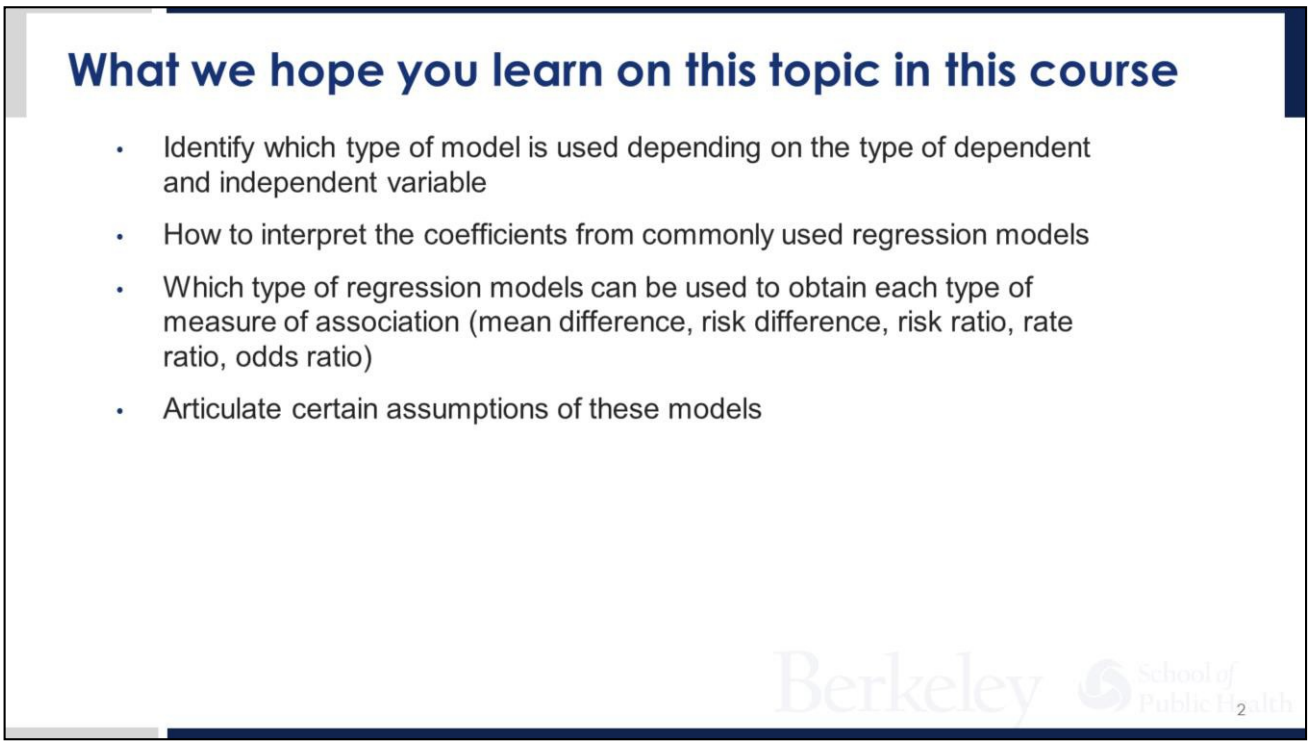
Multivariable log-linear and logistic regression models

PHW250 B – Andrew Mertens

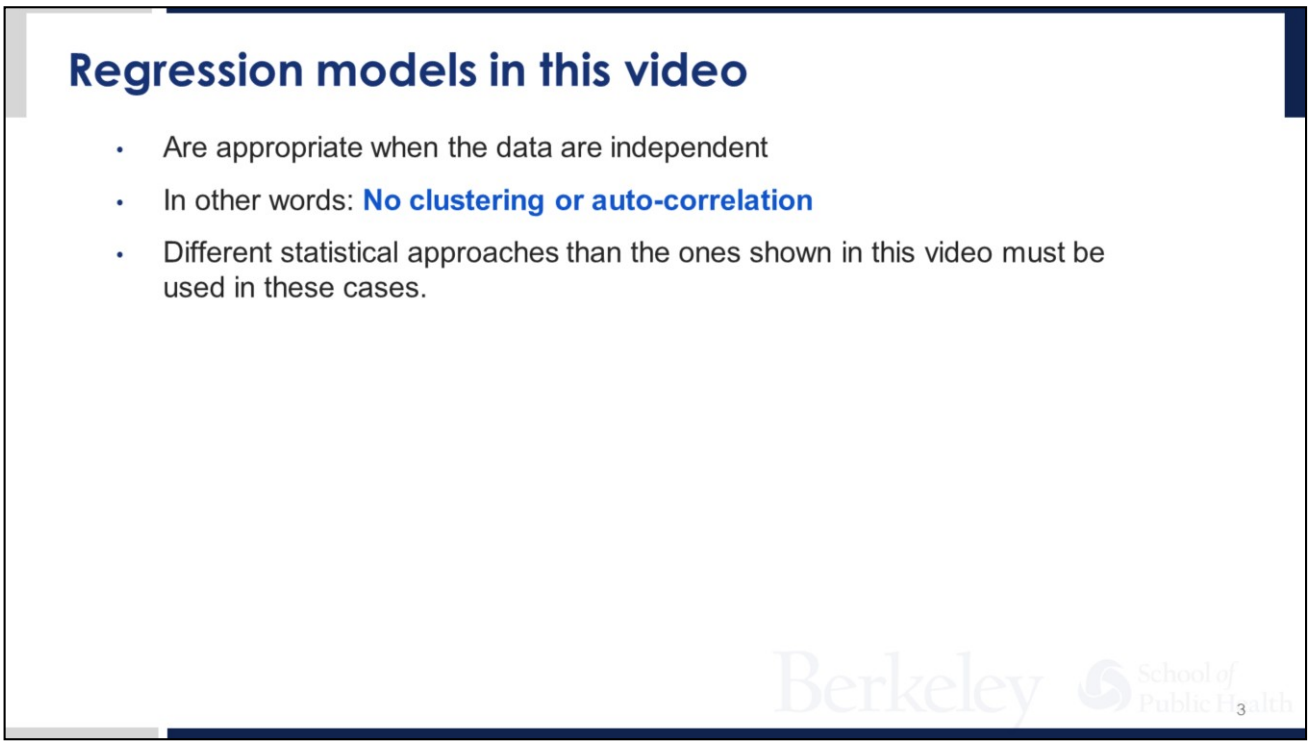
Andrew Mertens: This video will introduce you to multivariable log-linearand logistic regression models.



Here's our list of topics we've covered related to epidemiologic analysis. So now you've learned about linear regression models. And you've also learned about interaction terms in linear regression models. And those models focused on continuous outcome data. If you have binary outcome data or a categorical outcome data that you are willing to recode as binary or if you have rate information, logistic and log-linear models are your go tos. So that's the focus of our video.



And just to briefly reiterate this, you're going to see what feels like some technical notation here. And these are the real skills we hope you walk away with. We hope you can identify which type of model is appropriate to use depending on the type of dependent and independent variables you have, how to interpret the coefficients from this type of model, which model can be used to obtain the measure of association you want to estimate, and then some of the key assumptions of these models.



And I'll just briefly mentioned again all the models in this video assume the data are independent and that there's no clustering or auto-correlation. So video on linear regression models covered this in more detail. If this is the case, you still can estimate the measures of association I talk about in this video, but you need to use a different statistical approach to estimate those. You can't use logistic or log-linear models as I'm showing them here.

Summary of multivariable analyses

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method	Measure of association estimated
Continuous	>1	Continuous or binary	Multivariable	Linear regression	Mean difference
Binary or count	>1	Continuous or binary	Multivariable	Log-linear regression	Risk ratio or rate ratio
Binary	>1	Continuous or binary	Multivariable	Logistic regression	Odds ratio

Note: multivariable regression is also referred to as multiple regression

Berkeley School of Public Health

Here's the summary table I showed previously. And in this video, we're focusing on the bottom two rows where we have a binary or count outcome variable, more than one exposure and other variable. And our exposure can be continuous or binary. We're interested in the multivariable analysis. And if we want to estimate a risk ratio or rate ratio, we can use a log-linear regression. If we want to estimate an odds ratio, we can use a logistic regression. It's worth mentioning some people refer to multivariable regression as multiple regression. But we'll be using multivariable in this course.

Log-linear regression

$$\underbrace{\ln(E(Y|X=x))}_{\text{Log link}} = \underbrace{\beta_0 + \beta_1 x}_{\text{Coefficients}}$$

Note: you will see both **ln** and **log** used in the slides and the textbooks. In this course, you can always assume that **log** means **ln** (**log** with base e).

- For a binary exposure X, β_1 = the log risk or log rate ratio comparing the log risk or log rate when $X = 1$ and when $X = 0$
- Uses the **log link function**: we model the logarithm of the outcome as a function of the linear predictors
- Also known as "exponential risk" or "rate models"
- The log means the model can only predict positive values (no negative values).
- When modeling risk, some combinations of coefficients can lead to risks above 1, but typically not an issue for outcomes with relatively low risk.

Berkeley School of Public Health

Let's start with log-linear regression. It looks very similar to linear regression, except that on the left-hand side of the equation, we take the natural log of the mean outcome conditional on x. And as we indicated in this yellow box in the right-hand corner, you're going to see both ln and log-- l-o-g-- used in the slides and in the course materials, in the textbooks. And generally speaking, unless otherwise specified, in epidemiology and in this course in particular, if you see log, it means ln. It means log with base e. So don't let that trip you up. If we mean log 10, we always will put a 10 next to the log.

So what this means is we're taking the log of our mean outcome. And then we still have beta 0 plus beta 1 times X. For a binary exposure X, beta 1 is the log risk or log rate ratio comparing the log risk or log rate when X is 1 and when X is 0.

So we talked about link functions when we introduced linear multivariable regression. And we said we had an identity link for linear regression. Here we have a log link.

So the link function tells us how we're transforming the data on the left and right-hand side of the equation to connect the two. And in this case, we're taking the log on the left-hand side, not on the right-hand side. And so we call this a log link function. We model the logarithm of the outcome as a function of linear predictors, beta 0 plus beta 1x.

Log-linear regression

$$\underbrace{\ln(E(Y|X=x))}_{\text{Log link}} = \underbrace{\beta_0 + \beta_1 x}_{\text{Coefficients}}$$

Note: you will see both **ln** and **log** used in the slides and the textbooks. In this course, you can always assume that **log** means **ln** (**log** with base e).

- For a binary exposure X, β_1 = the log risk or log rate ratio comparing the log risk or log rate when $X = 1$ and when $X = 0$
- Uses the **log link function**: we model the logarithm of the outcome as a function of the linear predictors
- Also known as "exponential risk" or "rate models"
- The log means the model can only predict positive values (no negative values).
- When modeling risk, some combinations of coefficients can lead to risks above 1, but typically not an issue for outcomes with relatively low risk.

Berkeley School of Public Health

A convenient feature of this is that when we take the log it means the model can only predict positive values, no negative values. So that's nice because risks and rates can't be negative. Remember when we talked in the linear regression video about how sometimes you can actually use a linear regression model to model risks or rates, but the problem is that linear models don't constrain the estimates to be above 0 or to be between 0 and 1. So this is a nice feature of log linear models is that the estimates it returns are only going to be positive.

However, when we model risk, some combinations of coefficients are betas can lead to risks that are above 1. And so is typically not an issue if the outcomes have relatively low risk. But if the risk is higher, we have to think carefully about how to handle that. And again, that's beyond the scope of this class, but it is something to be aware of.

Obtaining the risk ratio (or rate ratio) from log-linear regression coefficients

1. Start with the log-linear model specification $\ln(E(Y|X=x)) = \beta_0 + \beta_1 x$

Berkeley School of Public Health

Now let's go through the steps of understanding how we obtain a risk ratio or a rate ratio from a log linear regression model. We start with our model specification, which we showed on the previous slide.

Obtaining the risk ratio (or rate ratio) from log-linear regression coefficients

1. Start with the log-linear model specification

$$\ln(E(Y|X=x)) = \beta_0 + \beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the risk as the difference in risk

$$\ln\left(\frac{E(Y|X=1)}{E(Y|X=0)}\right) = \ln(E(Y|X=1)) - \ln(E(Y|X=0))$$

Berkeley School of Public Health

So let's next use log rules to get our log risk or rate ratio. So here, we're going to focus on risk.

And what we can see is that using log rules, the log of the ratio of the mean outcome when X is 1 to the mean outcome when X is 0 equals the difference in the log of the mean of the outcome when x is 1 and the log of the mean of the outcome when X is 0. So this is just going back to log rules.

Obtaining the risk ratio (or rate ratio) from log-linear regression coefficients

1. Start with the log-linear model specification

$$\ln(E(Y|X=x)) = \beta_0 + \beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the risk as the difference in risk

$$\ln\left(\frac{E(Y|X=1)}{E(Y|X=0)}\right) = \ln(E(Y|X=1)) - \ln(E(Y|X=0))$$

3. Next we fill in the coefficients for each term in the difference in risk.

$$\ln(E(Y|X=1)) - \ln(E(Y|X=0)) = \beta_0 + \beta_1 \cdot (1) - \big(\beta_0 + \beta_1 \cdot (0)\big)$$

Berkeley School of Public Health

Next, let's take this difference and fill in the value for x using our betas. So we have beta 0 plus beta 1 times 1 for the first term minus beta 0 plus beta 1 times 0 for the second term. And these cancel out. And what we have left is beta 1.

Obtaining the risk ratio (or rate ratio) from log-linear regression coefficients

1. Start with the log-linear model specification

$$\ln(E(Y|X=x)) = \beta_0 + \beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the risk as the difference in risk

$$\ln\left(\frac{E(Y|X=1)}{E(Y|X=0)}\right) = \ln(E(Y|X=1)) - \ln(E(Y|X=0))$$

3. Next we fill in the coefficients for each term in the difference in risk.

$$\ln(E(Y|X=1)) - \ln(E(Y|X=0)) = \beta_0 + \beta_1 \cdot (1) - \big(\beta_0 + \beta_1 \cdot (0)\big)$$

4. After simplifying, we see that β_1 is equal to the natural log of the risk ratio.

$$\beta_1 = \ln\left(\frac{E(Y|X=1)}{E(Y|X=0)}\right)$$

Berkeley School of Public Health

So after we simplify, we see that beta 1 is equal to both the log difference in the risks, but also the log ratio of risks.

Obtaining the risk ratio (or rate ratio) from log-linear regression coefficients

1. Start with the log-linear model specification

$$\ln(E(Y|X=x)) = \beta_0 + \beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the risk as the difference in risk

$$\ln\left(\frac{E(Y|X=1)}{E(Y|X=0)}\right) = \ln(E(Y|X=1)) - \ln(E(Y|X=0))$$

3. Next we fill in the coefficients for each term in the difference in risk.

$$\ln(E(Y|X=1)) - \ln(E(Y|X=0)) = \beta_0 + \beta_1 \cdot (1) - \big(\beta_0 + \beta_1 \cdot (0)\big)$$

4. After simplifying, we see that β_1 is equal to the natural log of the risk ratio.

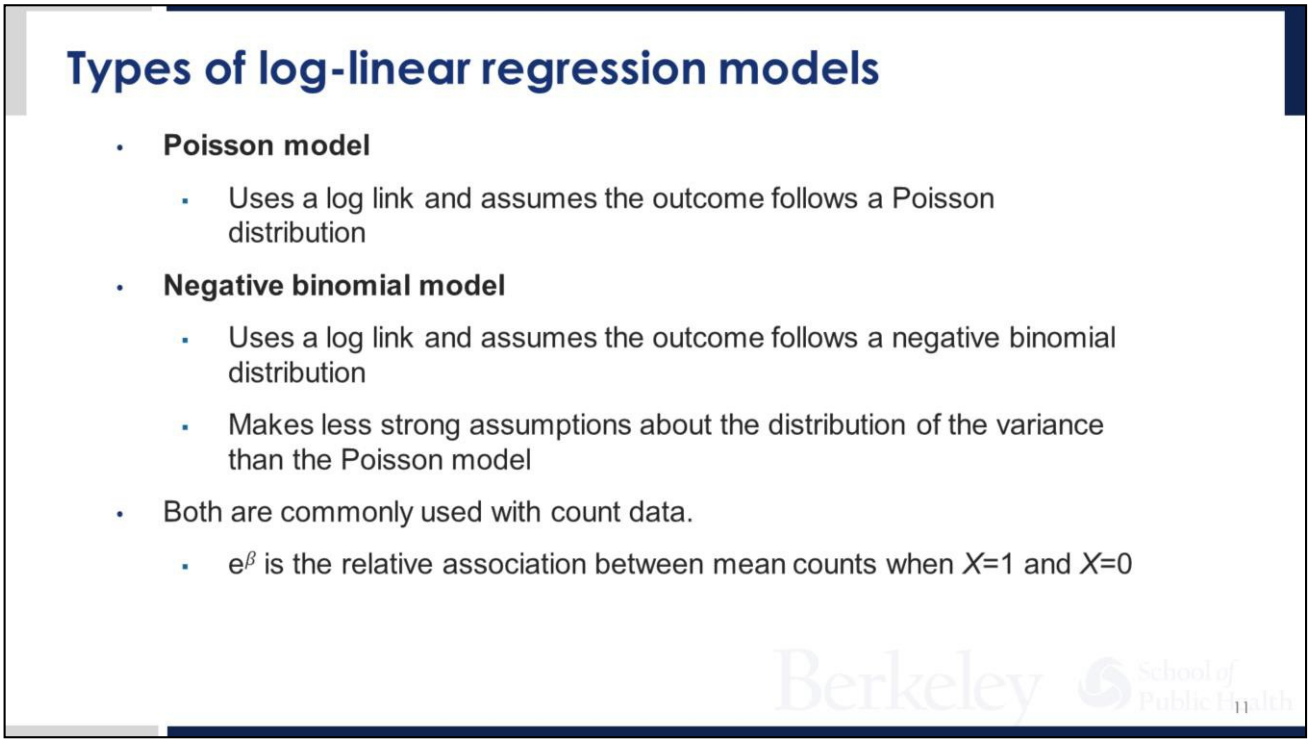
$$\beta_1 = \ln\left(\frac{E(Y|X=1)}{E(Y|X=0)}\right)$$

5. When we exponentiate β_1, it equals the risk ratio.

$$\exp(\beta_1) = \frac{E(Y|X=1)}{E(Y|X=0)}$$

Berkeley School of Public Health

And then if we want to get the risk ratio on its original scale-- so not on a log scale-- we can exponentiate beta 1. So exponential of beta 1 equals r risk ratio, comparing the risk when X is 1 to the risk when X is 0.



There are different types of log-linear models. So log-linear is a general term. And then within that category, we have Poisson models and negative binomial models, as well as other models. But these are the two most common to epidemiology. Poisson models use a log link and assume that the outcome follows a Poisson distribution. Negative binomial models are the same, but they assume the outcome follows a negative binomial distribution.

In statistics there are these different families of distributions that are known. So we know the shape of a Poisson distribution. To go back to more common one that you might be familiar with, a normal distribution, it follows a bell curve. So a Poisson distribution follows a different shaped curve. And a negative binomial distribution follows another shaped curve.

And we typically don't know whether our outcome data file a particular distribution or not in truth. We can look at the empirical distribution just by plotting a histogram of our outcome. But we don't know what it actually looks like.

Types of log-linear regression models

- **Poisson model**
  - Uses a log link and assumes the outcome follows a Poisson distribution
- **Negative binomial model**
  - Uses a log link and assumes the outcome follows a negative binomial distribution
  - Makes less strong assumptions about the distribution of the variance than the Poisson model
- Both are commonly used with count data.
  - e^{β} is the relative association between mean counts when X=1 and X=0

Berkeley School of Public Health

And so these models are saying, for the purpose of convenience, statistically, we're going to assume that the true distribution is a Poisson distribution, or the true distribution is a negative binomial distribution. So again, this is beyond the scope of this class. It's just good for you to have heard these names, because they'll come up in your future work potentially. These are the two most common types of models that are log-linear.

And negative binomial models make less strong assumptions about the distribution of the variance than Poisson models. And so sometimes they're used for that reason. Both are commonly used when the outcome is a count, so like the number of visits to a physician or the number of cases in a zip code. The exponential of the beta on the X term is the relative association and mean counts when X is 1 and X is 0 if we have a count outcome.

Using log-linear regression models to estimate rate ratios

$$\ln(E(Y|X=x)) = \beta_0 + \beta_1 x$$

$$\ln\left(\frac{E(Y|X=x)}{\text{Person-time}}\right) = \ln(E(Y|X=x)) - \ln(\text{Person-time})$$

$$= \ln(\text{Person-time}) + \beta_0 + \beta_1$$

- When the outcome is a count of events (e.g., a count of incident cases), we can estimate incidence rates by including a **person-time offset**.
- In log-linear models, this offset is the natural log of the person-time.

Berkeley School of Public Health

So the previous example focused on risk ratios. If we want to use log-linear regression to estimate a rate ratio, we need to use something called a person-time offset. That's because as you saw in the previous slide, we actually hadn't used any person-time. We didn't have rates. We had a risk. And if we want to estimate a rate ratio. This is how it works.

The top line here is the model specification when we're looking at risk. And the second line shows us what it looks like when we incorporate a person-time offset. So if we basically divide the mean of the outcome conditional on X by the person time, and that's inside the log, this can be rewritten as natural log of the mean outcome when x is x minus the natural log of the person-time. And then we can just move the natural log of the person-time to the right-hand side of the equation. And so that's called a person-time offset.

We're not going to go into much more detail beyond this in this course. But it's just helpful for you to have seen this to know that this is a nice way to estimate rate ratios if that's of interest to you, if that's the data structure that you have.

Example of log-linear regression

What risk factors are associated with incident coronary heart disease (CHD)?

TABLE 7-23 Data used for calculating the results shown in Table 7-22.

Cell	Male	Smok	Old age	Hyperten	Hypercho	Obese	PY	CHD	LogPY
1	0	0	0	0	0	0	1740.85	1	7.46213
2	0	0	0	0	0	1	1181.40	2	7.07446
3	0	0	0	0	1	0	539.97	0	6.29152
4	0	0	0	0	1	1	521.93	1	6.25754
…/…									
61	1	1	1	1	0	0	24.48	1	3.19804
62	1	1	1	1	0	1	37.41	0	3.62208
63	1	1	1	1	1	0	171.41	1	5.14405
64	1	1	1	1	1	1	85.37	5	4.44701

Berkeley School of Public Health

Szklo et al., 3rd Ed.

And here's a table from the Szklo textbook looking at the risk factors associated with incident coronary heart disease. So the goal was to calculate incident rate ratios. And we can see that there is 64 rows. The table is just showing a subset of these rows. And we have all these different variables. So the males, smoking, old age, all these variables with 0s and 1s are our independent variables.

And then we have person years in the PY column. CHD is a count of number of coronary heart disease events. And then we have the log of the person years in the final column. And all of these variables, except for the PY, are going to feed into our log-linear model specification to estimate a rate ratio.

Example of log-linear regression

$$\ln(\text{incidence}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

$$\ln(\text{incidence}) = -6.3473 + 1.1852 x_1 + 0.6384 x_2 + 0.2947 x_3 + 0.5137 x_4 + 0.6795 x_5 + 0.2656 x_6$$

TABLE 7-22 Results from a Poisson regression analysis of binary predictors of coronary heart disease (CHD) incidence in the Washington County cohort of the Atherosclerosis Risk in Communities (ARIC) Study, ages 45–64 years at baseline, 1987–1994.

Variable (unit)	Poisson regression coefficient	Rate ratio
Intercept	−6.3473	–
Gender (male = 1, female = 0)	1.1852	3.27
Smoking (yes = 1, no = 0)	0.6384	1.89
Older age* (yes = 1, no = 0)	0.2947	1.34
Hypertension† (yes = 1, no = 0)	0.5137	1.67
Hypercholesterolemia‡ (yes = 1, no = 0)	0.6795	1.97
Obesity§ (yes = 1, no = 0)	0.2656	1.30

*Age ≥ 55 years.
†Blood pressure ≥ 140 mm Hg systolic or ≥ 90 mm Hg diastolic or antihypertensive therapy.
‡Total serum cholesterol ≥ 240 mg/dL or lipid-lowering treatment.
§Body mass index ≥ 27.8 kg/m² in males and ≥ 27.3 kg/m² in females.

Binary covariate

Incidence rate ratio for CHD for males compared to females is $e^{1.1852} = 3.27$

Szklo et al., 3rd Ed.

14

So here's an example from the Szklo textbook. At the top, we have the typical model specification. And then this table here is showing us the coefficient results returned from the model process. And so this would be written like this at the top. The log incidence is equal to the intercept minus 6.3473 plus the coefficient on x1 times x1 plus the coefficient on 2 times x2 and so on. And so x1 is gender. x2 is smoking. x3 is old age and so on.

And to show you how to interpret one of these, if we want to get the rate ratio for gender-- so that's the incidence rate ratio for coronary heart disease comparing males to females-- we take the Poisson regression coefficient. That's 1.1852. And we exponentiate it, and that gives us 3.27. So that is the incidence rate ratio.

Logistic regression

$$\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right) = \underbrace{\text{logit}(E(Y|X=x))}_{\text{Logit link}} = \underbrace{\beta_0+\beta_1 x}_{\text{Coefficients}}$$

- For a binary exposure,
  - β_1 is the log odds ratio comparing the odds of the outcome when $X = 1$ and $X = 0$
  - $e^{\beta 1}$ is the odds ratio comparing the odds of the outcome when $X = 1$ and $X = 0$
- Uses the **logit link function**: we model the log-odds of the outcome as a function of the linear predictors
- The model only predicts values between 0 and 1.
  - This is the reason it is used so often even when measure of association under study is risk (not odds).
- Useful for case-control studies. The odds ratio produced by this model may estimate a CIR or IDR depending on the control sampling strategy.

Berkeley School of Public Health

15

Now let's move on to logistic regression. So logistic regression uses a different link function called a logit. So when we look inside the log in this term on the left here, what do we see? Well, this is an odds. So the mean of y over 1 minus the mean of y is an odds. So the logit is the log odds.

So for a binary exposure, beta 1 is the log odds ratio comparing the odds of the outcome when X is 1 and when X is 0. And if we exponentiate beta 1, it's equal to the odds ratio comparing the odds of the outcome when x is 1 an x is 0. So this is the type of model we use when we want to obtain an odds ratio.

This model constrains the predicted values to the range from 0 to 1. And this is probably the reason why it is used the most often when our measure of association is actually a risk ratio rather than an odds ratio. It's just this convenience mathematically that when we use this logit function, it's only going to return an estimate that ranges from 0 to 1 when we exponentiate our beta.

Unfortunately, odds ratios have been mistakenly interpreted as risk ratios countless times over the years. And so we encourage you to be extremely careful. If you're really interested in a risk ratio and you're going to use logistic regression, you need to be able to make the rare disease assumption that the odds ratio approximates the risk ratio. Otherwise, we would strongly encourage you to consider a log-linear model.

Logistic regression

$$\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right) = \underbrace{\text{logit}(E(Y|X=x))}_{\text{Logit link}} = \underbrace{\beta_0+\beta_1 x}_{\text{Coefficients}}$$

- For a binary exposure,
  - β_1 is the log odds ratio comparing the odds of the outcome when $X = 1$ and $X = 0$
  - $e^{\beta 1}$ is the odds ratio comparing the odds of the outcome when $X = 1$ and $X = 0$
- Uses the **logit link function**: we model the log-odds of the outcome as a function of the linear predictors
- The model only predicts values between 0 and 1.
  - This is the reason it is used so often even when measure of association under study is risk (not odds).
- Useful for case-control studies. The odds ratio produced by this model may estimate a CIR or IDR depending on the control sampling strategy.

Berkeley School of Public Health

15

Now, logistic regression is useful for case control studies because of the way that case control studies sample based on disease status. The odds ratio produced by logistic regression models in case control studies may estimate a cumulative incidence ratio, or an incidence density ratio depending on the control sampling strategy that was used.

Obtaining the odds ratio from logistic regression coefficients

1. Start with the logistic model specification

$$\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right)=\beta_0+\beta_1 x \qquad \ln(\text{Odds if } X=x)=\beta_0+\beta_1 x$$

Berkeley School of Public Health

Now let's go through how to obtain the odds ratio from logistic regression coefficient. So we're starting with this model specification. And for simplicity, I'm going to rewrite this quite complex term here as the log odds if X is x.

Obtaining the odds ratio from logistic regression coefficients

1. Start with the logistic model specification

$$\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right) = \beta_0 + \beta_1 x \qquad \ln(\text{Odds if } X=x) = \beta_0 + \beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the odds as the difference in odds

$$\ln\left(\frac{\text{Odds if } X=1}{\text{Odds if } X=0}\right) = \ln(\text{Odds if } X=1) - \ln(\text{Odds if } X=0)$$

Berkeley School of Public Health

Using log rules, we can rewrite the natural log of the ratio of the odds as a difference in odds. So this is just like what we did for log-linear regression. The log of a ratio is equal to the log of the numerator minus the log of the denominator.

Obtaining the odds ratio from logistic regression coefficients

1. Start with the logistic model specification

$$\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right) = \beta_0 + \beta_1 x \qquad \ln(\text{Odds if } X=x) = \beta_0 + \beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the odds as the difference in odds

$$\ln\left(\frac{\text{Odds if } X=1}{\text{Odds if } X=0}\right) = \ln(\text{Odds if } X=1) - \ln(\text{Odds if } X=0)$$

3. Next we fill in the coefficients for each term in the difference in log odds.

$$\ln(\text{Odds if } X=1) - \ln(\text{Odds if } X=0) = \beta_0 + \beta_1 \cdot (1) - \big(\beta_0 + \beta_1 \cdot (0)\big)$$

Berkeley School of Public Health

And then we can take the log difference of odds and plug in our regression coefficient. So we have beta 0 plus beta 1 times 1 when X is 1 minus beta 0 plus beta 1 times 0 when X is 0. And this cancels out to give us beta 1.

Obtaining the odds ratio from logistic regression coefficients

1. Start with the logistic model specification

$$\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right)=\beta_0+\beta_1 x \qquad \ln(\text{Odds if } X=x)=\beta_0+\beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the odds as the difference in odds

$$\ln\left(\frac{\text{Odds if } X=1}{\text{Odds if } X=0}\right)=\ln(\text{Odds if } X=1)-\ln(\text{Odds if } X=0)$$

3. Next we fill in the coefficients for each term in the difference in log odds.

$$\ln(\text{Odds if } X=1)-\ln(\text{Odds if } X=0)=\beta_0+\beta_1\cdot(1)-\big(\beta_0+\beta_1\cdot(0)\big)$$

4. After simplifying, we see that β_1 is equal to the natural log of the odds ratio.

$$\beta_1=\ln\left(\frac{\text{Odds if } X=1}{\text{Odds if } X=0}\right)=\ln(\text{OR})$$

Berkeley School of Public Health

So beta 1 is equal to the log odds ratio. Odds of X is 1 over odds of X is 0.

Obtaining the odds ratio from logistic regression coefficients

1. Start with the logistic model specification

$$\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right) = \beta_0 + \beta_1 x \qquad \ln(\text{Odds if } X=x) = \beta_0 + \beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the odds as the difference in odds

$$\ln\left(\frac{\text{Odds if } X=1}{\text{Odds if } X=0}\right) = \ln(\text{Odds if } X=1) - \ln(\text{Odds if } X=0)$$

3. Next we fill in the coefficients for each term in the difference in log odds.

$$\ln(\text{Odds if } X=1) - \ln(\text{Odds if } X=0) = \beta_0 + \beta_1 \cdot (1) - \big(\beta_0 + \beta_1 \cdot (0)\big)$$

4. After simplifying, we see that β_1 is equal to the natural log of the odds ratio.

$$\beta_1 = \ln\left(\frac{\text{Odds if } X=1}{\text{Odds if } X=0}\right) = \ln(\text{OR})$$

5. When we exponentiate β_1, it equals the odds ratio.

$$\exp(\beta_1) = \text{OR}$$

Berkeley School of Public Health

And we can exponentiate beta 1 to give us the odds ratio.

So this is how we can use a logistic regression model to get an odds ratio. We just exponentiate the coefficient on the variable of interest.

Example of logistic regression

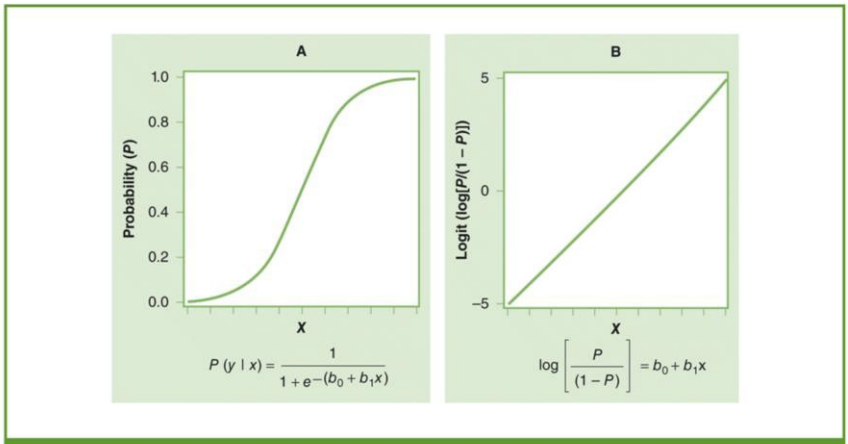


FIGURE 7-7 Two mathematically equivalent formulations of the logistic regression function.

Berkeley School of Public Health

So these graphs are assuming we have a continuous x. And what we can see is with a continuous x, under our logistic regression model, the probability does not follow a straight line. The formula for the relationship between x and y is shown below the plot. 1 over 1 plus e to the minus b0 plus b1x.

And while there are certain relationships in nature that do follow this kind of curved shape, that's not actually the reason why these models are typically used. It's really for statistical convenience.

And then panel B shows this that for this exposure x, if we fit a logistic regression model with this exposure and an outcome, it implies that the log odds of the outcome increases linearly with x. So we only see this linear increase in our relationship between x and y when we take the log odds of y.

Example of logistic regression

$$\ln(\text{odds ratio}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$
$$\ln(\text{odds ratio}) = -8.9502 + 1.3075 x_1 + 0.7413 x_2 + 0.0114 x_3 + 0.0167 x_4 + 0.0074 x_5 + 0.0240 x_6$$

TABLE 7-18 Results from a logistic regression analysis of binary and continuous predictors of coronary heart disease (CHD) incidence in the Washington County cohort of the Atherosclerosis Risk in Communities (ARIC) Study, ages 45–64 years at baseline, 1987–1994.

Variable (unit)	Logistic regression coefficient	Odds ratio
Intercept	−8.9502	–
Gender (male = 1, female = 0)	1.3075	3.70
Smoking (yes = 1, no = 0)	0.7413	2.10
Age (1 year)	0.0114	1.011
Systolic blood pressure (1 mm Hg)	0.0167	1.017
Serum cholesterol (1 mg/dL)	0.0074	1.007
Body mass index (1 kg/m^2)	0.0240	1.024

Binary covariate
Odds ratio for CHD for males compared to females is $e^{1.3075} = 3.70$.

Continuous covariate
Odds ratio for CHD for a 1 mm Hg increase in systolic blood pressure is $e^{0.0167} = 1.017$.

Berkeley School of Public Health

Let's go over an example. This is from the Szklo textbook. So here in this table, we're looking at predictors of coronary heart disease incidence. And it's important to mention that though the study is measuring incidence, when we use logistic regression, we're only estimating an odds ratio, which is different from a cumulative incidence ratio.

Here are the logistic regression coefficients from a model that included an outcome for coronary heart disease and various risk factors-- gender, smoking age, blood pressure, etc. Some of the risk factors were binary. For example, gender was operationalized as a binary variable in this study. And some were continuous, such as blood pressure.

And I've plugged in the coefficients to the model at the top. So if we want to go over two examples, to interpret the coefficient for a gender, it's binary. So with the odds ratio for the coronary heart disease for males compared to females is equal to the exponential of 1.3075. That's the coefficient on the gender variable. And that's equal to 3.7.

For continuous covariate, such as systolic blood pressure, that's measured in millimeters of mercury. The odds ratio for coronary heart disease for a 1 millimeters of mercury increase in systolic blood pressure is the exponential of 0.0167, which equals 1.017. So that's the odds ratio for just 1 unit increase in blood pressure.

Summary of key points

Log-linear models

- **Identify which type of model is used depending on the type of dependent and independent variable**
  - **Count / rate / binary outcome:** log-linear regression
- **How to interpret the coefficients from commonly used regression models**

$$\ln(E(Y|X = x)) = \beta_0 + \beta_1 x$$

  - RR = $e^{\beta 1}$
  - The exponentiated β_1 ($e^{\beta 1}$) is the risk or rate ratio comparing the risk or rate when x = 1 to x = 0
- **Key assumptions of log-linear models:** no clustering or autocorrelation

Berkeley School of Public Health

To summarize, here are the key things we'd like you to take away for log linear models. So when we're interested in using a model to estimate measures of association for a count or rate or binary outcome and we want to get a relative risk, let's use a log-linear regression model.

Here's our model specification. And if we exponentiate beta 1, the beta on the x term, that gives us the relative risk for x. The exponentiated term is the risk or rate ratio comparing the risk or rate when X is 1 to when X is 0. A key assumption of log linear models is that there is no clustering or auto-correlation in the data.

Summary of key points

Logistic models

- Identify which type of model is used depending on the type of dependent and independent variable
  - **Binary outcome:** logistic regression
- **How to interpret the coefficients from commonly used regression models**
  - $\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right) = \beta_0 + \beta_1 x$
  - OR = $e^{\beta 1}$
  - The exponentiated β_1 ($e^{\beta 1}$) is the odds ratio comparing the odds when x = 1 to x = 0
- **Key assumptions of logistic models:** no clustering or autocorrelation

Berkeley School of Public Health

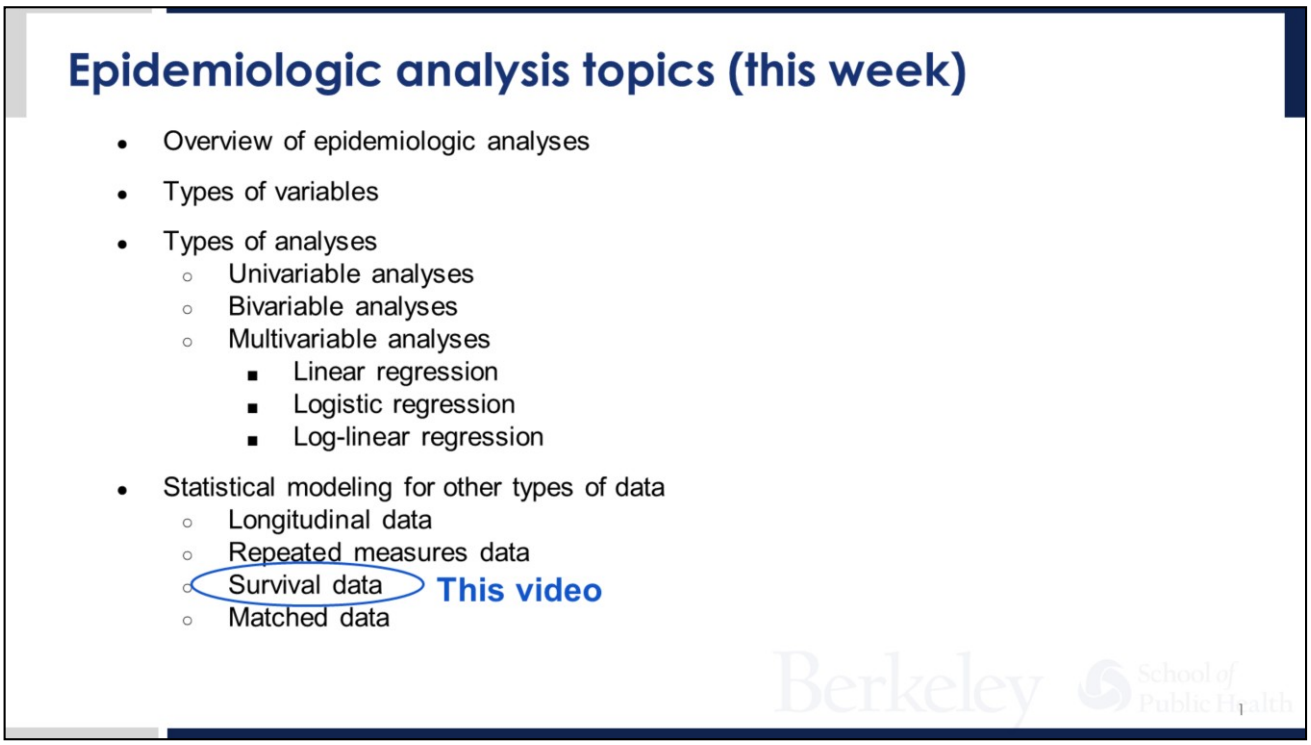
For logistic models, these are appropriate when we have a binary outcome and we're interested in estimating an odds ratio. And we can obtain this odds ratio by exponentiating the coefficient on the x variable of interest. So our exponentiated beta 1 is equal to the odds ratio comparing the odds when X is 1 to when X is 0. And these models also assume no clustering or auto-correlation.

Berkeley School of Public Health

Multivariable log-linear and logistic regression models

PHW250 B – Andrew Mertens

Andrew Mertens: This video will focus on analysis methods for a specialtype of data called survival data. It also can be called time-to-event data.



Here's the list of topics we've covered related to epidemiologic analyses. And as I just mentioned, we are focusing on survival data, which is a special type of data. And thus, it requires a special type of analysis.

Summary of multivariable analyses

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method	Measure of association estimated
Continuous	>1	Continuous or binary	Multivariable	Linear regression	Mean difference
Binary or count	>1	Continuous or binary	Multivariable	Log-linear regression	Risk ratio or rate ratio
Binary	>1	Continuous or binary	Multivariable	Logistic regression	Odds ratio
Time to event	>1	Continuous or binary	Multivariable	Cox proportional hazards model	Hazard ratio

Berkeley School of Public Health

So to add a row to the table we've been using to keep track of our different multivariable analyses, here's a row for time-to-event data. I'm about to define what that means. And so when we have an outcome variable that's a time-to-event variable, and we have more than one exposure and other variable, it's a multivariable model. And we're going to use something called Cox proportional hazards models to estimate hazard ratios with this data.

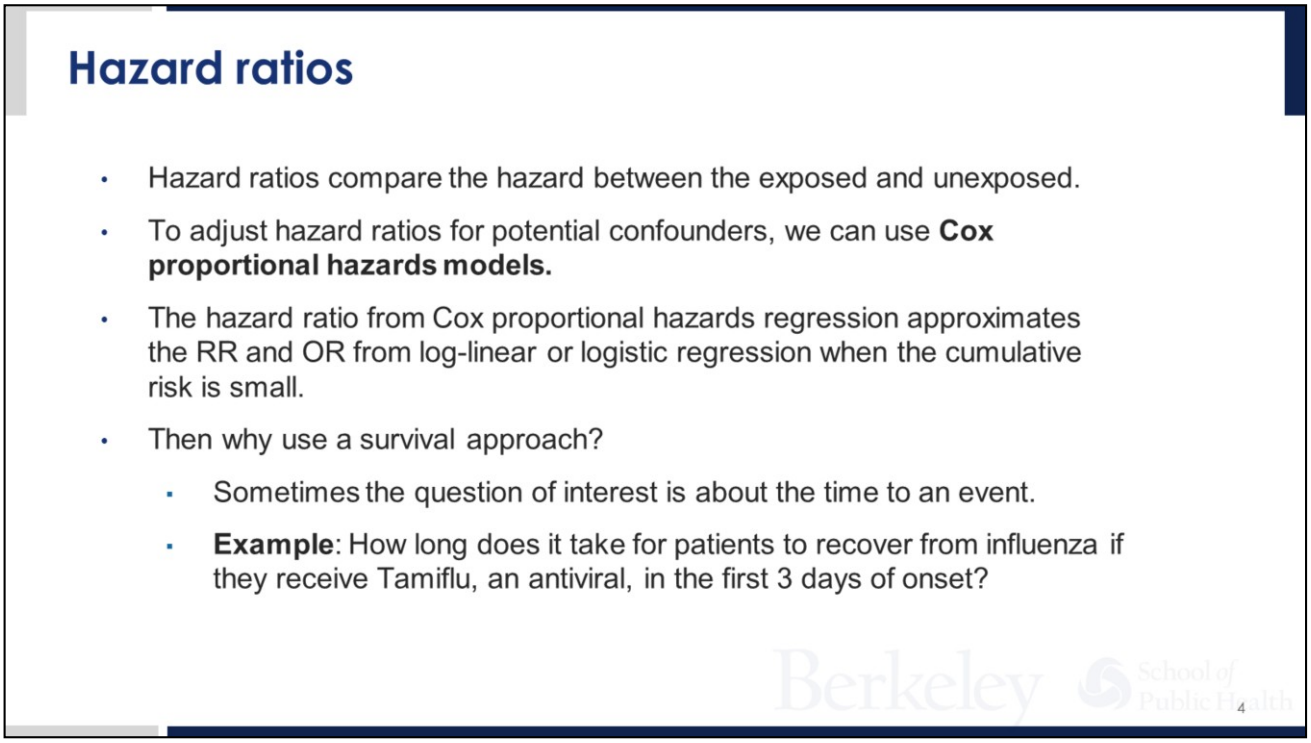
Survival data

- Survival data includes measurements of the time that passed before a person developed a certain disease or condition.
- Recall the concept of the **hazard** — the instantaneous potential for change in disease status per unit of time at time t relative to the size of the candidate (i.e., disease-free) population at time *t*
- **Examples of survival or "time-to-event" data:**
  - Among pancreatic cancer patients, the time till death
  - Among drivers who use their mobile phone while driving, the time till a car accident
  - Time to recovery for patients receiving a new form of hip replacement

Berkeley School of Public Health

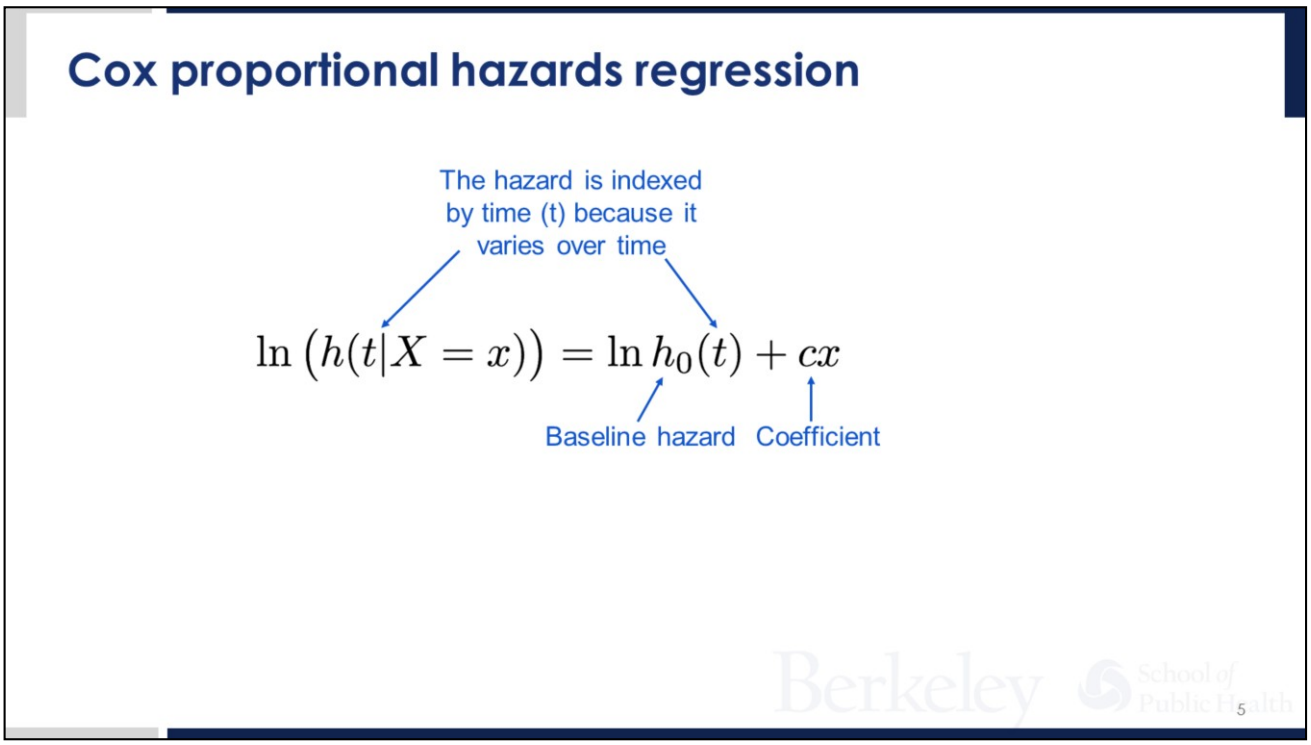
So what's survival data? Well, it includes measurements of time that passed before a person developed a certain disease or condition. Let's recap the concept of hazard, which we covered earlier in the course. Hazard is the instantaneous potential for change in disease status per unit of time at time t relative to the size of the candidate disease-free population at time t.

So it's similar to risk, but it's a measure of instantaneous risk. Here's some examples of survival or time-to-event data. Among pancreatic cancer patients, the time until death, among drivers who use their mobile phone while driving, the time till car accident, and time to recovery for patients receiving a new form of a hip replacement.



So we've talked about hazard ratios and how we can use them to compare the hazard between the exposed and the unexposed. And if we want to adjust for potential confounders when we estimate our hazard ratio, we need to use something called a Cox proportional hazards model. The hazard ratio estimated from these models approximate the relative risk or the odds ratio from a log linear or logistic regression when the cumulative risk is small.

So what's the value of using a survival approach if we're more interested in a risk or rate? Well, sometimes our research question really is about the time to an event. An example of this would be a study that wants to know how long it takes for patients to recover from influenza if they receive Tamiflu, which is an antiviral flu medication, in the first three days of onset. So this research question inherently involves time to recovery. And so for this kind of research question, a Cox proportional hazards model is the appropriate type of model because it's focusing our estimation on a hazard ratio.



Here's the notation for a Cox proportional hazards model. It's a little bit different from notation we've used for other models. So I just want to go over this piece by piece. So on the left-hand side of the equation, we have the log of h of t conditional on X. h indicates the hazard, and it's indexed by time t. And the reason we say it's indexed is because, when we say h parentheses t, that means that the hazard changes over time. When we use the notation h sub 0, that indicates the baseline hazard. And c is the letter we're going to use for our coefficient in this model.

Cox proportional hazards regression

$$\ln\left(h(t|X=x)\right) = \ln h_0(t) + cx$$

Baseline hazard Coefficient

$$h(t|X=x) = e^{(\ln h_0(t) + cx)} = h_0(t) \times e^{cx}$$

Berkeley School of Public Health

So the first specification I showed you showed everything on the log scale. If we exponentiate through, what we see is we get the hazard time t conditional on X is equal to the exponential of the log of the baseline hazard plus the coefficient. And using log rules, we can simplify this to be the baseline hazard times the exponential of the coefficient times X.

Cox proportional hazards regression

$$\ln\left(h(t|X=x)\right) = \ln h_0(t) + cx$$

Baseline hazard Coefficient

- Models the log of the hazard as a function of the log of the baseline hazard (the hazard where all X=0, $h_0(t)$) and covariates (Xs).
- For a binary exposure X,
  - c = the log hazard ratio comparing hazard when $X = 1$ to $X = 0$
  - e^c = hazard ratio = relative hazard at all times t comparing hazard when $X = 1$ to $X = 0$
- Called a proportional hazards model because it assumes that the ratio of the hazards is constant across the time interval.
- Cox devised a method for estimating regression coefficients without needing to specify an intercept.

Berkeley School of Public Health

7

So Cox proportional hazards models model the log of the hazard as a function of the log of the baseline hazard-- that's the hazard where all of X equals 0-- h0 of t, and the covariates. If our binary exposure is X, then c, our coefficient, is the log hazard ratio comparing the hazard when X is 1 and X is 0, and the exponential of c hazard ratio is the relative hazard at all times t comparing the hazard when X is 1 and X is 0.

We call this a proportional hazards model because it assumes that the ratio of the hazard is constant across the time interval. And Cox, who created this method, devised a way of estimating regression coefficients without needing to specify an intercept. So that's beyond the scope of this class, but it's sort of an interesting thing to know.

Proportional hazards

- Recall: Hazards cannot be directly calculated because they are defined for an infinitely small time interval

$$h(t) = \frac{P(\text{event in interval between } t \text{ and } [t + \Delta t] \mid \text{alive at } t)}{\Delta t}$$

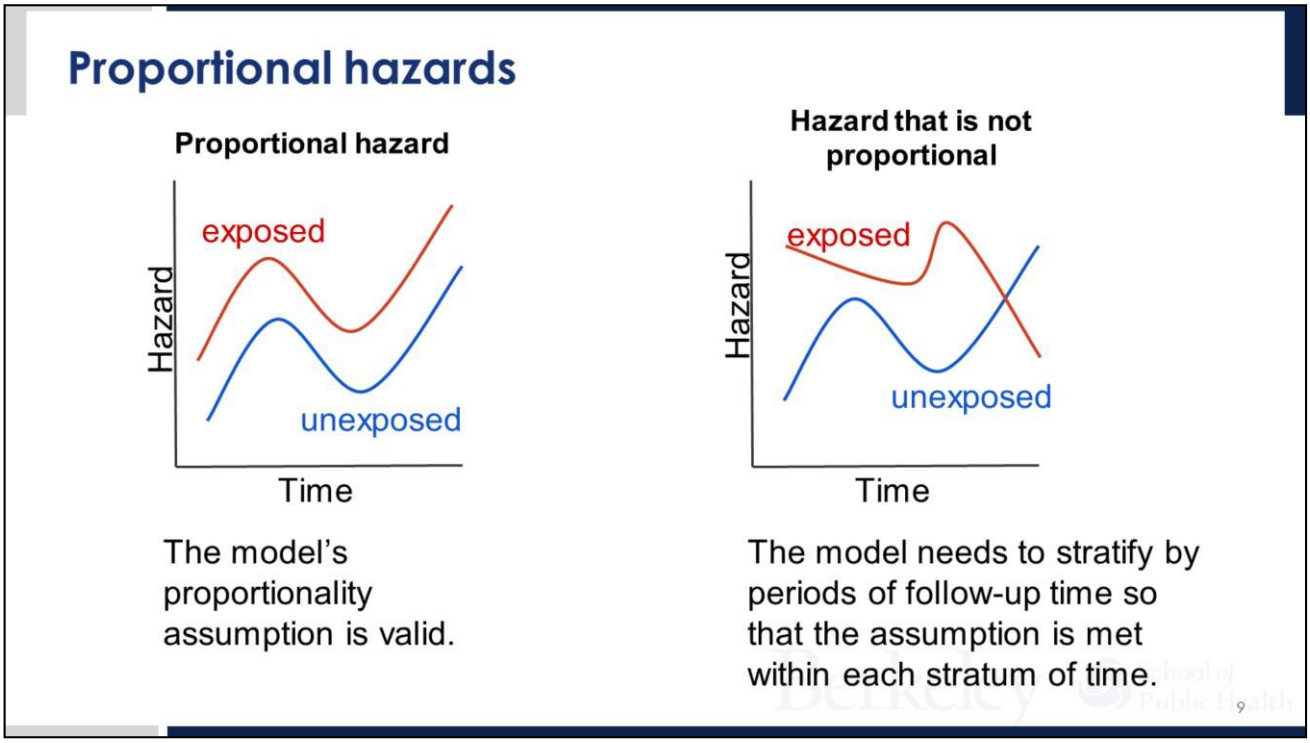
- However, if the relative hazard can be assumed to remain constant over an interval it can be modeled without the need to estimate the actual hazard.
- The exposure is associated with a fixed relative increase in the hazard of the disease compared with the baseline hazard level.
  - At any time t, the hazard of the exposed ($h_1(t)$) is a multiple of the baseline hazard level ($h_0(t)$)
- Implication: the model provides one hazard ratio for the entire follow-up period.

Berkeley School of Public Health

And let's talk more about this concept of proportional hazards. It's a key assumption of Cox proportional hazards model. Let's recall that hazards can't be directly calculated because they're defined for an infinitely small time interval. So here's the formula hazard at time t is equal to the probability that an event occurs in the interval between t and t plus delta t, conditional on being alive at t, divided by the delta t.

Usually, this time interval is so small that we can't get the exact hazard at a particular time, but if the relative hazard can be assumed to be constant over the interval, then we can model the hazard without having to actually estimate the hazard itself. So these models assess whether the exposure is associated with a fixed relative increase in the hazard of disease compared with a baseline hazard level.

It assumes that, at anytime t, the hazard to the exposed, which is each one of t, is a multiple of the baseline hazard, h0 of t. So baseline hazard is another way of saying the hazard among the unexposed, and h1 of t is the hazard among the exposed at time t. The implication is that, because the model makes these assumptions about the hazard being proportional over the duration of follow-up, the model only estimates one hazard ratio for the whole follow-up period.



What does this look like graphically? The plot on the left shows us hazard on the y-axis, time on the x-axis. The red curve is for the exposed, and the blue curve is for the unexposed. And in this curve, the hazard is following the same pattern over time for the exposed and unexposed, but one curve is just shifted up slightly higher than the other, but the pattern is the same. And so we call this proportional hazards. And in this situation, the model's assumption of proportionality is valid.

Now, in the figure on the right, we show something similar, but we see that the curves follow a different shape. And so the ratio of the hazard between the exposed and the unexposed varies over time. So the hazard is not proportional to each other over time. And thus, we need to do something special to our model to validly estimate the hazard ratio.

So in other words, we need to stratify by periods of follow-up time so that we meet this proportionality assumption within each stratum. Now, this is tough in a graph like this because the patterns are just so different, but if there's a little bit of variation, we can just stratify by time and then hope that, within that particular band of time, the pattern is similar to each other enough that the assumption of proportionality is reasonable.

Obtaining the hazard ratio from a Cox proportional hazards model

1. Start with the Cox model specification

$$\ln\left(h(t|X=x)\right) = \ln h_0(t) + cx$$

Berkeley School of Public Health

Here's how we obtain a hazard ratio from a Cox proportional hazard model coefficient. We start with our model specification…

Obtaining the hazard ratio from a Cox proportional hazards model

1. Start with the Cox model specification

$$\ln\big(h(t|X=x)\big) = \ln h_0(t) + cx$$

2. We use the notation from the model to write the formula for the hazard ratio

$$\ln\left(\frac{h(t|X=1)}{h(t|X=0)}\right) = \ln\big(h(t|X=1)\big) - \ln\big(h(t|X=0)\big)$$

Berkeley School of Public Health

…and then we're going to use our notation from the model to write the formula for the hazard ratio. And so if we take the log of the hazard at time t when X is 1 and the hazard of time t when X is 0, that's a log hazard ratio. And this can be rewritten using log rules as the difference in the log hazard when X is 1 and when X is 0.

Obtaining the hazard ratio from a Cox proportional hazards model

1. Start with the Cox model specification

$$\ln\big(h(t|X=x)\big) = \ln h_0(t) + cx$$

2. We use the notation from the model to write the formula for the hazard ratio

$$\ln\left(\frac{h(t|X=1)}{h(t|X=0)}\right) = \ln\big(h(t|X=1)\big) - \ln\big(h(t|X=0)\big)$$

3. Next we fill in the coefficients for each term in the difference.

$$= (\ln h_0(t) + c\cdot(1)) - (\ln h_0(t) + c\cdot(0))$$

Berkeley School of Public Health

And then we can plug in the coefficients from the formula at the top of the slide, our log baseline hazard at time t plus c times 1 minus our log baseline hazard at time t plus c times 0.

Obtaining the hazard ratio from a Cox proportional hazards model

1. Start with the Cox model specification

$$\ln\big(h(t|X=x)\big) = \ln h_0(t) + cx$$

2. We use the notation from the model to write the formula for the hazard ratio

$$\ln\left(\frac{h(t|X=1)}{h(t|X=0)}\right) = \ln\big(h(t|X=1)\big) - \ln\big(h(t|X=0)\big)$$

3. Next we fill in the coefficients for each term in the difference.

$$= (\ln h_0(t) + c\cdot(1)) - (\ln h_0(t) + c\cdot(0))$$

4. After simplifying, we see that c is equal to the log hazard ratio

$$= \ln\left(\frac{h(t|X=1)}{h(t|X=0)}\right) = c$$

Berkeley School of Public Health

And that leaves us with c. So c is our log hazard ratio.

Obtaining the hazard ratio from a Cox proportional hazards model

1. Start with the Cox model specification

$$\ln\big(h(t|X=x)\big) = \ln h_0(t) + cx$$

2. We use the notation from the model to write the formula for the hazard ratio

$$\ln\left(\frac{h(t|X=1)}{h(t|X=0)}\right) = \ln\big(h(t|X=1)\big) - \ln\big(h(t|X=0)\big)$$

3. Next we fill in the coefficients for each term in the difference.

$$= (\ln h_0(t) + c\cdot(1)) - (\ln h_0(t) + c\cdot(0))$$

4. After simplifying, we see that c is equal to the log hazard ratio

$$= \ln\left(\frac{h(t|X=1)}{h(t|X=0)}\right) = c$$

5. Exponentiating, e^c is equal to the hazard ratio

$$\frac{h(t|X=1)}{h(t|X=0)} = e^c$$

Berkeley School of Public Health

And if we exponentiate, we see that the exponential of c is our hazard ratio comparing the hazard when X is 1 to the hazard when X is 0.

Example of Cox proportional hazards model

$$\ln(\text{hazard ratio}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

Table 7-21 Results of a Cox proportional regression analysis of binary and continuous predictors of coronary heart disease (CHD) incidence in the Washington County cohort of the Atherosclerosis Risk in Communities (ARIC) Study, ages 45-64 years at baseline, 1987-1994

Variable (unit)	Cox regression coefficient	Hazard ratio
Gender (male = 1, female = 0)	1.2569	3.52
Smoking (yes = 1, no = 0)	0.7045	2.02
Age (1 year)	0.0120	1.012
Systolic blood pressure (1 mm Hg)	0.0152	1.015
Serum cholesterol (1 mg/dL)	0.0067	1.007
Body mass index (1 kg/m^2)	0.0237	1.024

Binary covariate
Hazard ratio for CHD for males compared to females is e1.2569 = 3.52

Continuous covariate
Hazard ratio for CHD for a 1 mm Hg increase in systolic blood pressure is $e^{0.0152} = 1.015$

Berkeley School of Public Health

Here's an example from the Szklo textbook. What we can see is this table is looking at the Cox proportional regression analysis of an outcome of coronary heart disease incidence, and different risk factors. So gender, smoking, age, blood pressure, et cetera. And in this column that says Cox regression coefficient, this is what's returned from the model.

Gender was operationalized as a binary variable. And so for a binary covariate, we can exponentiate the regression coefficient 1.2569, and that gives us a hazard ratio of 3.52. So that's the hazard ratio for coronary heart disease for males compared to females.

For a continuous covariate, such as systolic blood pressure, we can look at a 1 millimeter of mercury increase in blood pressure. And that's shown here. This Cox regression coefficient is 0.0152. If we exponentiate that, we get 1.015. That's our hazard ratio for a 1 millimeter of mercury increase.

Summary of key points

- **Identify which type of model is used depending on the type of dependent and independent variable**
  - **Time-to-event outcome**: Cox proportional hazards model
- **How to interpret the coefficients from commonly used regression models**
  - $\ln\left(h(t|X=x)\right) = \ln h_0(t) + cx$
  - The exponentiated coefficient *c* (e^c) is the hazard ratio comparing the hazard when x = 1 to x = 0
- **Articulate certain assumptions of these models**
  - Called a proportional hazards model because it assumes that the ratio of the hazards is constant across the time interval.

Berkeley School of Public Health

To summarize, we've really only covered this at a high level, but when we have time-to-event outcome data, Cox proportional hazards models are an appropriate type of model. And here is the model specification. These coefficient c can be exponentiated to obtain the hazard ratio comparing the hazard when x is 1 to X is 0.

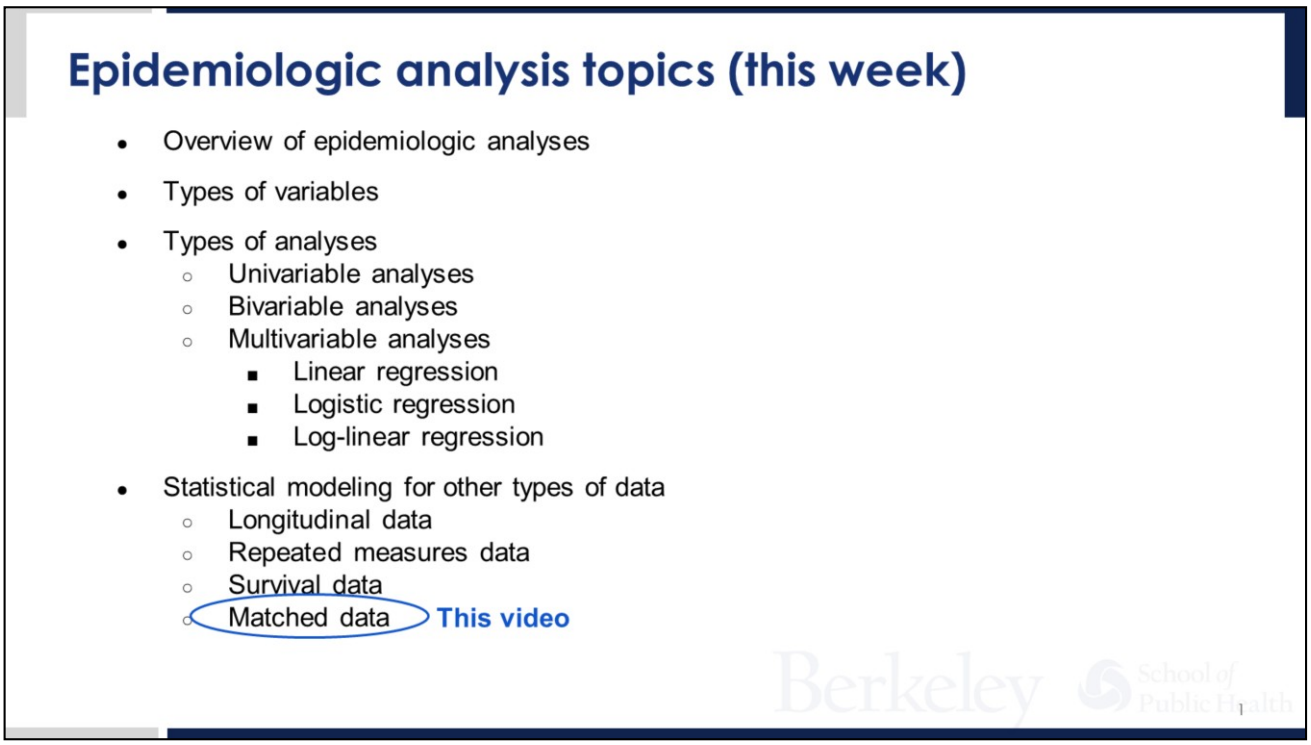
And the key assumption of this model is that it's called a proportional hazards model because it's assuming that the ratio of the hazards is constant across the time interval. And so we need to be very careful when that assumption is not valid. It's worth also briefly mentioning that this model assumes no clustering and no autocorrelation, like most of the other models we've talked about in this course.

Berkeley School of Public Health

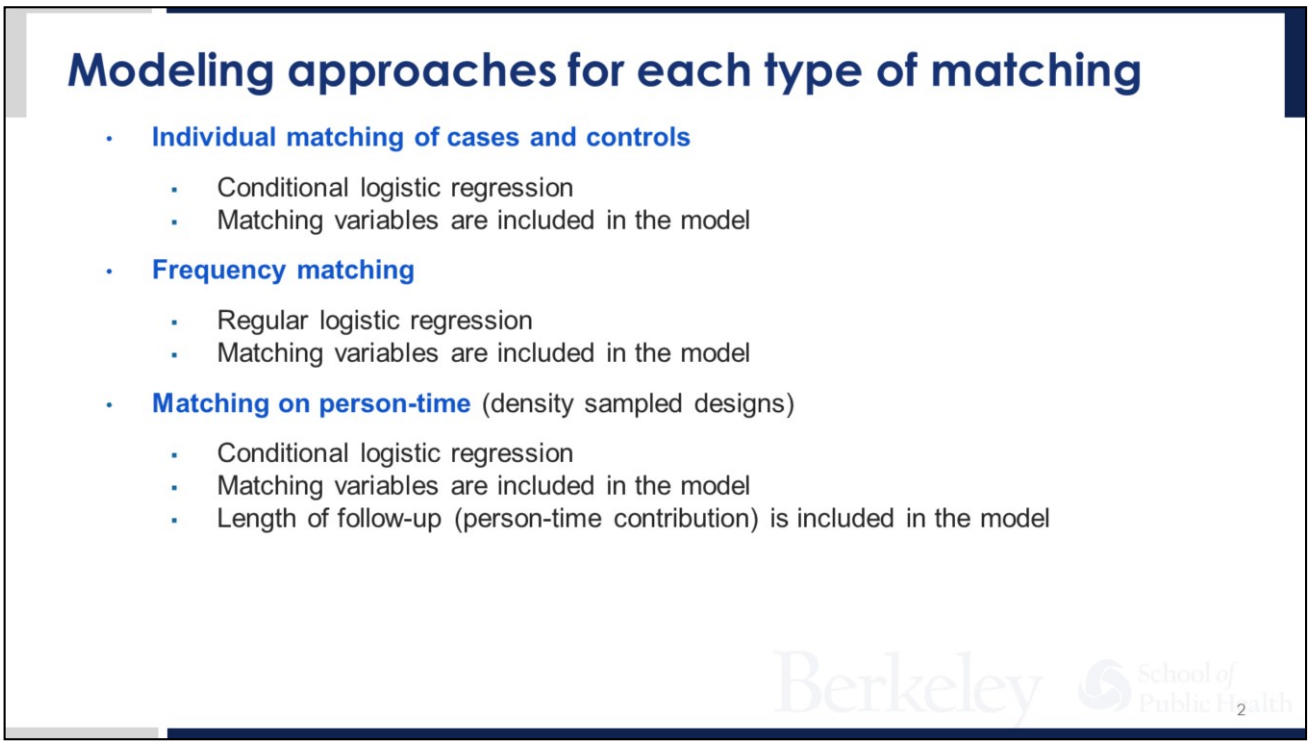
Models for matched case-control data

PHW250 B – Andrew Mertens

Andrew Mertens: This video will briefly talk about regression models formatched case control data.



So this is our last topic in the first portion of our epidemiologic analysis week.



We have several different types of matching we can use in a case control study. We have individual matching of cases and controls, frequency matching, and matching on person-time. And each of these different matching approaches requires a different form of analysis.

So in this video, we'll be talking about conditional logistic regression, which we can use when we have individual match data. And this regression modeling approach also requires us to condition on the matching variables in our model. If we're doing frequency matching with case control data, we could actually use a logistic regression model, a regular logistic regression model, as long as we include the matching variables in the model.

And then if we're doing a density sample design that matches on person-time, we can use conditional logistic regression. We'll want to also include the matching variables in the model. And length of follow-up, since it's what's matched on, as well, must also be included in the model. So our focus here is really going to be on conditional logistic regression, which we can use for individually matched case control studies, and those that match on person-time.

Recap: Odds ratio formula for matched pair data - case-control study

- When case and control both either exposed or unexposed we get no information about the exposure disease relation.
- The only information is in the discordant pairs.
- **Odds Ratio = B / C**
  - (See Jewell pg 261 for the derivation)
- **Intuition**: B and C are the pairs in which we have variation in the exposure – if no variation in exposure, cannot look at relation between exposure and disease
- This formula does not allow us to adjust for potential confounders

Concordant pairs
Discordant pairs

Table 16.3 *Exposure patterns in the four types of matched pairs*

		D	$\bar{D}$	
(1)	E	1	1	2
	$\bar{E}$	0	0	0
		1	1	

		D	$\bar{D}$	
(2)	E	1	0	1
	$\bar{E}$	0	1	1
	E	1	1	

		D	$\bar{D}$	
(3)	E	0	1	1
	$\bar{E}$	1	0	1
		1	1	

		D	$\bar{D}$	
(4)	E	0	0	0
	$\bar{E}$	1	1	2
	E	1	1	

Organization of matched pair data

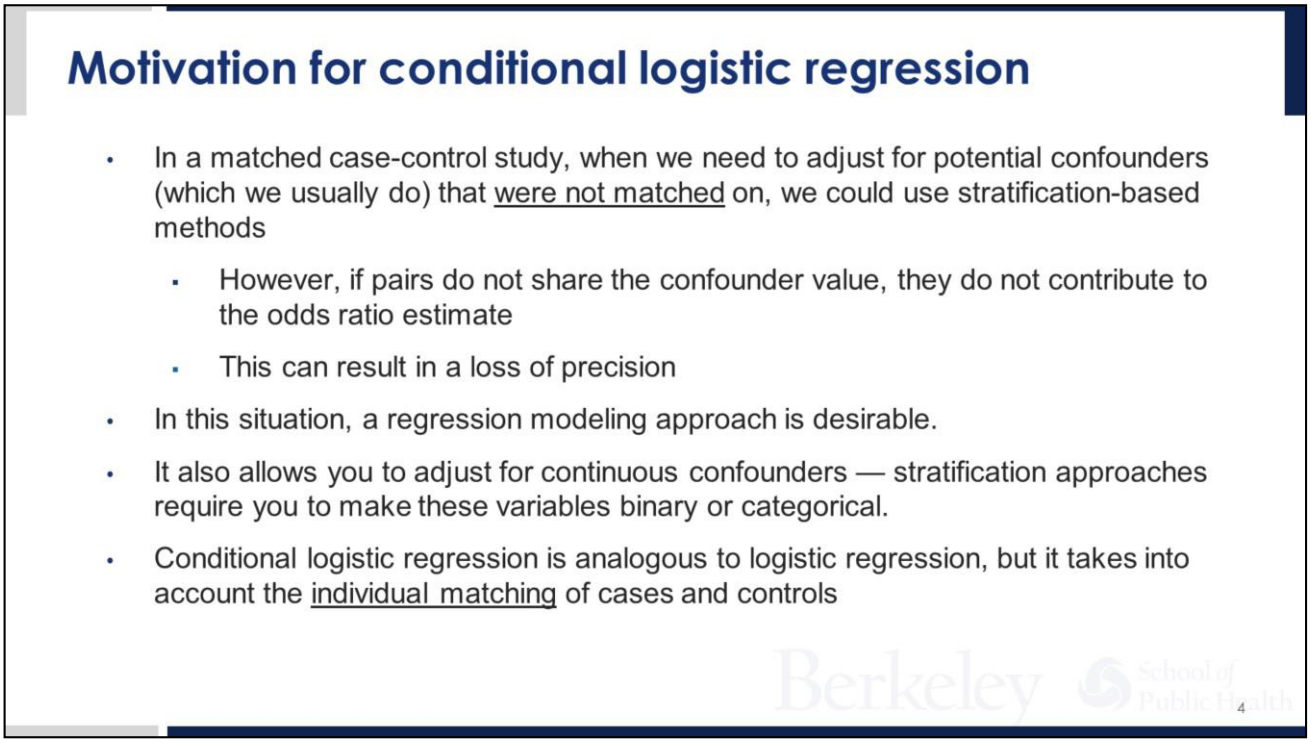
		Control	
		E	$\bar{E}$
Case	E	A	B
	$\bar{E}$	C	D

N

Jewell. *Statistics for Epidemiology*. 2004. 3

To briefly recap, we have a special odds ratio formula for case control studies that use individual pair matching. And that's because, if we recall from this figure on the right, it's the discordant pairs that really provide information. When case and controls both are exposed or unexposed, we get no new information about the exposure/disease relationship.

So the odds ratio formula in this type of study is B over C, instead of the usual AD over BC. And that intuition behind that is that the B and the C cells are those in which the pairs have variation in the exposure. This formula doesn't allow us to adjust for confounders, and so our focus if this video is how can we estimate the odds ratio adjusting for confounders in this type of study.



OK, so I just sort of recapped this motivation for conditional logistic regression. But this arises when we have a individually matched case control study and we want to adjust for confounders that were not matched on. We could use stratification-based methods, but if pairs don't share the confounder value, then they don't contribute to the odds ratio estimate. And this can result in a loss of precision. So that would affect the width of our confidence intervals by making them bigger.

So what do we mean by this? Well, if one of our confounders is age, and we didn't match on age, we matched on something else, if the ages of the two people in the pair are very different, they don't share that confounder value. And as a result, we can't include their data in an odds ratio estimate that's adjusted for age without using a model.

So this is where regression modeling is really helpful, because it allows you to adjust for these kinds of variables, and also for continuous confounders that would be difficult to use with a stratification-based approach, unless we were willing to make them into binary or categorical variables. Conditional logistic regression is analogous to logistic regression, but it takes into account the individual matching of cases and controls.

Conditional logistic regression

$$\text{logit}(E(Y|X=x, I=i)) = \underbrace{\beta_0^*}_{\text{Indicators for matching stratum}} + \underbrace{\beta_1 x}_{\text{Coefficients}} \qquad \beta_0^* = \beta_0 + \beta_i$$

- For a binary exposure X, β_1 = the difference in the log odds of the outcome Y when $X = 1$ and when $X = 0$
  - $e^{\beta 1}$ is the odds ratio comparing $X = 1$ to $X = 0$
- Intercept β_0^* is equal to $\beta_0 + \beta_i$, the sum of the overall intercept and each stratum-specific intercept.
- These odds ratios adjust for potential confounders as well as matching variables.
- This is analogous to including an indicator variable for each stratum of matching factors – in pair matching this would mean including an indicator variable for each pair.
- Must include matching factors as covariates in the model.
- Assumes no clustering or auto-correlation

Berkeley School of Public Health

5

So here's the model specification. And it looks quite a bit like logistic regression, but there's two key differences. The first is that we're taking the logit of the mean outcome conditional on x, as well as on the indicators for the matching stratum. So basically, if it's individually matched, each pair is in its own stratum. And so we have an indicator for each pair in the study.

And then our coefficients are slightly different. So we have this beta 0 star. And that's equal to the beta 0, the intercept, plus beta i, the intercept for the stratum-specific pair. For a binary exposure x, beta 1 is the difference in the log odds of the outcome y when x is 1 and when x is 0. And if we exponentiate beta 1, the odds ratio comparing x1 to x0. So it's very analogous to traditional logistic regression, but it's basically just taking into account the fact that we use matched data.

Conditional logistic regression

$$\text{logit}(E(Y|X = x, I = i)) = \beta_0^* + \beta_1 x \qquad \beta_0^* = \beta_0 + \beta_i$$

Indicators for matching stratum — Coefficients

- For a binary exposure X, β_1 = the difference in the log odds of the outcome Y when $X = 1$ and when $X = 0$
  - $e^{\beta 1}$ is the odds ratio comparing $X = 1$ to $X = 0$
- Intercept β_0^* is equal to $\beta_0 + \beta_i$, the sum of the overall intercept and each stratum-specific intercept.
- These odds ratios adjust for potential confounders as well as matching variables.
- This is analogous to including an indicator variable for each stratum of matching factors – in pair matching this would mean including an indicator variable for each pair.
- Must include matching factors as covariates in the model.
- Assumes no clustering or auto-correlation

Berkeley School of Public Health

5

The odds ratios estimated in a conditional logistic regression model adjust for potential confounders, as well as the matching variables. And recall that, when we covered matching in the previous unit, we learned that, in case control studies, matching actually introduces confounding that needs to be accounted for in the statistical analysis.

So what it's essentially doing is analogous to including an indicator variable for each stratum in the model. And in order for this to work, we do have to include these matching factors as covariates. It's worth briefly mentioning that this also assumes no clustering or autocorrelation, as most of the other models we've covered in this unit also have assumed.

Summary of key points

- **Identify which type of model is used depending on the type of dependent and independent variable**
  - **Individually matched case-control study with binary outcome**: conditional logistic regression
- **How to interpret the coefficients**
  - $\text{logit}(E(Y|X = x, I = i)) = \beta_0^* + \beta_1 x$
  - The exponentiated coefficient β_1 ($e^{\beta 1}$) is the odds ratio comparing x = 1 to x = 0
- **Articulate certain assumptions of these models**
  - No clustering or auto-correlation

Berkeley School of Public Health

To summarize, when we have individually matched case control study data with a binary outcome-- and this is also true for studies that match on person-time-- we need to use conditional logistic regression. And here's the model specification. And the exponentiated coefficient beta 1 is equal to the odds ratio comparing x equals 1 and x equals 0. And these models assume that there is no clustering or autocorrelation of the data.

Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural Bangladesh: a cluster randomised controlled trial

Stephen P Luby, Mahbubur Rahman, Benjamin F Arnold, Leanne Unicomb, Sania Ashraf, Peter J Winch, Christine P Stewart, Farzana Begum, Faruqe Hussain, Jade Benjamin-Chung, Elli Leontsini, Abu M Naser, Sarker M Parvez, Alan E Hubbard, Audrie Lin, Fosiul A Nizame, Kaniz Jannat, Ayse Ercumen, Pavani K Ram, Kishor K Das, Jaynal Abedin, Thomas F Clasen, Kathryn G Dewey, Lia C Fernald, Clair Null, Tahmeed Ahmed, John M Colford Jr

Summary

Background Diarrhoea and growth faltering in early childhood are associated with subsequent adverse outcomes. We aimed to assess whether water quality, sanitation, and handwashing interventions alone or combined with nutrition interventions reduced diarrhoea or growth faltering.

Methods The WASH Benefits Bangladesh cluster-randomised trial enrolled pregnant women from villages in rural Bangladesh and evaluated outcomes at 1-year and 2-years' follow-up. Pregnant women in geographically adjacent clusters were block-randomised to one of seven clusters: chlorinated drinking water (water); upgraded sanitation (sanitation); promotion of handwashing with soap (handwashing); combined water, sanitation, and handwashing; counselling on appropriate child nutrition plus lipid-based nutrient supplements (nutrition); combined water, sanitation, handwashing, and nutrition; and control (data collection only). Primary outcomes were caregiver-reported diarrhoea in the past 7 days among children who were in utero or younger than 3 years at enrolment and length-for-age *Z* score among children born to enrolled pregnant women. Masking was not possible for data collection, but analyses were masked. Analysis was by intention to treat. This trial is registered at ClinicalTrials.gov, number NCC01590095.

Findings Between May 31, 2012, and July 7, 2013, 5551 pregnant women in 720 clusters were randomly allocated to one of seven groups. 1382 women were assigned to the control group; 698 to water; 696 to sanitation; 688 to handwashing; 702 to water, sanitation, and handwashing; 699 to nutrition; and 686 to water, sanitation, handwashing, and nutrition. 331 (6%) women were lost to follow-up. Data on diarrhoea at year 1 or year 2 (combined) were available for 14 425 children (7331 in year 1, 7094 in year 2) and data on length-for-age *Z* score in year 2 were available for 4584 children (92% of living children were measured at year 2). All interventions had high adherence. Compared with a prevalence of 5·7% (200 of 3517 child weeks) in the control group, 7-day diarrhoea prevalence was lower among index children and children under 3 years at enrolment who received sanitation (61 [3·5%] of 1760; prevalence ratio 0·61, 95% CI 0·46–0·81), handwashing (62 [3·5%] of 1795; 0·60, 0·45–0·80), combined water, sanitation, and handwashing (74 [3·9%] of 1902; 0·69, 0·53–0·90), nutrition (62 [3·5%] of 1766; 0·64, 0·49–0·85), and combined water, sanitation, handwashing, and nutrition (66 [3·5%] of 1861; 0·62, 0·47–0·81); diarrhoea prevalence was not significantly lower in children receiving water treatment (90 [4·9%] of 1824; 0·89, 0·70–1·13). Compared with control (mean length-for-age *Z* score −1·79), children were taller by year 2 in the nutrition group (mean difference 0·25 [95% CI 0·15–0·36]) and in the combined water, sanitation, handwashing, and nutrition group (0·13 [0·02–0·24]). The individual water, sanitation, and handwashing groups, and combined water, sanitation, and handwashing group had no effect on linear growth.

Interpretation Nutrient supplementation and counselling modestly improved linear growth, but there was no benefit to the integration of water, sanitation, and handwashing with nutrition. Adherence was high in all groups and diarrhoea prevalence was reduced in all intervention groups except water treatment. Combined water, sanitation, and handwashing interventions provided no additive benefit over single interventions.

Funding Bill & Melinda Gates Foundation.

Copyright © The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

Lancet Glob Health 2018; 6: e302–15

Published Online January 29, 2018 http://dx.doi.org/10.1016/S2214-109X(17)30490-4

See Comment page e236

See Articles page e316

Division of Infectious Diseases and Geographic Medicine, Stanford University, Stanford, CA, USA (Prof S P Luby MD); **International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh** (M Rahman MBBS, L Unicomb PhD, S Ashraf MPH, F Begum MPH, F Hussain MSS, A M Naser MBBS, S M Parvez MPH, F A Nizame MA, K Jannat MBBS, K K Das MS, J Abedin MS, T Ahmed PhD); **School of Public Health University of California Berkeley, Berkeley, CA, USA** (B F Arnold PhD, J Benjamin-Chung PhD, Prof A L Hubbard PhD, A Lin PhD, A Ercumen PhD, Prof L C Fernald, Prof J M Colford Jr MD); **Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA** (Prof P J Winch MD, E Leontsini MD); **Department of Nutrition, University of California Davis, Davis, CA, USA** (C P Stewart PhD, Prof K G Dewey PhD); **School of Public Health and Health Professions, University of Buffalo, Buffalo, NY, USA** (P K Ram MD); **and Rollins School of Public Health, Emory University, Atlanta, GA, USA** (Prof T F Clasen PhD, C Null PhD)

Correspondence to: Dr Stephen P Luby, Infectious Diseases and Geographic Medicine, Stanford University, Stanford, CA 94305 **sluby@stanford.edu**

Introduction

Over 200 million children born in low-income countries are at risk of not reaching their development potential.[1] Poor linear growth in early childhood is a marker for chronic deprivation that is associated with increased mortality, impaired cognitive development, and reduced adult income.[2] Nutrition-specific interventions have been shown to improve child growth

Research in context

Evidence before this study

Although malnutrition and diarrhoeal disease in children have been known for decades to impair child health and growth, there is little evidence on interventions that are successful at improving growth and reducing diarrhoea. Several observational analyses noted positive associations between improvements in water, sanitation, and handwashing conditions and child growth, but at the time this study was conceived there were no published randomised controlled trials specifically powered to evaluate the effect of such interventions on child growth as a primary outcome. Subsequent published trials of sanitation interventions have reported mixed results. Systematic reviews of complementary feeding interventions have reported small but significant improvements in child growth. More recent evidence from lipid-based nutrient supplementation trials has been mostly consistent with these earlier systematic reviews. Chronic enteric infection might affect children's capacity to respond to nutrients; however, we found no published studies comparing the effect on child growth of nutritional interventions alone versus nutritional interventions plus water, sanitation, and handwashing interventions. Although many programmatic interventions target multiple pathways of enteric pathogen transmission, systematic reviews have found no greater reduction in diarrhoea with combined versus single water, sanitation, and handwashing interventions. There is little direct evidence comparing interventions that target a single versus multiple pathways. Only three randomised controlled trials compared single versus combined interventions in comparable populations at the same time. None of these trials found a significant reduction in diarrhoea among children younger than 5 years who received combined versus the most effective single intervention.

Added value of this study

This trial was designed to compare the effects of individual and combined water quality, sanitation, hygiene, and nutrient supplementation plus infant and young child feeding counselling interventions on diarrhoea and growth when given to infants and young children in a setting where child growth faltering was common. The trial had high intervention adherence, low attrition, and ample statistical power to detect small effects. Children receiving interventions with nutritional components had small growth benefits compared with those in the control cluster. Water quality, sanitation, and handwashing interventions did not improve child growth, neither when delivered alone nor when combined with the nutritional interventions. Children receiving sanitation, handwashing, nutrition, and combined interventions had less reported diarrhoea. Combined interventions showed no additional reduction in diarrhoea beyond single interventions.

Implications of all the available evidence

The modest improvements observed in growth faltering with nutritional supplementation and counselling are consistent with other trials that report similar levels of efficacy in some contexts. By contrast to observational studies that report an association between growth faltering and water, sanitation, and hygiene assessments, this intervention trial provides no evidence that household drinking water quality, sanitation, or handwashing interventions consistently improve growth. This trial further supports findings from smaller trials that combined individual water, sanitation, and handwashing interventions are not consistently more effective in the prevention of diarrhoea than are single interventions.

but they have only corrected a small part of the total growth deficit.[3]

Environmental enteric dysfunction is an abnormality of gut function that might explain why most nutrition interventions fail to normalise early childhood growth.[4] Environmental contaminants are thought to induce the chronic intestinal inflammation, loss of villous surface area, and impaired barrier function that combine to impair food and nutrient uptake. Several observational studies find that children living in communities where most people have access to a toilet are less likely to be stunted than are children who live in communities where open defecation is more common.[5] Intervention trials to reduce exposure to human faeces can resolve questions of confounding in the relationship between toilet access and child growth and evaluate potential interventions. Improvements to drinking water quality, sanitation, and handwashing might improve the effectiveness of nutrition interventions and thereby help to tackle a larger portion of the observed growth deficit.

In addition to asymptomatic infections and subclinical changes to the gut, episodes of symptomatic diarrhoea accounted for about 500 000 deaths of children younger than 5 years in 2015.[6] Approaches to reduce diarrhoea include treated drinking water, improved sanitation, and increased handwashing with soap. Although funding a single intervention for a larger population might improve health more than multiple interventions that target a smaller population, data to inform such decisions are scarce.

Interventions that combine nutrition and water, sanitation, and handwashing might provide multiple benefits to children, but there is little evidence that directly compares the effects of individual and combined interventions on diarrhoea and growth of young children.[7,8]

We aimed to investigate whether individual water, sanitation, handwashing, or nutrition interventions can reduce linear growth faltering; to assess whether combined water, sanitation, and handwashing interventions are more effective at reducing diarrhoea than individual interventions; and to investigate whether the combination of water, sanitation, handwashing, and nutrition interventions reduces growth faltering more

than each individual intervention. A companion trial in Kenya evaluated the same objectives.[9]

Methods

Study design

The WASH Benefits Bangladesh study was a cluster-randomised trial conducted in rural villages in Gazipur, Kishoreganj, Mymensingh, and Tangail districts of Bangladesh (appendix p 2). We grouped pregnant women who lived near enough to each other into a cluster to allow delivery of interventions by a single community promoter. We hypothesised that the interventions would improve the health of the index child in each household. Each measurement round lasted about 1 year and was balanced across treatment arms and geography to minimise seasonal or geographical confounding when comparing outcomes across groups. We chose areas with low groundwater iron and arsenic (because these affect chlorine demand) and where no major water, sanitation, or nutrition programmes were ongoing or planned by the government or large non-government organisations. The study design and rationale have been published previously.[10]

The latrine component of the sanitation intervention was a compound level intervention. The drinking water and handwashing interventions were household level interventions. The nutrition intervention was a child-specific intervention. We assessed the diarrhoea outcome among all children in the compound who were younger than 3 years at enrolment, which could underestimate the effect of interventions targeted only to index households (drinking water, and handwashing) or index children (nutrition). After the study results were unmasked, we analysed diarrhoea prevalence restricted to index children (ie, children directly targeted by each intervention).

The study protocol was approved by the Ethical Review Committee at The International Centre for Diarrhoeal Disease Research, Bangladesh (PR-11063), the Committee for the Protection of Human Subjects at the University of California, Berkeley (2011-09-3652), and the institutional review board at Stanford University (25863).

Participants

Rural households in Bangladesh are usually organised into compounds where patrilineal families share a common courtyard and sometimes a pond, water source, and latrine. Research assistants visited compounds in candidate communities. If compound residents reported no iron taste in their drinking water nor iron staining of their water storage vessels,[11] and if a woman reported being in the first two trimesters of pregnancy, research assistants recorded the global positioning system coordinates of her household. We reviewed maps of plotted households and made clusters of eight expectant women who lived close enough to each other for a single community promoter to readily walk to each compound. We used a 1 km buffer around each cluster to reduce the potential for spillover between clusters (median buffer distance 2·6 km [IQR 1·8–3·7]). Participants gave written informed consent before enrolment.

The in utero children of enrolled pregnant women (index children) were eligible for inclusion if their mother was planning to live in the study village for the next 2 years, regardless of where she gave birth. Only one pregnant woman was enrolled per compound, but if she gave birth to twins, both children were enrolled. Children who were younger than 3 years at enrolment and lived in the compound were included in diarrhoea measurements.

See Online for appendix

Randomisation and masking

Clusters were randomly allocated to treatment using a random number generator by a coinvestigator at University of California, Berkeley (BFA). Each of the eight geographically adjacent clusters was block-randomised to the double-sized control arm or one of the six interventions (water; sanitation; handwashing; water, sanitation, and handwashing; nutrition; or water, sanitation, handwashing, and nutrition). Geographical matching ensured that arms were balanced across locations and time of measurement.

Interventions included distinct visible components so neither participants nor data collectors were masked to intervention assignment, although the data collection and intervention teams were different individuals. Two investigators (BFA and JBC) did independent, masked statistical analyses from raw datasets to generate final estimates, with the true group assignment variable replaced with a re-randomised uninformative assignment variable. The results were unmasked after all analyses were replicated.

Procedures

We used the Integrated Behavioural Model for Water Sanitation and Hygiene to develop the interventions over 2 years of iterative testing and revision.[12] This model addresses contextual, psychosocial, and technological factors at the societal, community, interpersonal, individual, and habitual levels.

Community promoters delivered the interventions. These promoters were women who had completed at least 8 years of formal education, lived within walking distance of an intervention cluster, and passed a written and oral examination. Promoters attended multiple training sessions, including quarterly refreshers. Training addressed technical intervention issues, active listening skills, and strategies for the development of collaborative solutions with study participants. Promoters were instructed to visit intervention households at least once weekly in the first 6 months, and then at least once every 2 weeks. Promoters who delivered more complex interventions received longer formal training (table 1).

	Water	Sanitation	Handwashing	Nutrition	Water, sanitation, and handwashing	Water, sanitation, handwashing, and nutrition
Training*						
Duration of initial training	4 days	4 days	4 days	5 days	5 days	9 days
Duration of refresher training	1 day	1 day	1 day	1 day	1 day	1 day
Implementation†						
Technology and supplies provided	Insulated storage container for drinking water; Aquatabs (Medentech, Ireland)	Sani-scoop; potty; double-pit pour flush improved latrine	Handwashing station; storage bottle for soapy water; laundry detergent sachets for preparation of soapy water	LNS (Nutriset, France); storage container for LNS	Same as individual water, sanitation, and handwashing interventions	Same as individual water, sanitation, handwashing, and nutrition interventions
Key behavioural recommendations delivered by promoters	Targeted children drink treated, safely stored water	Family use double pit latrines; potty train children; safely dispose of faeces into latrine or pit	Family wash hands with soap after defecation and during food preparation	Exclusive breastfeeding up to 180 days; introduce diverse complementary food at 6 months; feed LNS from 6–24 months	Same as individual water, sanitation, and handwashing interventions	Same as individual water, sanitation, handwashing, and nutrition interventions
Population targeted	Children younger than 5 years living in index households	Whole compound for latrines; index households for potty training and safe faeces disposal	Residents of index households	Index children (targeted through mother)	Same as individual water, sanitation, and handwashing interventions	Same as individual water, sanitation, handwashing, and nutrition interventions
Emphasis during visits after refresher training	Safe storage of water, children drink only treated and safely stored water	Latrine cleanliness; maintenance; pit switching	Handwashing before food preparation	Dietary diversity during complementary feeding; provide LNS even if child is unwell	Same as individual water, sanitation, and handwashing interventions	Same as individual water, sanitation, handwashing, and nutrition interventions

LNS=lipid-based nutrient supplement. *Common across all arms: roles and responsibilities, introduction to behaviour change principles, and interpersonal and counselling communication skills. Specific for each intervention: technology installation and use, onsite demonstration of use in the home, resupplying and restocking, problem solving challenges to technology use, and adoption of behaviours. Refresher training was done 12–15 months after start of intervention; content was based on analysis of reasons for gap between goals for uptake and actual uptake and addressed reasons for low uptake (specific to each intervention). †Promoter visits were intended to teach participants how to use technologies and how to use and restock products; arrange for social support; communicate benefits of use and practice and changes in social norms; congratulate and encourage; problem-solve as needed; and inspire. Techniques used included counselling via flipcharts and cue cards, onsite demonstrations of technologies and products, video dramas, storytelling, games, and songs. Promoter's guides detailed the visit objective, target audience, and the specific steps and materials to be used.

Table 1: **Training of community health promoters and content of home visits for the six intervention groups**

After the hardware was installed, household visits involved promoters greeting target household members, checking for the presence and functionality of hardware and signs of use, observing any of the recommended practices, and then following a structured plan for that visit. For each visit, a promoter's guide detailed the visit objective, the target audience, the specific steps, and materials to be used. Discussions, video dramas, storytelling, games, songs, and training on hardware maintenance were included in different visits. The breadth of the curriculum varied by the complexity of the intervention. Promoters delivering combined interventions were expected to spend sufficient time to cover all of the behavioural objectives with target households. Promoters did not visit control households. Promoters received a monthly stipend equivalent to US$20, comparable to the local compensation for 5 days of agricultural labour.

The water intervention, which was modelled on a successful intervention from a previous trial,[11] provided a 10 L vessel with a lid, tap, and regular supply of sodium dichloroisocyanurate tablets (Medentech, Wexford, Ireland) to the household of index children. Households were encouraged to fill the vessel, add one 33 mg tablet, and wait 30 min before drinking the water. All household members, but especially children younger than 5 years, were encouraged to drink only chlorine-treated water. Non-index households in the compound did not receive the water intervention.

The latrine component of the sanitation intervention targeted all households in the compound. All latrines that did not have a slab, a functional water seal, or a construction that prevented surface runoff of a faecal stream into the community were replaced. If the index household did not have their own latrine, the project built one. The standard project intervention latrine was a double pit latrine with a water seal.[13] Each pit had five concrete rings that were 0·3 m high. When the initial pit filled, the superstructure and slab could be moved to the second pit. In the less than 2% of cases where there was insufficient space for a second pit or the water table was too high for a pit that was 1·5 m deep, the design was adapted. Nearly all households (99%) provided labour and modest financial contributions towards the construction of the latrines. All households in sanitation intervention compounds also received a sani-scoop, which is a hand tool for the removal of faeces from the compound,[14] and child potties if they had any children younger than 3 years.[15] Promoters encouraged mothers to teach their children to use the potties, to safely dispose of faeces in latrines, and to regularly remove animal and human faeces from the compound.

The handwashing intervention targeted households with index children. These households received

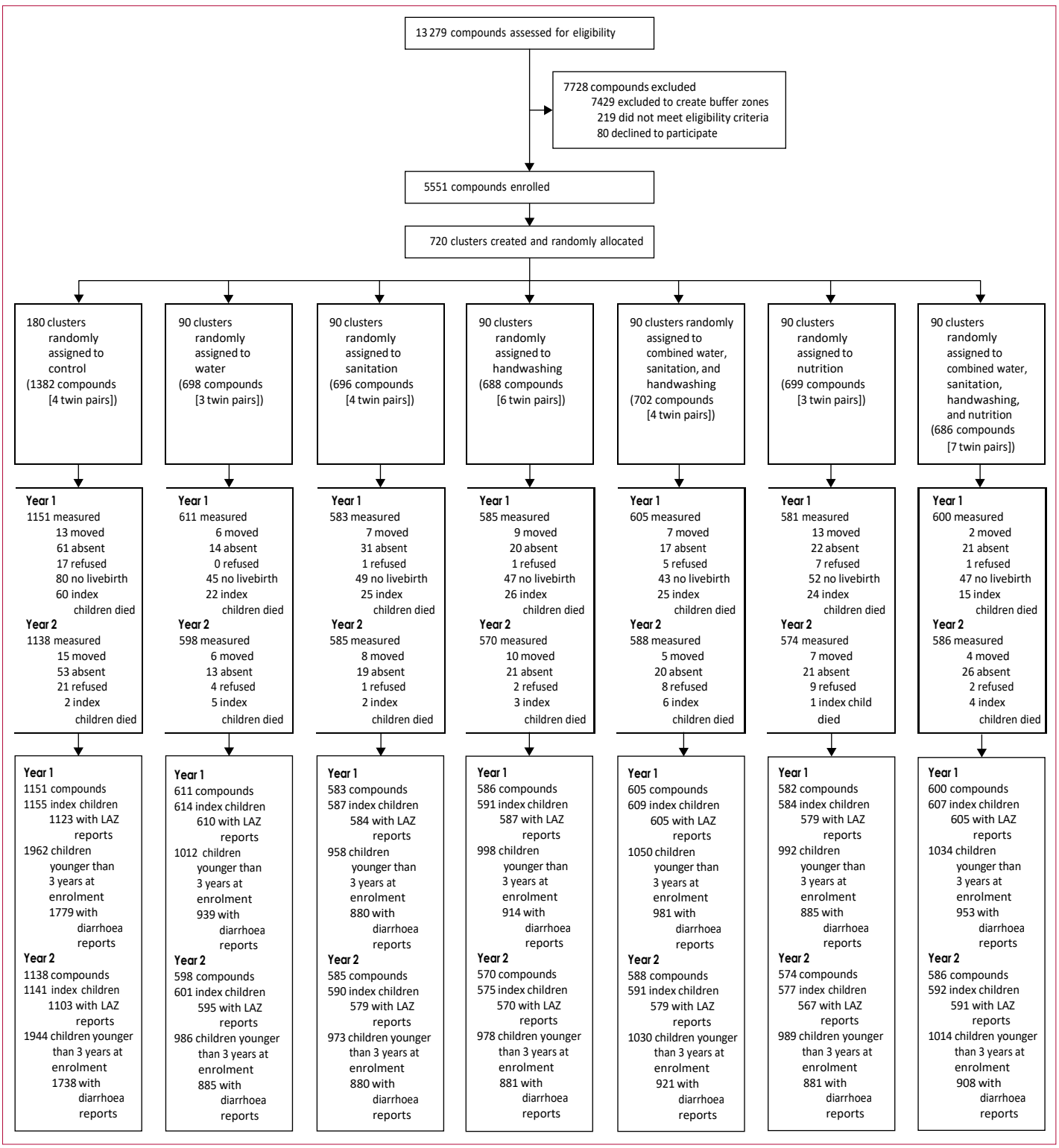


Figure 1: **Trial profile and analysis populations for primary outcomes**
LAZ=length-for-age *Z* scores.

	Control (n=1382)	Water treatment (n=698)	Sanitation (n=696)	Handwashing (n=688)	Water, sanitation, and handwashing (n=702)	Nutrition (n=699)	Water, sanitation, and handwashing, and nutrition (n=686)
Maternal							
Age (years)	24 (5)	24 (5)	24 (5)	24 (5)	24 (5)	24 (5)	24 (6)
Years of education	6 (3)	6 (3)	6 (3)	6 (3)	6 (3)	6 (3)	6 (3)
Paternal							
Years of education	5 (4)	5 (4)	5 (4)	5 (4)	5 (4)	5 (4)	5 (4)
Works in agriculture	414 (30%)	224 (32%)	204 (29%)	249 (36%)	216 (31%)	232 (33%)	207 (30%)
Household							
Number of people	5 (2)	5 (2)	5 (2)	5 (2)	5 (2)	5 (2)	5 (2)
Has electricity	784 (57%)	422 (60%)	408 (59%)	405 (59%)	426 (61%)	409 (59%)	412 (60%)
Has a cement floor	145 (10%)	82 (12%)	85 (12%)	55 (8%)	77 (11%)	67 (10%)	72 (10%)
Acres of agricultural land owned	0·15 (0·21)	0·14 (0·20)	0·14 (0·22)	0·14 (0·20)	0·15 (0·23)	0·16 (0·27)	0·14 (0·38)
Drinking water							
Shallow tubewell is primary water source	1038 (75%)	500 (72%)	519 (75%)	482 (70%)	546 (78%)	519 (74%)	504 (73%)
Has stored water at home	666 (48%)	353 (51%)	341 (49%)	347 (50%)	304 (43%)	301 (43%)	331 (48%)
Reported treating water yesterday	4 (0%)	1 (0%)	0 (0%)	1 (0%)	0 (0%)	0 (0%)	2 (0%)
Sanitation							
Daily defecating in the open							
Adult men	97 (7%)	39 (6%)	52 (8%)	64 (9%)	54 (8%)	59 (9%)	50 (7%)
Adult women	62 (4%)	18 (3%)	33 (5%)	31 (5%)	29 (4%)	39 (6%)	24 (4%)
Children aged 8 to <15 years	53 (10%)	25 (9%)	28 (9%)	43 (15%)	30 (10%)	23 (8%)	28 (10%)
Children aged 3 to <8 years	267 (38%)	141 (37%)	137 (38%)	137 (39%)	137 (38%)	129 (39%)	134 (37%)
Children aged 0 to <3 years	245 (82%)	112 (85%)	117 (84%)	120 (85%)	123 (79%)	128 (85%)	123 (88%)
Latrine							
Owned*	750 (54%)	363 (52%)	374 (54%)	372 (54%)	373 (53%)	377 (54%)	367 (53%)
Concrete slab	1251 (95%)	644 (95%)	610 (92%)	613 (93%)	620 (93%)	620 (94%)	621 (94%)
Functional water seal	358 (31%)	183 (31%)	177 (30%)	162 (28%)	152 (26%)	183 (31%)	155 (27%)
Visible stool on slab or floor	625 (48%)	350 (53%)	332 (52%)	335 (52%)	289 (44%)	331 (51%)	298 (46%)
Owned a child potty	61 (4%)	27 (4%)	28 (4%)	35 (5%)	27 (4%)	36 (5%)	30 (4%)
Human faeces observed in the							
House	114 (8%)	65 (9%)	56 (8%)	70 (10%)	48 (7%)	58 (8%)	49 (7%)
Child's play area	21 (2%)	6 (1%)	6 (1%)	8 (1%)	7 (1%)	8 (1%)	7 (1%)
Handwashing location							
Within six steps of latrine							
Has water	178 (14%)	83 (13%)	81 (13%)	63 (10%)	67 (10%)	62 (10%)	72 (11%)
Has soap	88 (7%)	50 (8%)	48 (8%)	34 (5%)	42 (7%)	32 (5%)	36 (6%)
Within six steps of kitchen							
Has water	118 (9%)	51 (8%)	51 (8%)	45 (7%)	61 (9%)	61 (9%)	60 (9%)
Has soap	33 (3%)	18 (3%)	14 (2%)	13 (2%)	15 (2%)	23 (4%)	18 (3%)
Nutrition							
Household is food secure†	932 (67%)	495 (71%)	475 (68%)	475 (69%)	482 (69%)	479 (69%)	485 (71%)

Data are n (%) or mean (SD). Percentages were estimated from slightly smaller denominators than those shown at the top of the table for the following variables due to missing values: mother's age; father's education; father works in agriculture; acres of land owned; open defecation; latrine has a concrete slab; latrine has a functional water seal; visible stool on latrine slab or floor; ownership of child potty; observed faeces in the house or child's play area; and handwashing variables. *Households in these communities who do not own a latrine typically share a latrine with extended family members who live in the same compound. †Assessed by the Household Food Insecurity Access Scale.

Table 2: **Baseline characteristics by intervention group**

two handwashing stations, one with a 40 L water reservoir placed near the latrine and a 16 L reservoir for the kitchen. Each handwashing station included a basin to collect rinse water and a soapy water bottle.[16] Promoters also provided a regular supply of detergent sachets for making soapy water. Promoters encouraged residents to wash

	Control	Water	Sanitation	Handwashing	Washing, sanitation, and handwashing	Nutrition	Washing, sanitation, handwashing, and nutrition
Number of compounds assessed							
Enrolment	1382 (100%)	698 (100%)	696 (100%)	688 (100%)	702 (100%)	699 (100%)	686 (100%)
Year 1	1151 (83%)	611 (88%)	583 (84%)	585 (85%)	605 (86%)	581 (83%)	600 (87%)
Year 2	1138 (82%)	598 (86%)	585 (84%)	570 (83%)	588 (84%)	574 (82%)	586 (85%)
Stored drinking water							
Enrolment	666 (48%)	353 (51%)	341 (49%)	347 (50%)	304 (43%)	301 (43%)	331 (48%)
Year 1	503 (44%)	587 (96%)	245 (42%)	266 (45%)	588 (97%)	229 (39%)	577 (96%)
Year 2	485 (43%)	567 (95%)	260 (44%)	267 (47%)	558 (95%)	225 (39%)	569 (97%)
Stored drinking water has detectable free chlorine (>0·1 mg/L)							
Enrolment	..	..	..	..	..	..	..
Year 1	..	467 (78%)	..	..	467 (79%)	..	472 (80%)
Year 2	..	488 (84%)	..	..	471 (81%)	..	501 (87%)
Latrine with a functional water seal							
Enrolment	358 (31%)	183 (31%)	177 (30%)	162 (28%)	152 (26%)	183 (31%)	155 (27%)
Year 1	308 (29%)	151 (27%)	554 (95%)	144 (27%)	573 (95%)	149 (28%)	564 (94%)
Year 2	324 (31%)	184 (33%)	568 (97%)	165 (32%)	567 (97%)	163 (31%)	561 (96%)
No visible faeces on latrine slab or floor							
Enrolment	625 (48%)	350 (53%)	332 (52%)	335 (52%)	289 (44%)	331 (51%)	298 (46%)
Year 1	658 (60%)	358 (61%)	516 (89%)	324 (58%)	522 (86%)	333 (60%)	527 (88%)
Year 2	612 (56%)	338 (58%)	502 (86%)	324 (60%)	484 (82%)	313 (58%)	495 (85%)
Handwashing location has soap							
Enrolment	294 (23%)	153 (24%)	155 (25%)	134 (22%)	155 (24%)	152 (24%)	149 (23%)
Year 1	283 (28%)	165 (30%)	158 (30%)	533 (91%)	546 (90%)	172 (34%)	536 (89%)
Year 2	320 (28%)	177 (30%)	180 (31%)	527 (92%)	531 (90%)	195 (34%)	540 (92%)
LNS sachets consumed (% expected)*							
Enrolment	..	..	..	..	..	..	..
Year 1	..	..	..	..	..	93%	94%
Year 2	..	..	..	..	..	94%	93%

Data are n (%) or %. Free chlorine in drinking water and LNS consumption were not measured at enrolment and were only measured in a subset of groups. LNS=lipid-based nutrient supplement. *LNS adherence measured as proportion of 14 sachets consumed in the past week among index children ages 6–24 months (reported).

Table 3: **Measures of intervention adherence by study group at enrolment and at 1-year and 2-years follow-up**

their hands with soapy water before preparing food, before eating or feeding a child, after defecating, and after cleaning a child who has defected.

We aimed to deploy interventions so that index children were born into households with the interventions in place. In the combined intervention arms, the sanitation intervention was implemented first, followed by handwashing and then water treatment.

The nutrition intervention targeted index children. Promoters gave study mothers with children aged 6–24 months two 10 g sachets per day of lipid-based nutrient supplement (LNS; Nutriset; Malaunay, France) that could be mixed into the child's food. Each sachet provided 118 kcal, 9·6 g fat, 2·6 g protein, 12 vitamins, and ten minerals. Promoters explained that LNS should not replace breastfeeding or complementary foods and encouraged caregivers to exclusively breastfeed their children during the first 6 months and to provide a diverse, nutrient-dense diet using locally available foods for children older than 6 months. Intervention messages were adapted from the Alive & Thrive programme in Bangladesh.[17]

Outcomes

Primary outcomes were caregiver-reported diarrhoea among all children who were in utero or younger than 3 years at enrolment in the past 7 days (based on all data from year 1 and year 2) and length-for-age Z score at year 2 in index children. Secondary outcomes included length-for-age Z score at year 1; weight-for-length Z score, weight-for-age Z score, head circumference-for-age Z score at year 1 and year 2; and prevalence of moderate stunting (length-for-age Z score less than −2), severe stunting (length-for-age Z score less than −3) underweight (weight-for-age Z score less than −2), and wasting (weight-for-age Z score less than −2). All-cause mortality among index children was a tertiary outcome.[10] Full details on exclusion criteria, measurement protocols, and outcome definitions are in the appendix (p 21–27).

	N	Mean* prevalence	Unadjusted† prevalence difference (95% CI)	Adjusted‡ prevalence difference (95% CI)
Control *vs* intervention				
Control	3517	5·7%	..	..
Water	1824	4·9%	−0·6 (−1·9 to 0·6)	−0·8 (−2·2 to 0·6)
Sanitation	1760	3·5%	−2·2 (−3·4 to −1·0)	−2·3 (−3·5 to −1·1)
Handwashing	1795	3·5%	−2·3 (−3·4 to −1·1)	−2·5 (−3·6 to −1·3)
Water, sanitation, and handwashing	1902	3·9%	−1·7 (−2·9 to −0·6)	−1·8 (−3·1 to −0·4)
Nutrition	1766	3·5%	−2·0 (−3·1 to −0·8)	−2·1 (−3·5 to −0·8)
Water, sanitation, handwashing, and nutrition	1861	3·5%	−2·2 (−3·3 to −1·0)	−2·2 (−3·4 to −1·0)
Water, sanitation, and handwashing *vs* individual groups				
Water, sanitation, and handwashing	1902	3·9%	..	..
Water	1824	4·9%	−1·2 (−2·5 to 0·2)	−0·9 (−2·2 to 0·5)
Sanitation	1760	3·5%	0·4 (−0·8 to 1·7)	0·5 (−0·8 to 1·8)
Handwashing	1795	3·5%	0·3 (−1·0 to 1·5)	0·7 (−0·6 to 1·9)

Among children younger than 3 years at enrolment. *Post-intervention measurements in years 1 and 2 combined. †Unadjusted estimates were estimated using a pair-matched Mantel-Haenszel analysis. ‡Adjusted for prespecified covariates using targeted maximum likelihood estimation with data-adaptive model selection: field staff who collected data, month of measurement, household food insecurity, child age, child sex, mother's age, mothers height, mothers education level, number of children younger than 18 years in the household, number of individuals living in the compound, distance in minutes to the primary water source, household roof, floor, wall materials, and household assets.

Table 4: **Diarrhoea prevalence 1 and 2 years (combined) after intervention**

For more on the **preregistered analysis protocol and full replication files** see https://osf.io/wvyn4

Outcome and adherence was assessed by a team of university graduates who were not involved in the delivery or promotion of interventions. They received a minimum of 21 days of formal training. The mother of the index child answered the interview questions.

We defined diarrhoea as at least three loose or watery stools within 24 h or at least one stool with blood.[18] We assessed diarrhoea in the preceding 7 days among index children and among children who lived in enrolled compounds and who were younger than 3 years at enrolment and so would be expected to remain under 5 years of age throughout the trial. Diarrhoea was assessed at about 16 months and 28 months after enrolment. We included caregiver-reported bruising or abrasion as a negative control outcome.[19]

We calculated Z scores for length for age, weight for length, weight for age, and head circumference for age using the WHO 2006 child growth standards. Child mortality was assessed at the two follow-up evaluation visits based on caregiver interview. Length-for-age Z scores were measured at about 28 months after enrolment when index children would average about 24 months of age. Trained anthropometrists followed standard protocols[20] and measured recumbent length (to 0·1 cm) and weight without clothing in duplicate; if the two values disagreed (>0·5 cm for length, 0·1 kg for weight) they repeated the measure until replicates fell within the error tolerance. We excluded children from Z-score analyses if their measurements were outside biologically plausible ranges according to WHO recommendations.[20]

Statistical analyses

Sample size calculations for the two primary outcomes were based on a relative risk of diarrhoea of 0·7 or smaller (assuming a 7-day prevalence of 10% in the control group[21]) and a minimum detectable effect of 0·15 length-for-age Z score for comparisons of any intervention against control, accounting for repeated measures within clusters. The calculations assumed a type I error (α) of 0·05 and power (1–β) of 0·8, a one-sided test for a two-sample comparison of means, and 10% loss to follow-up. Sample size calculations indicated 90 clusters per group, each with eight children. Full details are given in appendix 4 of our study protocol.[10]

We analysed participants according to their randomised assignment (intention to treat), regardless of adherence to the intervention. Since randomisation was geographically pair-matched in blocks of eight clusters, we estimated unadjusted prevalence differences and ratios using a pooled Mantel-Haenszel estimator that stratified by matched pair.

We used paired t tests and cluster-level means for unadjusted Z score comparisons. For each comparison, we calculated two p values (two-sided): one for the test that mean differences were different from zero and a second to test for any difference between groups in the full distribution using permutation tests with the Wilcoxon signed-rank statistic. Secondary adjusted analyses controlled for prespecified, prognostic baseline covariates using data-adaptive, targeted maximum likelihood estimation. To assess whether interventions affected nearby clusters, we estimated the difference in primary outcomes between control compounds at different distances from intervention compounds. We did not adjust for multiple comparisons.[22]

Analyses were done in *R* (version 3.2.3). We tested for the presence of between-cluster spillover effects using a non-parametric method described in the prespecified analysis plan.

The trial is registered at ClinicalTrials.gov, number NCT01590095. The International Centre for Diarrhoeal Disease Research, Bangladesh convened a data and safety monitoring board and oversaw the study.

Role of the funding source

The funders of the study approved the study design, but had no role in data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all data in the study and had final responsibility for the decision to submit for publication.

Results

Fieldworkers identified 13 279 compounds with a pregnant woman in her first or second trimester; over half were excluded to create 1 km buffer zones between intervention areas. Between May 31, 2012, and July 7, 2013, we randomly allocated 720 clusters and enrolled 5551 pregnant women in 5551 compounds to an

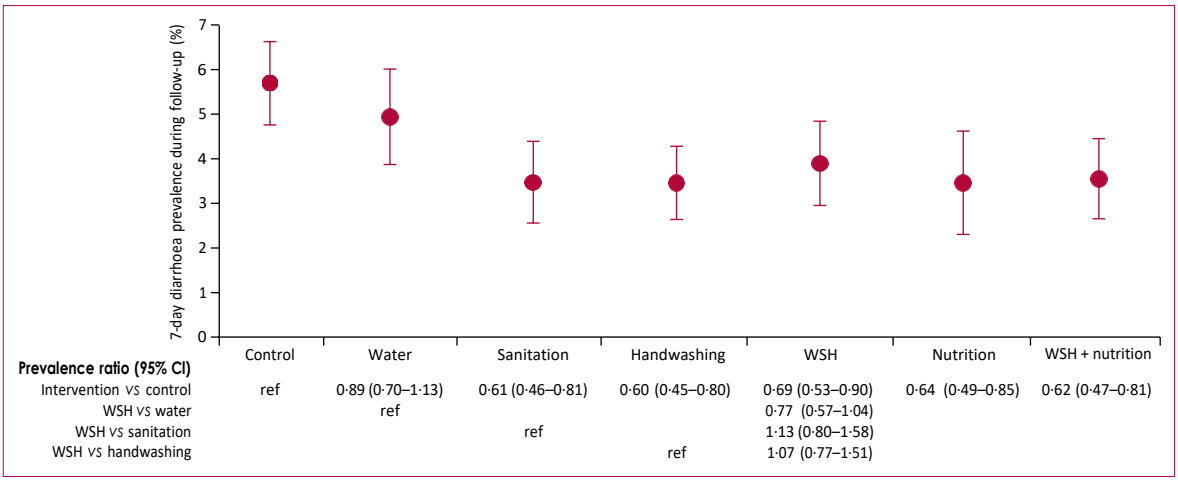


Figure 2: **Intervention effects on diarrhoea prevalence in index children and children younger than 3 years at enrolment 1 and 2 years after intervention**
Data are mean (95% CI). ref=reference. WSH=water, sanitation, and handwashing.

intervention or the control group (figure 1). Index children in 912 (16%) enrolled compounds did not complete follow-up, most commonly because they were not born alive (361 [7%]) or died before the final assessment (220 [4%]). 109 (2%) households moved, 175 (3%) were absent on repeated follow-up, and 47 (<1%) withdrew (figure 1). 4667 (93%) of 4999 surviving index children were measured at year 2, with length-for-age *Z* scores for 4584 (92%) children.

There were a median of two households (IQR 1–3, range 1–11) per compound. Most index households (4108 [74%] of 5551) collected drinking water from shallow tubewells. At enrolment, about half (2976 [54%] of 5551) of households owned their own latrine; most (4979 [90%] of 5551 households) used a latrine that had a concrete slab, and a quarter (1370 [25%] of 5551) had a functional water seal. Baseline characteristics of enrolled households were similar across groups (table 2).

Measures of intervention adherence included presence of stored drinking water with detectable free chlorine (>0·1 mg/L), a latrine with a functional water seal, presence of soap at the primary handwashing location, and reported consumption of LNS sachets. Intervention-specific adherence measures were all greater than 75% in households assigned to the relevant intervention and were substantially higher than practices in the control group. Adherence was similar in the single water, sanitation, handwashing, and nutrition intervention groups compared with the two groups that combined interventions (table 3). Adherence was similar at 1-year and 2-year follow-up.

Diarrhoea prevalence in the control group was substantially below the 10% we had anticipated in our sample size calculations (table 4). Diarrhoea prevalence was particularly low during the first 9 months of observations, with evidence of seasonal epidemics in the control group during the monsoon seasons (appendix p 3).

Compared with the control group, index children and children who were younger than 3 years at enrolment and living in compounds where an index child received any intervention except water treatment had significantly decreased prevalence of diarrhoea at 1-year and 2-year follow-up (figure 2, table 4). The reductions in diarrhoea prevalence in the combined water, sanitation, and handwashing group were no larger than in the individual water, sanitation, or handwashing groups.

Secondary adjusted analyses showed similar effect estimates of interventions on reported diarrhoea (table 4). The effect of intervention was similar among the index children in targeted households (appendix p 10–11) compared with the analysis that included both index children and children younger than 3 years at enrolment who lived in the compound (figure 2); however, the point estimates of the prevalence ratio suggested that water or handwashing interventions did not have a notable effect on non-index children (appendix p 10–11).

There was no difference in prevalence of caregiver-reported bruising or abrasion between children in the control group and any of the intervention groups (appendix p 4).

After 2 years of intervention (median age 22 months, IQR 21–24), mean length-for-age *Z* score in the control group was –1·79 (SD 1·01); children who received the nutrition intervention had an average increase of 0·25 (95% CI 0·15–0·36) in length-for-age *Z* scores; and children who received the water, sanitation, handwashing, and nutrition intervention had an average increase of 0·13 (0·02–0·24) in length-for-age *Z* scores (figure 3). After about 1 year of intervention (median age 9 months, IQR 8–10), children in the nutrition only group (but not children in the water, sanitation, handwashing, and nutrition group) were significantly taller than control children (appendix p 5).

Compared with control children, there was no significant difference in length-for-age *Z* scores in children

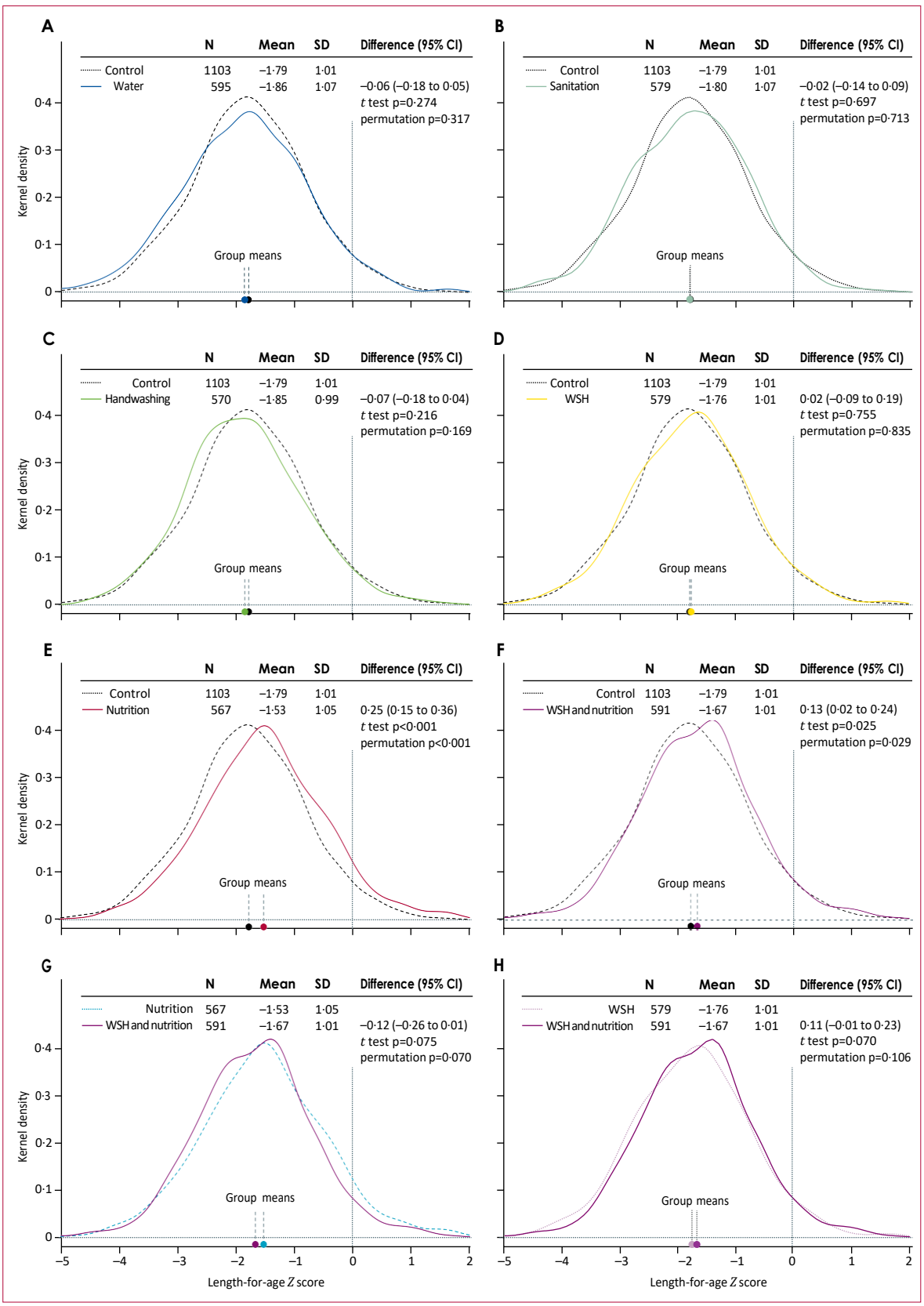


Figure 3: **Intervention effects on length-for-age *Z* scores in 4584 children after 2 years of intervention**

Kernel density plots show the distribution of length-for-age *Z* scores among index children who were born into the study and were aged 18–28 months (median 22, IQR 21–24) at the time of measurement. Dashed lines are the comparison group distribution and solid lines are the active comparator distribution. (A) Water *vs* control. (B) Sanitation *vs* control. (C) Handwashing *vs* control. (D) WSH *vs* control. (E) Nutrition *vs* control. (F) WSH and nutrition *vs* control. (G) WSH and nutrition *vs* nutrition. (H) WSH and nutrition *vs* WSH. WSH=water, sanitation, and handwashing.

receiving the water treatment (length-for-age *Z* score difference –0·06 [95% CI –0·18 to 0·05]), sanitation (–0·02 [–0·14 to 0·09]), handwashing (–0·07 [–0·18 to 0·04]), or water, sanitation, and handwashing interventions (0·02 [–0·09 to 0·13]; figure 3). Length-for-age *Z* scores were similar for children who received water, sanitation, handwashing, and nutrition and those who received nutrition only intervention (–0·12 [–0·26 to 0·01]).

After 2 years of intervention, children in the nutrition only or the water, sanitation, handwashing, and nutrition intervention had higher *Z* scores for length for age, weight for length, weight for age, and head circumference for age than did children in the control group (table 5). Children in the water treatment, sanitation, handwashing, or combined water, sanitation, and handwashing interventions had *Z* scores for length for age, weight for length, weight for age, and head circumference for age that were similar to controls (table 5).

Compared with children living in control households, children enrolled in the nutrition only intervention were less likely to be stunted after 2 years; children enrolled in the water, sanitation, handwashing, and nutrition intervention were less likely to be severely stunted, or underweight (table 6). The proportion of children who were wasted was similar between the intervention and control groups.

Prespecified adjusted analyses found similar effect estimates on anthropometric outcomes with similar efficiency (appendix p 12–15). There was no evidence of between-cluster spillover effects (appendix p 8, 9 and 17–20).

In the control group, the cumulative incidence of child mortality was 4·7% (figure 1). Mortality in the individual water, sanitation, and handwashing groups and combined water, sanitation, and handwashing group was similar to controls. The two groups with a nutrition intervention had lower mortality: 3·8% for the nutrition group and 2·9% for the water, sanitation, handwashing, and nutrition group; this difference was significant for the combined group (risk difference water, sanitation, handwashing, and nutrition *vs* control –1·9% [95% CI –3·6 to –0·1]; p=0·0371; 38% relative reduction; appendix p 16).

	N	Mean (SD)	Difference *vs* control	Difference *vs* (95% CI)	Difference *vs* sanitation, and handwashing (95% CI)
Weight-for-age *Z* score					
Control	1121	–1·54 (1·00)	..	..	..
Water	599	–1·61 (1·04)	–0·07 (–0·19 to 0·04)	..	..
Sanitation	588	–1·52 (1·06)	–0·00 (–0·11 to 0·11)	..	..
Handwashing	573	–1·57 (1·00)	–0·04 (–0·16 to 0·08)	..	..
Water, sanitation, and handwashing	586	–1·53 (1·05)	0·00 (–0·09 to 0·10)	..	..
Nutrition	573	–1·29 (1·07)	0·24 (0·12 to 0·35)	..	..
Water, sanitation, handwashing, and nutrition	592	–1·42 (0·99)	0·13 (0·04 to 0·22)	–0·11 (–0·23 to 0·02)	0·12 (0·01 to 0·23)
Weight-for-height *Z* score					
Control	1104	–0·88 (0·93)	..	..	..
Water	596	–0·92 (0·97)	–0·04 (–0·14 to 0·05)	..	..
Sanitation	580	–0·85 (0·95)	0·01 (–0·09 to 0·11)	..	..
Handwashing	570	–0·86 (0·94)	0·00 (–0·11 to 0·12)	..	..
Water, sanitation, and handwashing	580	–0·88 (1·01)	0·00 (–0·10 to 0·11)	..	..
Nutrition	567	–0·71 (1·00)	0·15 (0·04 to 0·26)	..	..
Water, sanitation, handwashing, and nutrition	591	–0·79 (0·94)	0·09 (0·00 to 0·18)	–0·06 (–0·17 to 0·05)	0·09 (–0·03 to 0·21)
Head circumference-for-age *Z* score					
Control	1118	–1·61 (0·94)	..	..	..
Water	594	–1·63 (0·91)	–0·04 (–0·14 to 0·06)	..	..
Sanitation	584	–1·61 (0·86)	–0·01 (–0·10 to 0·09)	..	..
Handwashing	571	–1·56 (0·93)	0·05 (–0·06 to 0·15)	..	..
Water, sanitation, and handwashing	584	–1·59 (0·91)	0·03 (–0·07 to 0·12)	..	..
Nutrition	570	–1·45 (0·94)	0·16 (0·04 to 0·27)	..	..
Water, sanitation, handwashing, and nutrition	590	–1·51 (0·90)	0·11 (0·01 to 0·20)	–0·05 (–0·17 to 0·07)	0·08 (–0·04 to 0·19)

All three secondary outcomes were prespecified.

Table 5: **Child growth *Z* scores at 2-year follow-up**

Discussion

In the WASH Benefits Bangladesh cluster-randomised controlled trial, the linear growth of children whose households had a chlorinated drinking water intervention, sanitation improvements, or handwashing intervention alone or in combination was no different than children in randomly assigned control households that received no intervention. Children in the nutrient supplement and counselling group grew somewhat taller than controls. Children in households that received a combination of water, sanitation, handwashing, and nutrition had no greater growth benefit than those receiving the nutrition-only intervention. Compared with control households, caregiver-reported diarrhoea prevalence was significantly decreased in households that received any of the interventions, except those who received only the drinking water treatment.

The trial's statistical power to detect small effects and high adherence to the interventions suggest that the absence of improvement in growth with water, sanitation, and handwashing interventions was a genuine null effect. These results suggest either that the hypothesis that exposure to faecal contamination contributes importantly to child growth faltering in Bangladesh is flawed or that the hypothesis remains valid but the water, sanitation, and handwashing interventions used in this trial did not reduce exposure to environmental pathogens sufficiently to reduce growth faltering. Future articles from our group will describe the effects of intervention on environmental contamination with faecal indicator bacteria and on the prevalence and concentration of

	n/N (%)	Difference *vs* control (95% CI)	Difference *vs* washing, sanitation, and handwashing (95% CI)	Difference *vs* nutrition (95% CI)
Stunting*				
Control	451/1103 (41%)	..	..	..
Water	255/595 (43%)	2·4 (–2·6 to 7·3)	..	..
Sanitation	232/579 (40%)	–0·4 (–5·3 to 4·6)	..	..
Handwashing	263/570 (46%)	5·3 (0·2 to 10·3)	..	..
Water, sanitation, and handwashing	232/579 (40%)	–0·5 (–5·5 to 4·4)	..	..
Nutrition	186/567 (33%)	–7·7 (–12·4 to –2·9)	..	..
Water, sanitation, handwashing, and nutrition	221/591 (37%)	–3·8 (–8·6 to 1·1)	–2·8 (–8·4 to 2·8)	4·0 (–1·6 to 9·6)
Severe stunting†				
Control	124/1103 (11%)	..	..	..
Water	86/595 (15%)	3·3 (–0·1 to 6·7)	..	..
Sanitation	65/579 (11%)	0·1 (–3·0 to 3·3)	..	..
Handwashing	65/570 (11%)	0·2 (–3·0 to 3·4)	..	..
Water, sanitation, and handwashing	59/579 (10%)	–1·0 (–4·1 to 2·1)	..	..
Nutrition	47/567 (8%)	–2·8 (–5·7 to 0·2)	..	..
Water, sanitation, handwashing, and nutrition	50/591 (9%)	–3·0 (–5·9 to 0·0)	–1·9 (–5·2 to 1·4)	–0·3 (–3·5 to 3·0)
Wasting†				
Control	118/1104 (11%)	..	..	..
Water	73/596 (12%)	1·8 (–1·4 to 5·0)	..	..
Sanitation	65/580 (11%)	0·9 (–2·3 to 4·0)	..	..
Handwashing	60/570 (11%)	0·1 (–3·1 to 3·2)	..	..
Water, sanitation, and handwashing	69/580 (12%)	1·4 (–1·8 to 4·6)	..	..
Nutrition	50/567 (9%)	–1·6 (–4·5 to 1·3)	..	..
Water, sanitation, handwashing, and nutrition	52/591 (9%)	–1·7 (–4·7 to 1·2)	–2·8 (–6·3 to 0·7)	0·2 (–3·0 to 3·5)
Underweight†				
Control	344/1121 (31%)	..	..	..
Water	213/599 (36%)	5·3 (0·7 to 10·0)	..	..
Sanitation	179/588 (30%)	0·3 (–4·3 to 4·9)	..	..
Handwashing	197/573 (34%)	3·9 (–0·9 to 8·7)	..	..
Water, sanitation, and handwashing	192/586 (33%)	2·2 (–2·4 to 6·8)	..	..
Nutrition	149/573 (26%)	–4·2 (–8·6 to 0·3)	..	..
Water, sanitation, handwashing, and nutrition	148/592 (25%)	–5·8 (–10·2 to –1·4)	–7·8 (–12·9 to –2·6)	–1·7 (–6·6 to 3·3)

*Prespecified secondary outcome. †Prespecified tertiary outcome.

Table 6: **Prevalence of children stunted, severely stunted, wasted, and underweight at 2-year follow-up**

enteric pathogens in stool specimens from children and thus provide insight on how effectively the interventions altered environmental contamination and enteropathogen transmission.

The effect of the nutrition intervention, which corrected one sixth of the growth deficit compared with international norms of healthy growth, was consistent with other randomised controlled trials of postnatal LNS that have reported variable and generally small effects on linear growth.[23–27] This variation is probably because of contextual factors that affect a population's capacity to respond to an intervention. The water, sanitation, and handwashing intervention did not affect crucial contextual factors to amplify the effect of the nutrition interventions in rural Bangladesh. Continued research should explore interventions to reduce growth faltering.

Although intervention households generally reported less diarrhoea, people who received the intervention might have been grateful and, out of courtesy, reported less diarrhoea.[28] However, compared with control households, intervention households reported no reduction in bruising or abrasions (negative control outcomes), so there was no evidence of systematic under-reporting of all health outcomes. It also seems unlikely that courtesy bias would affect each of the interventions except the drinking water intervention. The nutrition intervention might have led to improvements in breastfeeding practices or in essential fatty acids or micronutrient status, which could have contributed to improved gut epithelial immune response and thus less diarrhoea.[29]

The finding that drinking water treatment intervention had no notable effect on diarrhoea contrasts with our previous study of the identical intervention done between October, 2011, and November, 2012 in nearby communities that found a 36% reduction in reported diarrhoea.[11] Restriction of the analysis to WASH Benefits index children who were targeted for the drinking water intervention led to a stronger treatment effect estimate (prevalence ratio 0·80 [95% CI 0·60–1·07]). Diarrhoea prevalence in the WASH Benefits control group (6%) was substantially lower than the 10% prevalence noted in a large prior study[21] and the 11% prevalence in the control group of our previous study.[11] Diarrhoeal prevalence characteristically varies substantially in nearby locations and from year to year.[30] Diarrhoea prevalence in the control group of this WASH Benefits trial in rural Bangladesh was similar to diarrhoea prevalence among cohorts of children aged 1–4 years in the USA.[31] At the time of the study, rotavirus immunisation had not been introduced into the Bangladesh national immunisation programme. The unexpectedly low diarrhoea prevalence among control children suggests decreased transmission of diarrhoea-causing pathogens during the WASH Benefits trial compared with recent evaluations. This low transmission provided less opportunity to interrupt transmission and less statistical power to show that interruption.

Combining interventions to improve drinking water quality, sanitation, and handwashing provided no additive benefit for the reduction of diarrhoea over single interventions. The unexpectedly low diarrhoea prevalence suggests low transmission of enteric pathogens through some of the pathways, which might have prevented any additive benefit from the combined interventions. Combined interventions did not compromise observed adherence to recommended practices. If a substantial proportion of the reduced diarrhoea was because of

courtesy bias, this bias might mask subtle additive benefits. The only previous randomised controlled evaluations of multiple interventions versus single interventions also found no additive benefit of multiple components of water, sanitation, and handwashing on reported diarrhoea among children younger than 5 years.[7,32,33] Because transmission pathways of enteropathogens vary by time and location, this absence of an additive effect with combined interventions is unlikely to generalise to all locations. However, these findings suggest that focusing resources on a single low-cost high-uptake intervention to a larger population might reduce diarrhoea prevalence more than would similar spending on more comprehensive approaches to smaller populations.

Children who received both the nutrition and the combined water, sanitation, and handwashing intervention were 38% less likely to die than children in the control group. Mortality was not a primary study outcome. Although the confidence limits are broad and the p value is borderline (p=0·037), a causal relationship from the interventions is plausible, since diarrhoea and poor nutrition are risk factors for death among young children in this setting. Notably, reduced mortality was only seen in the intervention groups that saw improved growth (nutrition groups), which were the groups with objective indicators of biological effect. Forthcoming investigations of the timing and causes of death assessed by verbal autopsy, distribution of enteropathogens among intervention groups, and effect of interventions on respiratory disease will provide additional evidence to assess the biological plausibility of a causal relationship between the combined water, sanitation, handwashing, and nutrition intervention and reduced mortality.

The randomised design, balanced groups, and high adherence suggests that the absence of an association between water, sanitation, and handwashing interventions and growth is internally valid, but this intervention was implemented in one socioecological zone (rural Bangladesh) during a time of low diarrhoea prevalence. Reducing faecal exposure through household water, sanitation, and handwashing interventions might affect growth in settings with a different prevalence of gastrointestinal disease or mix of pathogens.[34] Notably, water, sanitation, and handwashing interventions did not prevent growth faltering in this context where stunting is a prevalent public health issue and where adherence to the interventions was substantially higher than in typical programmatic interventions.[21,35,36]

The objective measures of uptake reflected the availability of infrastructure and supplies, but might over-represent actual use. Future articles from our group will include structured observation and other measures of uptake. Although more intensive interventions could lead to even better practices, it seems unlikely that large-scale routine programmes could implement interventions with such intensity.

Because the sanitation intervention targeted compounds with pregnant women, these interventions only reached about 10% of residents in villages where interventions were implemented. If a higher threshold of sanitation coverage is necessary to achieve herd protection, then this study design would preclude the detection of this effect. We used compounds as the unit of intervention because they enabled us to deliver intensive interventions with high adherence for thousands of newborn children. In addition, we expected compound-level faecal contamination to represent the dominant source of exposure for index children because of the physical separation of compounds, and because children younger than 2 years of age in these communities spent nearly all of their time in their own compound.

The combined water, sanitation, handwashing, and nutrition intervention had sustained high levels of adherence. Although the full range of benefits of these successfully integrated interventions are yet to be fully elucidated, our findings suggest there might be a survival benefit. Forthcoming articles by our group will report the effects of intervention on biomarkers of environmental enteric dysfunction, soil-transmitted helminth infection, enteric pathogen infection, biomarkers of inflammation and allostatic load, anaemia and nutritional biomarkers, and child language, motor development, and social skills.

Contributors

SPL drafted the research protocol and manuscript with input from all coauthors and coordinated input from the study team throughout the project. PJW, EL, FB, FH, MR, LU, PKR, FAN, and TFC developed the water, sanitation, and handwashing intervention. CPS, KJ, KGD, and TA developed the nutrition intervention and guided the analysis and interpretation of these results. MR, LU, SA, FB, FH, AMN, SMP, KJ, AL, AE, KKD, and JA oversaw piloting and subsequent study implementation, contributed to refinements in interventions and measurements, and responded to threats to validity. BFA, JB-C, AEH, and JMC developed the analytical approach, did the statistical analysis, constructed the tables and figures, and helped interpret the results. CN and LCF helped to develop the study design and interpret of results.

Declaration of interests

We declare no competing interests.

Acknowledgments

We appreciate the time, patience, and good humour of the study participants and the remarkable dedication to quality of the field team who delivered the intervention and assessed the outcomes. This research was financially supported by a global development grant (OPPGD759) from the Bill & Melinda Gates Foundation to the University of California, Berkeley, CA, USA.

References

1 Lu C, Black MM, Richter LM. Risk of poor development in young children in low-income and middle-income countries: an estimation and analysis at the global, regional, and country level. *Lancet Glob Health* 2016; **4:** e916–22.

2 Black MM, Walker SP, Fernald LC, et al. Early childhood development coming of age: science through the life course. *Lancet* 2016; **389:** 77–90.

3 Dewey KG, Adu-Afarwuah S. Systematic review of the efficacy and effectiveness of complementary feeding interventions in developing countries. *Matern Child Nutr* 2008; **4** (suppl 1)**:** 24–85.

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

5 Cumming O, Cairncross S. Can water, sanitation and hygiene help eliminate stunting? Current evidence and policy implications. *Matern Child Nutr* 2016; **12** (suppl 1)**:** 91–105.

6 Wang H, Naghavi M, Allen C, et al. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* 2016; **388:** 1459–544.

7 Fewtrell L, Kaufmann RB, Kay D, Enanoria W, Haller L, Colford JM Jr. Water, sanitation, and hygiene interventions to reduce diarrhoea in less developed countries: a systematic review and meta-analysis. *Lancet Infect Dis* 2005; **5:** 42–52.

8 Waddington H, Snilstveit B. Effectiveness and sustainability of water, sanitation, and hygiene interventions in combating diarrhoea. *J Dev Effect* 2009; **1:** 295–335.

9 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; published online Jan 29. http://dx.doi.org/10.1016/S2214-109X(18)30005-6.

10 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.

11 Ercumen A, Naser AM, Unicomb L, Arnold BF, Colford J, Luby SP. Effects of source- versus household contamination of tubewell water on child diarrhea in rural Bangladesh: a randomized controlled trial. *PLoS One* 2015; **10:** e0121907.

12 Dreibelbis R, Winch PJ, Leontsini E, et al. The integrated behavioural model for water, sanitation, and hygiene: a systematic review of behavioural models and a framework for designing and evaluating behaviour change interventions in infrastructure-restricted settings. *BMC Public Health* 2013; **13:** 1015.

13 Hussain F, Clasen T, Akter S, et al. Advantages and limitations for users of double pit pour-flush latrines: a qualitative study in rural Bangladesh. *BMC Public Health* 2017; **17:** 515.

14 Sultana R, Mondal UK, Rimi NA, et al. An improved tool for household faeces management in rural Bangladeshi communities. *Trop Med Int Health* 2013; **18:** 854–60.

15 Hussain F, Luby SP, Unicomb L, et al. Assessment of the acceptability and feasibility of child potties for safe child feces disposal in rural Bangladesh. *Am J Trop Med Hyg* 2017; **97:** 469–76.

16 Hulland KR, Leontsini E, Dreibelbis R, et al. Designing a handwashing station for infrastructure-restricted communities in Bangladesh using the integrated behavioural model for water, sanitation and hygiene interventions (IBM-WASH). *BMC Public Health* 2013; **13:** 877.

17 Menon P, Nguyen PH, Saha KK, et al. Combining intensive counseling by frontline workers with a nationwide mass media campaign has large differential impacts on complementary feeding practices but not on child growth: results of a cluster-randomized program evaluation in Bangladesh. *J Nutr* 2016; **146:** 2075–84.

18 Baqui AH, Black RE, Yunus M, Hoque AR, Chowdhury HR, Sack RB. Methodological issues in diarrhoeal diseases epidemiology: definition of diarrhoeal episodes. *Int J Epidemiol* 1991; **20:** 1057–63.

19 Arnold BF, Ercumen A, Benjamin-Chung J, Colford JM Jr. Brief report: negative controls to detect selection bias and measurement bias in epidemiologic studies. *Epidemiology* 2016; **27:** 637–41.

20 de Onis M, Onyango AW, Van den Broeck J, Chumlea WC, Martorell R. Measurement and standardization protocols for anthropometry used in the construction of a new international growth reference. *Food Nutr Bull* 2004; **25** (suppl 1)**:** S27–36.

21 Huda TM, Unicomb L, Johnston RB, Halder AK, Yushuf Sharker MA, Luby SP. Interim evaluation of a large scale sanitation, hygiene and water improvement programme on childhood diarrhea and respiratory disease in rural Bangladesh. *Soc Sci Med* 2012; **75:** 604–11.

22 Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990; **1:** 43–46.

23 Maleta KM, Phuka J, Alho L, et al. Provision of 10-40 g/d lipid-based nutrient supplements from 6 to 18 months of age does not prevent linear growth faltering in Malawi. *J Nutr* 2015; **145:** 1909–15.

24 Hess SY, Abbeddou S, Jimenez EY, et al. Small-quantity lipid-based nutrient supplements, regardless of their zinc content, increase growth and reduce the prevalence of stunting and wasting in young burkinabe children: a cluster-randomized trial. *PLoS One* 2015; **10:** e0122242.

25 Iannotti LL, Dulience SJ, Green J, et al. Linear growth increased in young children in an urban slum of Haiti: a randomized controlled trial of a lipid-based nutrient supplement. *Am J Clin Nutr* 2014; **99:** 198–208.

26 Dewey KG, Mridha MK, Matias SL, et al. Lipid-based nutrient supplementation in the first 1000 d improves child growth in Bangladesh: a cluster-randomized effectiveness trial. *Am J Clin Nutr* 2017; **105:** 944–57.

27 Christian P, Shaikh S, Shamim AA, et al. Effect of fortified complementary food supplementation on child growth in rural Bangladesh: a cluster-randomized trial. *Int J Epidemiol* 2015; **44:** 1862–76.

28 Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 2008; **336:** 601–05.

29 Veldhoen M, Ferreira C. Influence of nutrient-derived metabolites on lymphocyte immunity. *Nat Med* 2015; **21:** 709–18.

30 Luby SP, Agboatwalla M, Hoekstra RM. The variability of childhood diarrhea in Karachi, Pakistan, 2002–2006. *Am J Trop Med Hyg* 2011; **84:** 870–77.

31 Arnold BF, Wade TJ, Benjamin-Chung J, et al. Acute gastroenteritis and recreational water: highest burden among young US children. *Am J Publ Health* 2016; **106:** 1690–97.

32 Luby SP, Agboatwalla M, Painter J, et al. Combining drinking water treatment and hand washing for diarrhoea prevention, a cluster randomised controlled trial. *Trop Med Int Health* 2006; **11:** 479–89.

33 Lindquist ED, George CM, Perin J, et al. A cluster randomized controlled trial to reduce childhood diarrhea using hollow fiber water filter and/or hygiene-sanitation educational interventions. *Am J Trop Med Hyg* 2014; **91:** 190–97.

34 Pickering AJ, Djebbari H, Lopez C, Coulibaly M, Alzua ML. Effect of a community-led sanitation intervention on child diarrhoea and child growth in rural Mali: a cluster-randomised controlled trial. *Lancet Glob Health* 2015; **3:** e701–11.

35 Clasen T, Boisson S, Routray P, et al. Effectiveness of a rural sanitation programme on diarrhoea, soil-transmitted helminth infection, and child malnutrition in Odisha, India: a cluster-randomised trial. *Lancet Glob Health* 2014; **2:** e645–53.

36 Patil SR, Arnold BF, Salvatore AL, et al. The effect of India's total sanitation campaign on defecation behaviors and child health in rural Madhya Pradesh: a cluster randomized controlled trial. *PLoS Med* 2014; **11:** e1001709.

American Journal of Epidemiology
© The Author(s) 2017. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journalpermissions@oup.com.
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.

Downloaded from https://academic.oup.com/aje/article-abstract/186/7/866/3813104
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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,

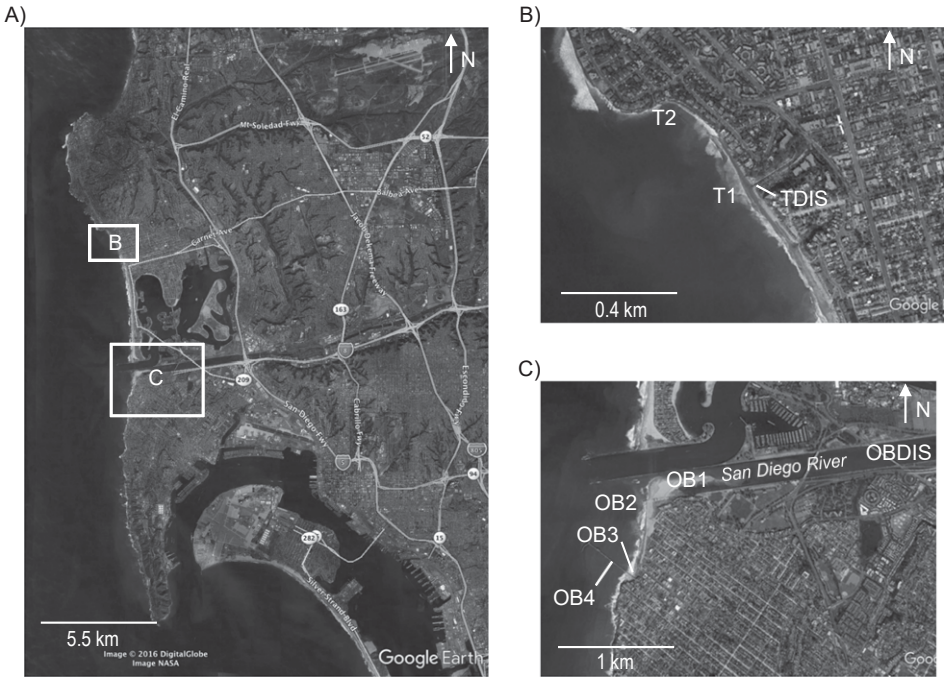


Figure 1. Monitoring beach water quality sampling locations in San Diego, California, winters of 2013–2014 and 2014–2015. Shown are the locations of the 2 monitored beaches along the San Diego coastline (A) and the water quality sampling sites at Tourmaline Surfing Park (B) and Ocean Beach (C). Samples 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.

Downloaded from https://academic.oup.com/aje/article-abstract/186/7/866/3813104
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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

Characteristic	Beach[a]		Online[a]		Total	
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

Downloaded from https://academic.oup.com/aje/article-abstract/186/7/866/3813104
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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 indi- cator 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 indica- tor level. We also estimated the incidence rate ratio associated with exposures to water above versus below US Environ- mental 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,

Downloaded from https://academic.oup.com/aje/article-abstract/186/7/866/3813104
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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).

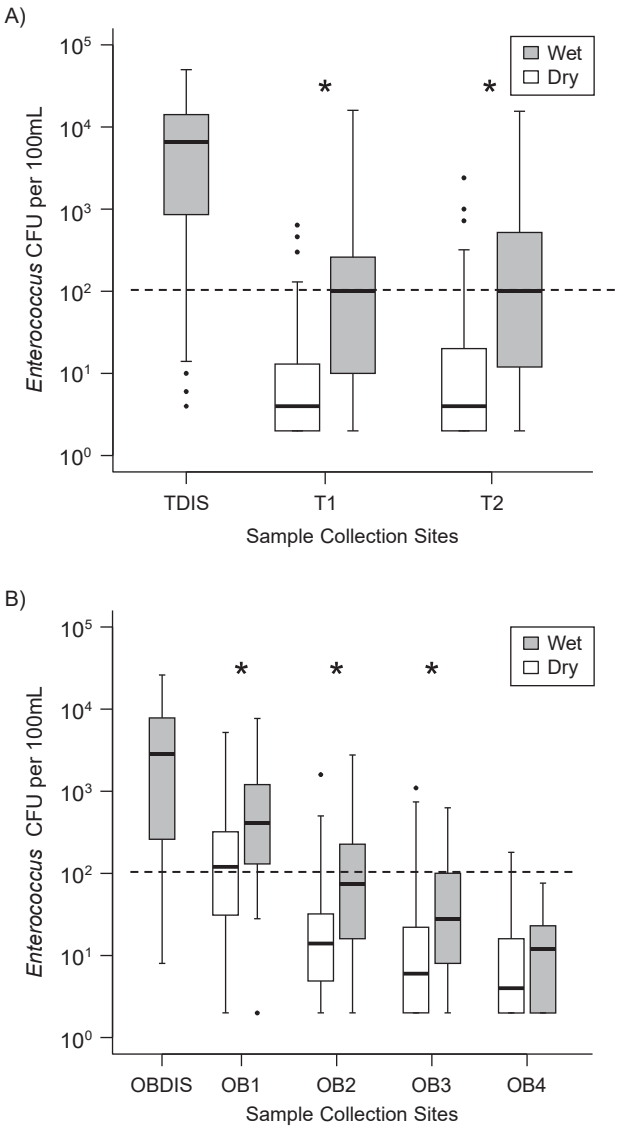


Figure 2. *Enterococcus* levels during dry and wet weather at the sampling locations at Tourmaline Surfing Park (A) and Ocean Beach (B) mapped in Figure 1. Boxes mark interquartile ranges, vertical lines mark 1.5 times the interquartile range, and points mark outliers. Horizontal dashed lines mark the 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.

Downloaded from https://academic.oup.com/aje/article-abstract/186/7/866/3813104
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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

Downloaded from https://academic.oup.com/aje/article-abstract/186/7/866/3813104
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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.

Downloaded from https://academic.oup.com/aje/article-abstract/186/7/866/3813104
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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.

Downloaded from https://academic.oup.com/aje/article-abstract/186/7/866/3813104
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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.

Downloaded from https://academic.oup.com/aje/article-abstract/186/7/866/3813104
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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.

Downloaded from https://academic.oup.com/aje/article-abstract/186/7/866/3813104
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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 Groupt

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; 950Jo confidence interval = 7-120), and risk increased with increasing tampon absorbency (odds ratio = 1.34 per gram increase in absorbency; 950Jo 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 [IO].

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.

t 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); Ors. **M.** Donawa and C. Gaffey (U.S. Food and Drug Administration); and Ors. **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.

Downloaded from https://academic.oup.com/cid/article-abstract/11/Supplement_1/S35/306648
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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 l02°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, 90Jo; day 2, 140Jo; day 3, 170Jo; day 4, 290Jo; day 5, 120Jo; day 6, 130Jo, day 7, 20Jo; and day 8, 40Jo).

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 (lOJo). 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 (98OJo) 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 (660Jo) had used tampons

Downloaded from https://academic.oup.com/cid/article-abstract/11/Supplement_1/S35/306648
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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

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

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

Downloaded from https://academic.oup.com/cid/article-abstract/11/Supplement_1/S35/306648
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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

Tampon use	Odds ratio*/950Jo 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 340/o for every 1-g increase in absorbency (odds ratio per gram increase = 1.34; 950/o 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; 950/o 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	950/o confidence interval
	Patients	Combined controls		
None	2	128		
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		0	
Total	90	347		

* Single brand and style use only.

Downloaded from https://academic.oup.com/cid/article-abstract/11/Supplement_1/S35/306648
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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

Brand and style of tampon	No. (OJo) using brand/style in indicat d group		Odds ratio/950Jo 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)	I (<l)	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	950Jo 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.

Downloaded from https://academic.oup.com/cid/article-abstract/11/Supplement_1/S35/306648
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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)	I.I	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*	I (I)	2 (<I)		
Any spermicide	6 (6)	22 (6)		
Intrauterine device	2 (2)	7 (2)		
Tubal ligation	6 (6)	31 (8)		
Hysterectomy	I (I)	I (<I)		
Rhythm	2 (2)	0		
Withdrawal	2 (2)	I (<I)		
Cervical cap*	0	I (<I)		

• **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 (660Jo) 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 rv65% 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. (OJo) taking medication in indicated group: Patients (*n* = 108)	Controls (*n* = 372)	Odds ratio	950Jo 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)	I.I	0.3-3.6
Premesyn	3 (3)	2 (I)	*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 rvl2,000 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-

Downloaded from https://academic.oup.com/cid/article-abstract/11/Supplement_1/S35/306648
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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

I. 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 111, 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 OW. 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-I by *Staphylococcus aureus.* J Infect Dis 1987;155:812-5

8. Mills JT, Parsonnet J, Kass EH. Production of toxic- shock-syndrome toxin-I: 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-I (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 **l):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-

Downloaded from https://academic.oup.com/cid/article-abstract/11/Supplement_1/S35/306648
by University of California School of Law (Boalt Hall) user
on 13 August 2018

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 revie.w. 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 Toon. 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 l 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 NJ3 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 kno . 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.

Downloaded from https://academic.oup.com/cid/article-abstract/11/Supplement_1/S35/306648
by University of California School of Law (Boalt Hall) user
on 13 August 2018