PHW250B Week 11 Reader

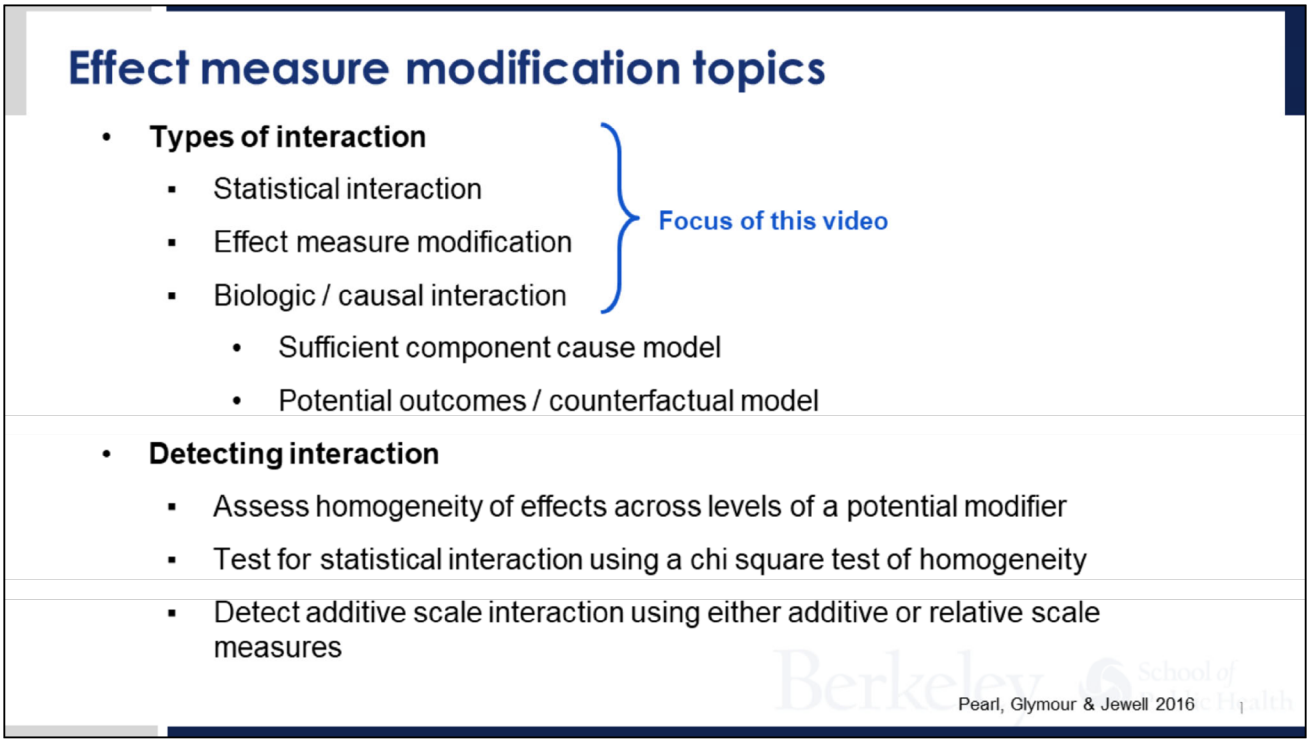
Topic 1: Biological Interaction, Statistical Interaction, Effect Modification

Topic 2: Casual Perspective on Effect Modification

Topic 3: Additive Scale Measures of Interaction

Journal Club

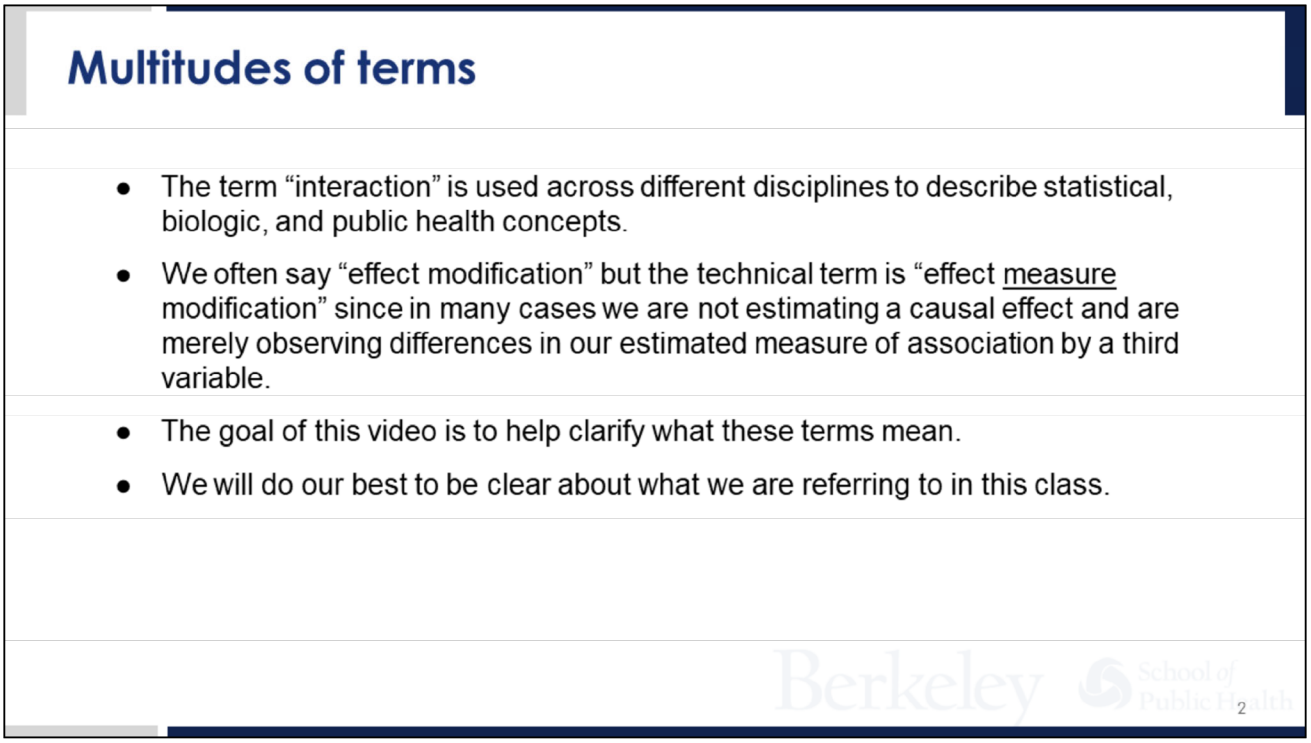
JACK COLFORD: We are going to talk now about a concept that goes by many different names including effect modification, effect measure modification, interaction, synergy, antagonism, and so forth. So we'll go through each of these in this lesson.



If you think about this concept as types of interaction within data, we might split that concept up into three different major broad categories-- statistical interaction, effect measure modification, and biologic or causal interaction. We're going to focus on these in this video, and then other videos we'll talk about how to evaluate the sufficient component cause model and the potential outcomes or counterfactual model for the specific case of biologic or causal interaction.

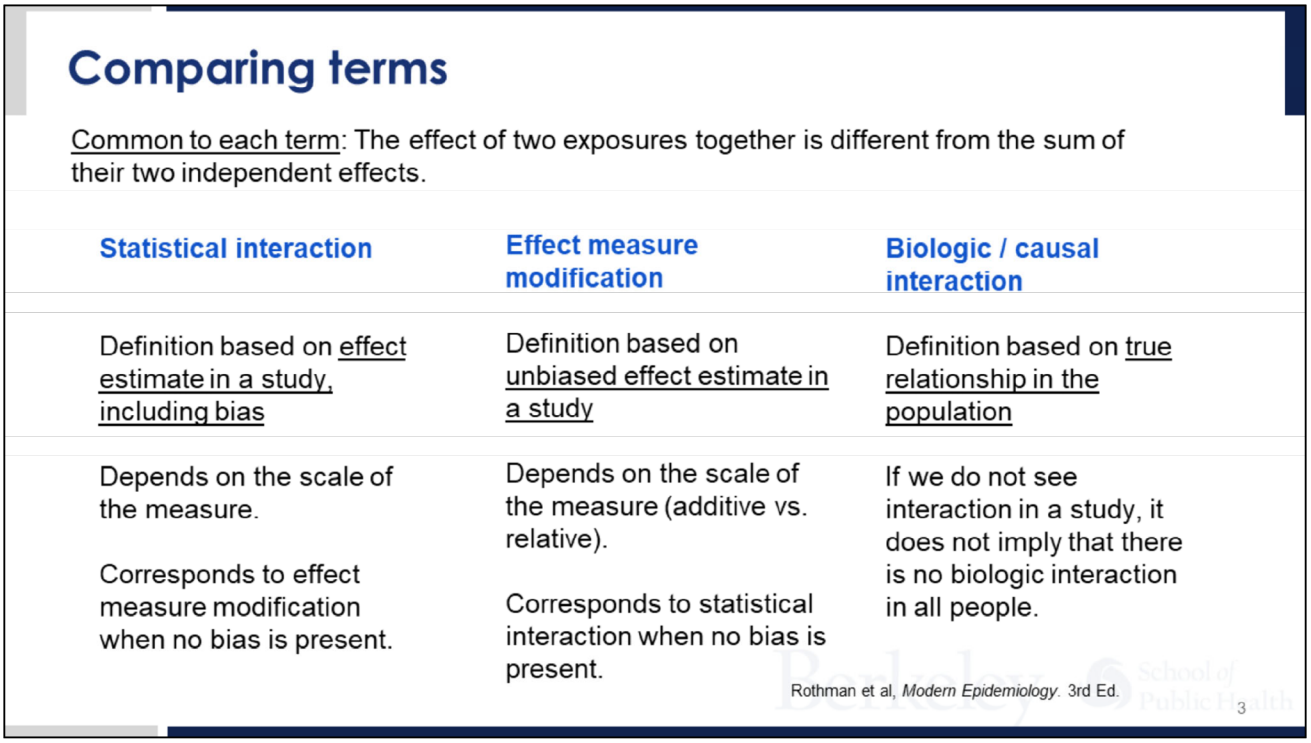
There are different ways to detect the presence of interaction data. One is to assess the homogeneity of effects across levels of a potential modifier. So for example, if I'm looking at the relationship between smoking and lung cancer, I might want to look at the relationship between smoking and lung cancer stratified by gender. So I'd look at among men and among women and see if it has the same relationship or whether gender modifies the relationship between smoking and lung cancer.

We can test for a statistical interaction using a chi square test of homogeneity. And on the additive scale, we can detect interaction using additive or relative scale measures actually.



The term interaction is used across different disciplines to describe statistical, biologic, and public health concepts. In epidemiology, we often say effect modification, but the technical term is effect measure of modification since, in many cases, we are not estimating a causal effect and are merely observing differences in our estimated measure of association by a third variable.

So again, in the example with looking at the relationship between smoking and lung cancer, we're seeing whether that effect is, in itself, changed by attention to a third variable, in this case, gender. So the goal of this video is to help clarify what these terms mean. And we'll do our best to be clear about what we're referring to specifically in this class.



Now, it helps to look at the three different types of interaction we're talking about and compare and contrast them. So basically, what we're saying is common to all of these terms, as it appears that the effect of two exposures together, like smoking and gender, is different from the sum of their two independent effects, as if we looked at smoking and lung cancer, and then separately looked at gender and lung cancer.

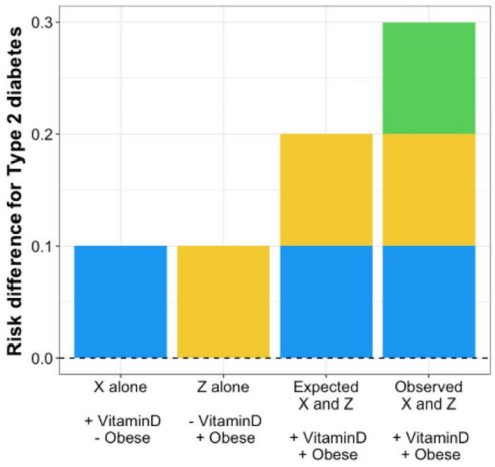
So for statistical interaction, the definition is based on an effect estimate in a study that includes any bias that's present. So we don't know whether the bias is there or not. But if it is there, it's included in our estimate, our statistical estimate. And statistical estimates of interaction depend on the scale of the measure that's being used. And we'll talk more about that. And for statistical interaction, this corresponds to the effect measure modification when no bias is present. So if you calculated effect measure modification with no bias present, that would be statistical interaction.

Effect measure modification has a definition that's based on an unbiased effect estimate in a study. And it depends on the scale of the measure. So are we on the additive scale or the relative scale? And effect measure modification corresponds to statistical interaction when no bias is present.

And finally, biologic or causal interaction has a definition that's based on a true relationship in the population. If we do not see interaction in a study, it does not imply that there is no biologic interaction in all the people in that population.

Statistical interaction & Effect measure modification

- **Statistical interaction:** Departure from additivity of estimated effects on the chosen outcome scale.
- **Effect measure modification:** Departure from additivity of the effects in the study population on the chosen outcome scale
- The green bar in the graph to the right indicates the "departure from additivity" since it implies risk above and beyond the total risk for each exposure on its own.
- Methods for assessing statistical interaction can also be used to assess effect measure modification if we assume that there is no bias in our effect estimates.



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

4

Let's look at a figure to help illustrate some of these concepts. Let me orient you to this figure a little bit. We're looking at the risk for diabetes, or type 2 diabetes, in a population. And we're looking at the risk for diabetes based on the presence of vitamin D, it's presence or absence, and the presence or absence of obesity.

So our two risk factors that we're considering are vitamin D and obesity. And X alone and Z alone refer to either vitamin D alone or obesity alone. And the comparison being made here-- and this is important to understand-- in the first column, we're comparing-- we say X alone-- we're comparing the presence of vitamin D alone to people with neither vitamin D nor obesity. And in the column that says Z alone, we're comparing people with obesity to those with neither vitamin D nor obesity. So in other words, the comparison throughout this figure will always be people with neither risk factor.

In the first column, we're looking at people with one risk factor, vitamin D, but not the other. In the second column, we're looking at people with one risk factor, obesity, but not vitamin D. In the third column, we're looking at people with both vitamin D and obesity in terms of what we expect to see based on their X alone and Z alone estimates. In the final column, we're looking at what we actually observe.

So we've got one column that shows us the risk for one variable. The other column shows us the risk for the other variable. The third column shows us what we would

expect with both of the variables present. And the fourth column shows us what we actually see in our data. So statistical interaction is defined as a departure from additivity of estimated effects on the chosen outcome scale. We'll go through what this means. And effect measure modification is departure from additivity of the effects in the study population on the chosen outcome scale.

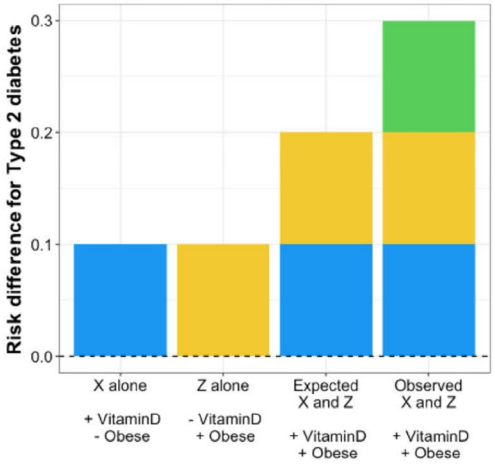
So in the figure to the right, the green bar in the graph on the fourth column, indicates the departure from additivity, since it applies risk above and beyond the total risk for each exposure on its own. Let's talk this through a little bit. The third column is showing us, if we added the risks together, what we would expect to see-- that expected Y and Z column.

But in our observed data, we're seeing risk higher than that. And it's higher by the amount of that green portion at the very top of the fourth column. So that's the departure from additivity. That's the effect measure modification.

Methods for assessing statistical interaction can also be used to assess effect measure modification if we assume that there's no bias in our effect estimates.

What does it mean if we find statistical interaction?

- The presence of additive or relative scale statistical interaction describe patterns on that scale in the population subgroups with the specific combination of exposure.
- Statistical interaction does not necessarily imply biologic or causal interaction.
- In the example to the right, the results imply that on the additive scale, there is a greater risk of type 2 diabetes among population subgroups who are obese and take vitamin D supplements than among other population subgroups.



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

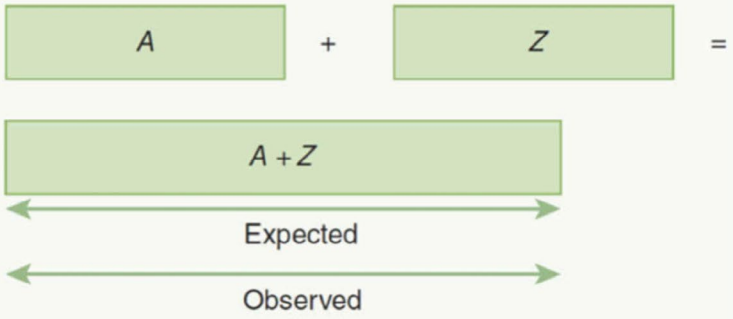
5

The presence of additive or relative scale statistical interaction describes patterns on that scale in the population subgroups with the specific combination of exposure. So this is what this graph, again, is all about. We're asking when we have both of them present compared to neither of them present, do we see more or less risk than we expect to see when both of them are present compared to neither of them present.

So in the example to the right, here in this figure, the results imply that on the additive scale, there is a greater risk of type 2 diabetes among population subgroups who are obese and take vitamin D supplements than among other population subgroups.

Assessing interaction by comparing observed vs. expected joint effects

A. When there is *no* interaction, the *observed* joint effect of risk factors *A* and *Z* equals the sum of their independent effects:



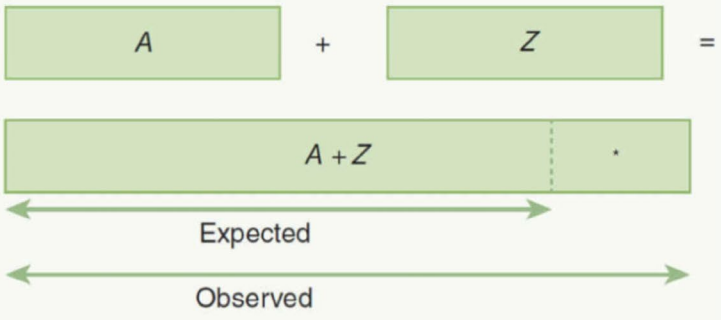
Szklo & Nieto 4th Ed.

6

We can assess interaction by comparing observed versus expected joint effects just as we were doing in the figure a moment ago. Let's look at this figure here where we're constructing an example with no interaction. So we have the risk of some disease from a risk factor A and the risk of some disease from a risk factor Z. If we put those two risk factors together and add up their-- conceptually, add up their risk-- so A plus Z-- that's our expected risk. If we observed that risk, and the observed risk were the same as the expected risk, then there is no interaction.

Synergism

B. When there is *positive* interaction (*synergism*), the *observed* joint effect of risk factors *A* and *Z* is *greater* than the expected on the basis of summing their independent effects:



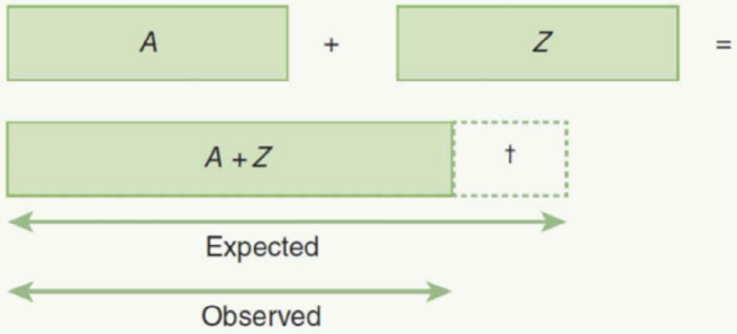
Szklo & Nieto 4th Ed.

Now another concept related here is the term synergism or synergy. And when there's positive interaction or synergism, the observed joint effect of risk factors A and Z is greater than the expected risk on the basis of summing their independent effects or compared to summing their independent effects.

So we have risk factor A, and it confers a certain level of risk. And we have risk factor Z, and it confers a certain level of risk. If we put those two risks together, A plus Z, we expect to see the green bar illustrating risk up to the dotted line. But if we actually observe more risk, then this is called synergism or positive interaction because we're seeing more risk in our observed data than we expected to see by just the individual risks alone.

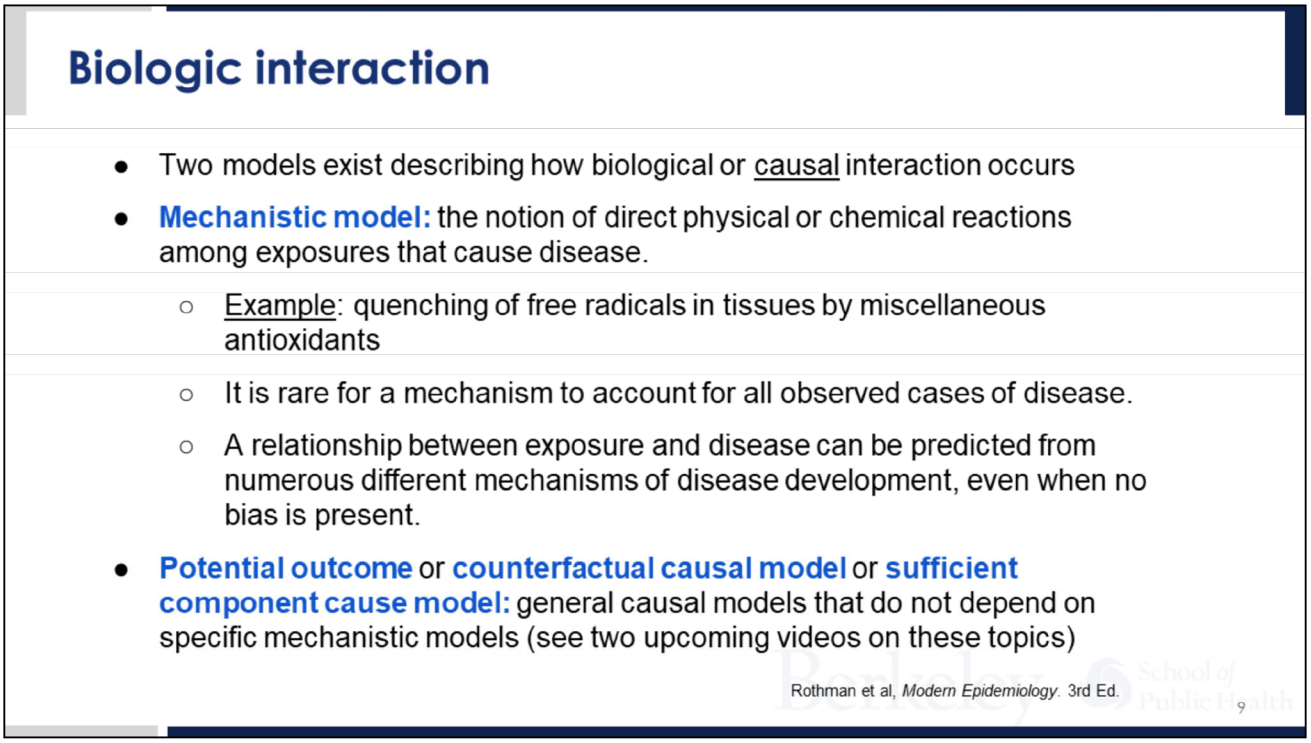
Antagonism

C. When there is *negative* interaction (*antagonism*), the *observed* joint effect of risk factors *A* and *Z* is *smaller* than the *expected* on the basis of summing their independent effects:



Szklo & Nieto 4th Ed.

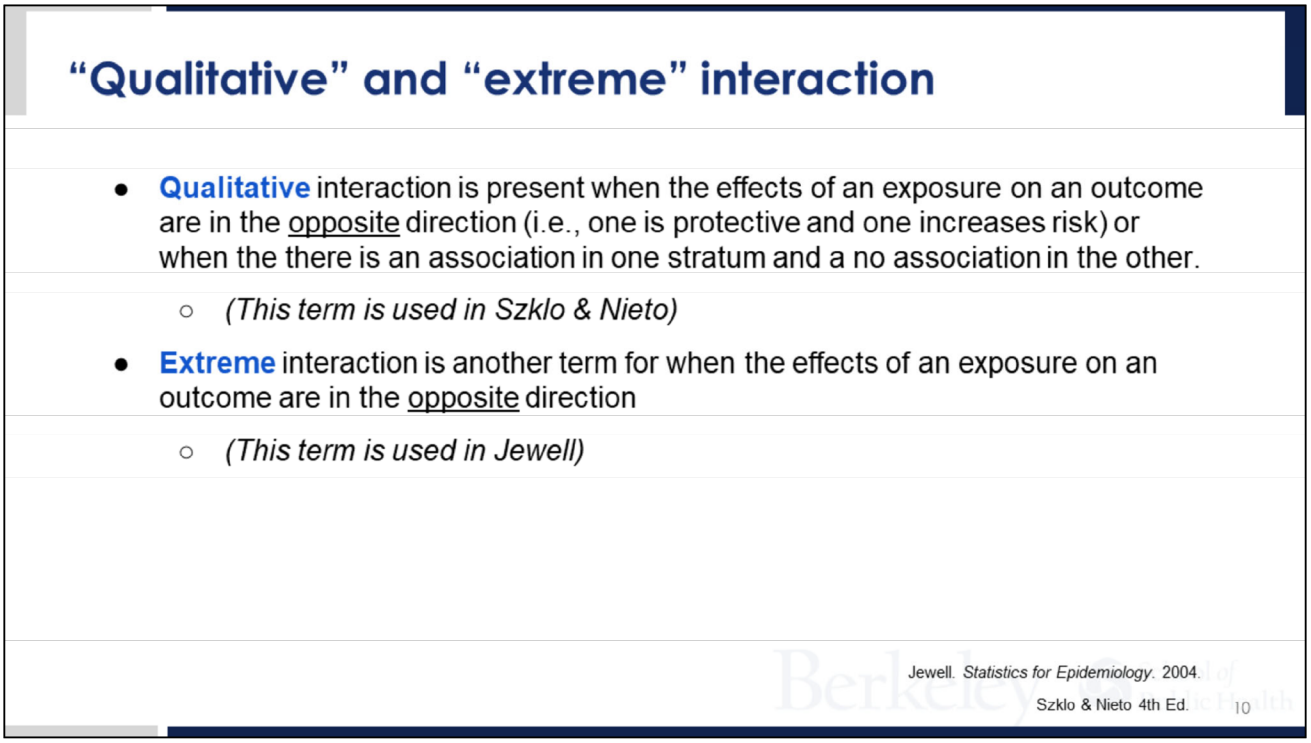
A related concept is essentially the opposite, which is antagonism or negative interaction. And here, the observed joint effect of risk factors A and Z is smaller than the expected risk on the basis of summing their independent effects. So if we sum A and Z here, we would expect to see the full length of the bar including the open area. That's our expected risk. But if we observe risk of A plus Z that's less than that sum, then we say there's negative interaction or antagonism.



There are two models that exist to describe how biologic or causal interaction occurs. This is the mechanistic model and the potential outcome model. So the mechanistic model is that the notion of direct physical or chemical reactions among exposures cause disease.

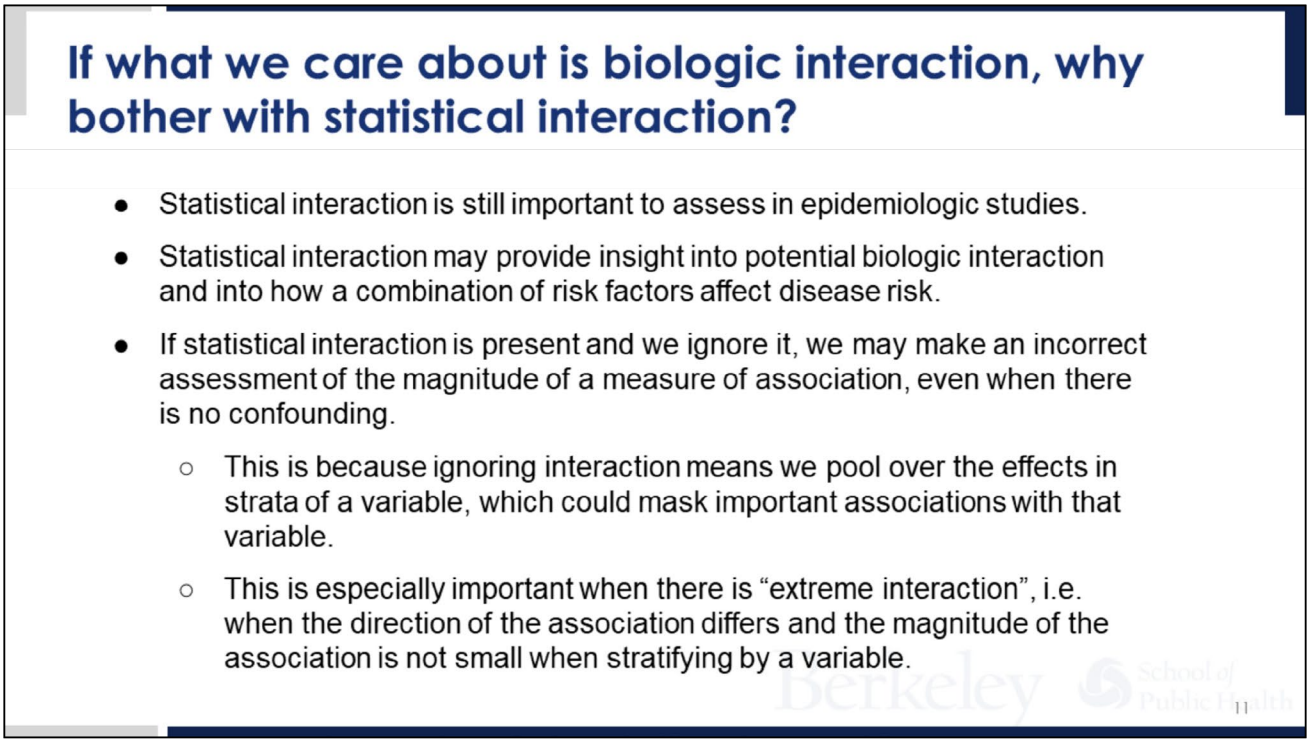
So for example, if free radicals in tissues are quenched by miscellaneous antioxidants, that's an example of a mechanistic way in which interaction is occurring. It's rare for a mechanism to account for all observed cases of disease. And a relationship between exposure and disease can be predicted from numerous different mechanisms of disease development, even when there's no bias present.

Another model for biologic interaction, or to understand biologic interaction, is the potential outcome also called the counterfactual causal model or the sufficient component cause model. These general causal models do not depend on specific mechanistic models. And we'll have upcoming videos on these topics later in the course.



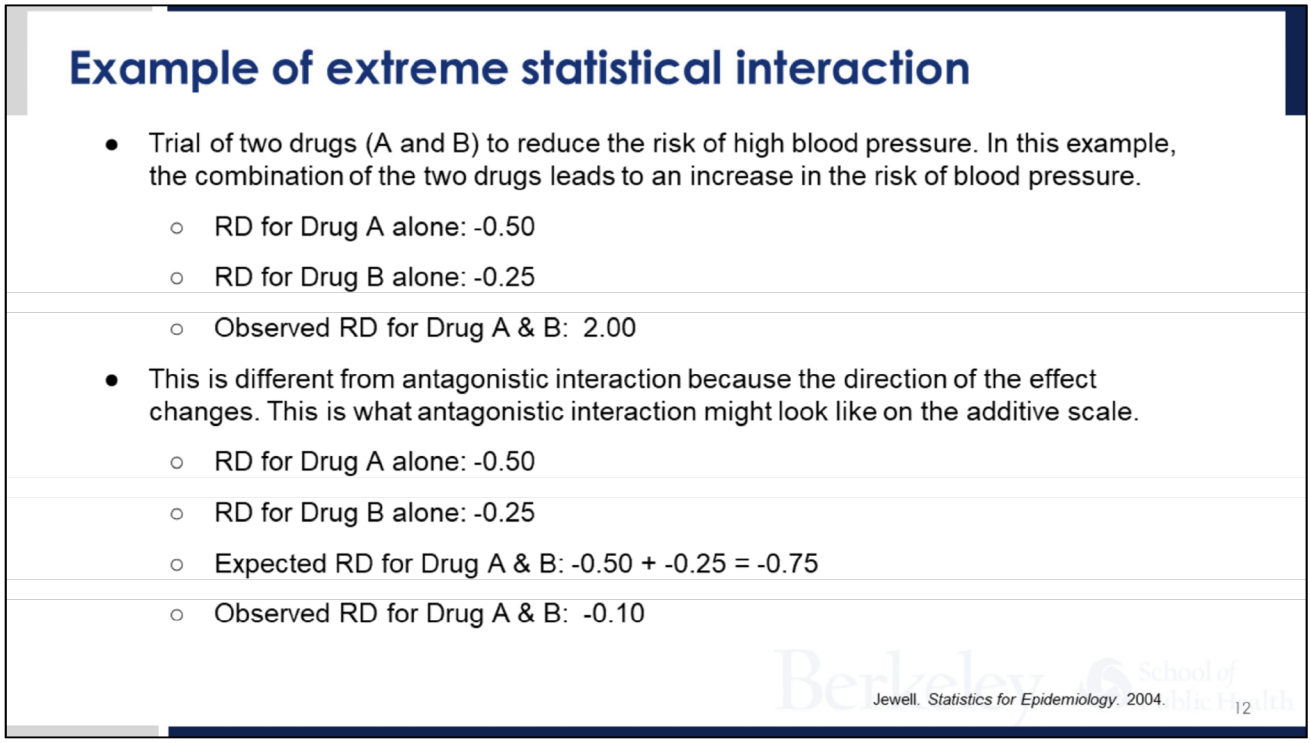
Another two terms to know are qualitative and extreme interaction. Qualitative interaction refers to the situation when the effects of an exposure on an outcome move in opposite directions. That is, one exposure is protective, and one increases risk, or when there is an association in one stratum or no association in another strata. This is the term used for this situation in Szklo and Nieto.

Another way this phenomenon is referred to in other textbooks, like the Jewell textbook, is extreme interaction. This is another term for this situation when effects of an exposure on an outcome are moving in opposite directions. So that's called extreme interaction.



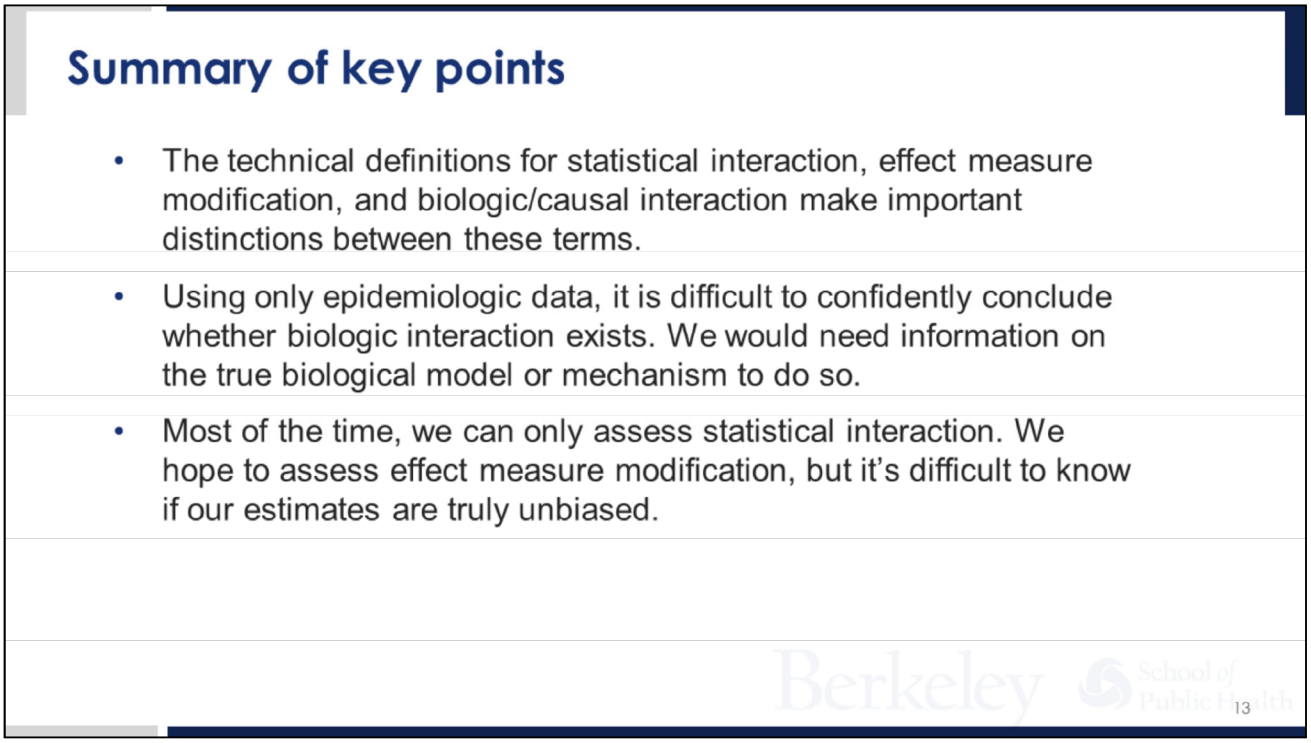
So if we mostly care about biologic interaction, why do we bother with statistical interaction? Well, it's still important to assess statistical interaction in epi studies, and it might provide insight into biologic interaction and into how a combination of risk factors might affect disease risk. If statistical interaction is present and we ignore it, we might make an incorrect assessment of the magnitude of a measure of association even when no confounding is present.

This is true, because ignoring interaction means that we pool over the effects in the strata of a variable, and that could mask important associations with that variable. So stratification can be very important. And this is especially true when there's extreme interaction. That is when the direction of the association differs in the strata, and the magnitude of the association is not small when stratifying by that variable.

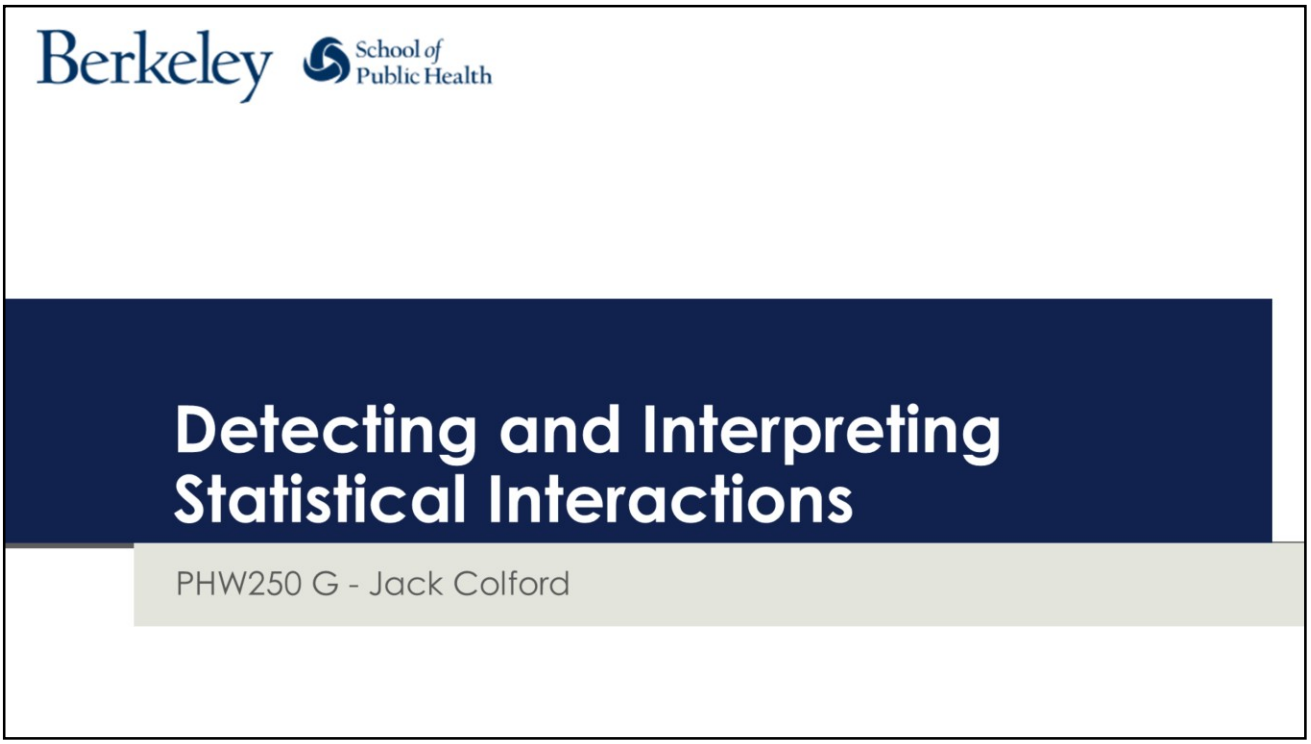


Let's review an example of extreme statistical interaction. Imagine that we're doing a trial of two different drugs, A and B versus some control, and our goal is to reduce the risk of high blood pressure. In this example, the combination of the two drugs leads to an increase in the risk of blood pressure. So for example, if the risk difference for drug A versus control alone was minus 0.50 and the risk difference for drug B alone was minus 0.25, but we actually observed when we combined drug A and B versus control a risk difference of 2, this is an example of extreme statistical interaction. And this is different from antagonistic interaction, because the direction of the effect change.

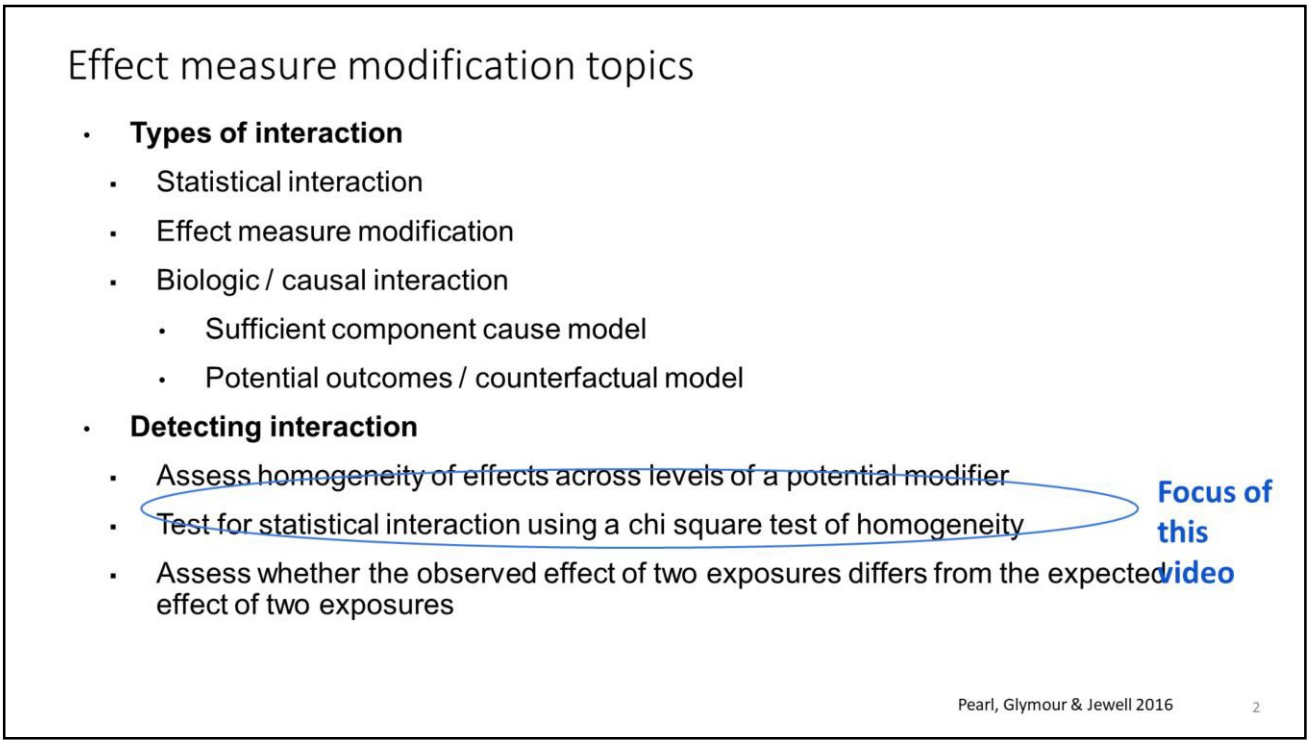
In contrast, what antagonistic interaction might look on the additive scale is in this next example where the risk difference for drug A alone imagine that that's minus 0.5, the risk difference for drug B alone minus 0.25, the expected risk difference for the two drugs together A and B would be minus 0.5 plus minus 0.25, which would equal minus 0.75. But if we observed the risk difference for drug A and B and saw a minus 0.10, that would just be an example of antagonistic interaction. It's still in the same direction, but it's less.



So in summary the technical definitions for statistical interaction, effect measure modification, and biologic or causal interaction make important distinctions between these terms. Using only epi data, it's difficult to competently conclude whether biologic interaction exists. We would need information on the true biological model or the mechanism in order to do so. And finally, most of the time we can only assess statistical interaction. We hope to assess effect measure modification, but it's difficult to know if our estimates are truly unbiased.



PRESENTER: In this video, Professor Colford will go over detecting and interpreting statistical interactions as they relate to effect measure modification.



JACK COLFORD: We've been talking about effect modification or effect measure modification. And we started off by discussing how there are different types of interaction, including statistical interaction, effect measure modification, and biologic or causal interaction. There are several ways to detect whether interaction is present in data.

And they include assessing whether the homogeneity of effects exist across levels of a potential modifier. That will be the focus of this video. I can also test for statistical interaction between the strata using a chi square test of homogeneity. And finally, I can assess whether the observed effect of two exposures differs from the expected effect of two exposures.

Interpreting interactions

An observed interaction in a dataset could reflect any of the following:

- A true biologic, causal interaction
- Heterogeneity due to random variability
- Heterogeneity due to confounding
- Heterogeneity due to bias
- Heterogeneity due to differential intensity of exposure

Szklo & Nieto 4th Ed.

3

So when we see evidence that there is interaction in a dataset, there could be a number of different things going on. This could be a true biologic causal interaction. Or it might be due to heterogeneity from random variability. Or it might be due to heterogeneity from confounding. Or it might be due to heterogeneity from bias. Or it might be due to heterogeneity from a differential intensity of exposure. And we'll talk through each of these.

Heterogeneity due to random variability

- Random variability can be produced by the stratification of a potential effect modifier
- This is most likely to occur when an investigator obtains null results in a study and wonders whether a result occurs in certain subpopulations, so he carries out a stratified analysis (e.g., estimate the association by gender or education level).
- The sample size is smaller to estimate the same measure of association in subgroup analyses.
- As a result, precision is lower, increasing the probability of finding heterogeneous measures of association due to chance rather than a true causal interaction.
- This is why it is best to pre-specify potential effect modifiers.

Szklo & Nieto 4th Ed.

4

Let's first talk about heterogeneity possibly due to the random variability. Random variability can be produced by stratification of a potential effect modifier. And this is most likely to occur when an investigator obtains null results in a study and wonders whether a result might have been different in certain subpopulation. So the investigator carries out a stratified analysis and then examines the association by, for example, gender or education level or some potential effect-modifying variable.

The sample size is smaller in each of these strata. And so, when you estimate the measure of association in each of the strata, you have a much smaller sample size for doing these subgroup analyses. Because of the smaller sample size, the precision is lower, and this increases the probability of finding heterogeneous measures of association that are due just to random chance rather than to a true causal interaction. That's why it's best to prespecify potential effect modifiers so you don't go on a hunt and just keep stratifying until you find relationships that might just be due to random chance.

Heterogeneity due to confounding

- Differential confounding across strata may explain observed heterogeneity in the measure of association.
- Confounding could either exaggerate or decrease heterogeneity.
- For this reason, adjusting for confounders while assessing effect modification is important. This can be done using multivariate statistical models (more on this later in the course).

Szklo & Nieto 4th Ed.

5

All right. Another reason that interaction might appear to be present is heterogeneity due to confounding. So if a dataset has differential confounding across strata, that might explain observed heterogeneity in the measure of association. The confounding could either exaggerate or decrease the heterogeneity. And for this reason, adjusting for confounders while assessing the effect modification is important. This can be done using multivariate statistical models. And we'll talk more about this later in the course.

Example: Heterogeneity due to confounding

Gender/smoking	Coffee intake	Cases	Controls	Odds ratio
Female/ nonsmoker	Yes	10	10	
	No	90	90	1.0
	Total	100	100	
Male/total	Yes	38	22	
	No	62	78	2.2
	Total	100	100	
Male/smoker	Yes	35	15	
	No	35	15	1.0
	Total	70	30	
Male/ nonsmoker	Yes	3	7	
	No	27	63	1.0
	Total	30	70	

Assume that smoking causes cancer *Y*, 50% of smokers but only 10% of nonsmokers drink coffee, coffee intake is not independently related to cancer *Y*, all females are nonsmokers, and 70% of male cases and 30% of male controls are smokers.

- Comparing the stratified to pooled ORs for males, we see that smoking is a confounder in males.
- This causes apparent heterogeneity in the ORs for males vs. females (OR=2.2 vs. 1.0).
- If the OR for males were adjusted for confounding, the apparent heterogeneity would disappear.

Szklo & Nieto 4th Ed.

6

Let's talk through this example of heterogeneity that's brought about or due to confounding.

PRESENTER: Let's assume that smoking causes cancer Y. 50% of smokers, but only 10% of nonsmokers drink coffee. Coffee intake is not independently related to cancer Y. All females are nonsmokers, and 70% of male cases and 30% of male controls are smokers. We are looking at the odds ratios between coffee intake and cancer.

JACK COLFORD: So in this table, when we compare the stratified pooled odds ratios for males, we see that smoking is a confounder in males. And the way we can tell this is that the male smokers have a relative risk or odds ratio of 1. The male nonsmokers have an odds ratio of 1. So the adjusted estimate for males would be 1.

But that's quite different than the unadjusted estimate for males, which is 2.2. If the odds ratio for males were adjusted for confounding, the apparent heterogeneity would disappear. So if we adjust for confounding among the males, this heterogeneity disappears, and there's no relationship between gender and cancer.

Heterogeneity due to bias

- Apparent heterogeneity may be due to differential bias between strata
- Bias could either exaggerate or decrease heterogeneity.
- Example: study of race, education, and miscarriage

	White		Black		
	Number	Risk/100	Number	Risk/100	Black/white ratio
Total	325	7.7	93	5.5	0.7
Mother's years of education					
< 9	12	10.4	0	–	–
10–11	52	8.0	15	4.5	0.6
12	111	6.3	44	4.7	0.7
≥ 13	150	9.2	33	9.5	1.0

- Only looking at the "Total" row, black women had a lower risk of miscarriage than white women.
- The authors believe this was due to underascertainment of miscarriage among blacks.
- Stratifying by education level, the absence of data among women with <9 years of

Szklo & Nieto 4th Ed. 7

Another situation can occur is when we have heterogeneity due to bias. So the apparent heterogeneity here may be due to differential bias between the strata. And this bias could either exaggerate or decrease heterogeneity. So as an example, we might study race, education, and miscarriage. And so if we only look at the total row here, the black women had a lower risk of miscarriage than the white women.

The authors believed that this was due to under ascertainment of miscarriage among blacks because, when they stratified by education level, there was an absence of data among black women with less than nine years of education. So it seems that there must have been systematic underreporting in this category because it's not really credible to believe that there would be no miscarriages in that group of black women.

Heterogeneity due to differential intensity of exposure

- Heterogeneity can occur when the level of exposure to a risk factor is associated with the potential effect modifier's levels.
- Example: study of asthma and airborne soy dust
  - RR = 4.4 when wind speed was <12 miles per hour
  - RR = 1.7 when wind speed was ≥ 12 miles per hour
  - Slow wind speed may have caused heavier exposure to soy dust
  - If so, this does not reflect a true biological interaction between soy dust and wind speed.
  - This does not mean that this information is not useful — it can inform public health intervention to reduce exposure to soy dust.

Szklo & Nieto 4th Ed.

8

Heterogeneity can also arise from a differential intensity of exposure. And by this, we mean that heterogeneity can occur when the level of exposure to a risk factor is associated with the potential effect modifiers' levels. Here's an example. Some investigators were studying asthma in airborne soy dust. When the wind speed was less than 12, the relative risk for the relationship between the dust and asthma was 4.4.

But when the wind speed was higher, greater than 12 miles per hour, the relative risk was 1.7. So they hypothesized that slow wind speed may have caused heavier exposure to soy dust. If so, this doesn't reflect a true biological interaction between soy dust and wind speed. This doesn't mean that the information is not useful. It can still inform public health interventions to reduce exposure to soy dust. But this apparent interaction was due only to wind speed, not to a true biologic relationship.

Detecting statistical interaction / effect measure modification

Three methods covered in this course:

1. Assess homogeneity of effects across levels of a potential modifier
2. Test for statistical interaction using a chi square test of homogeneity
3. Assess whether the observed effect of two exposures differs from the expected effect of two exposures
    - Modern causal approaches to assessing effect modification use this method

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

9

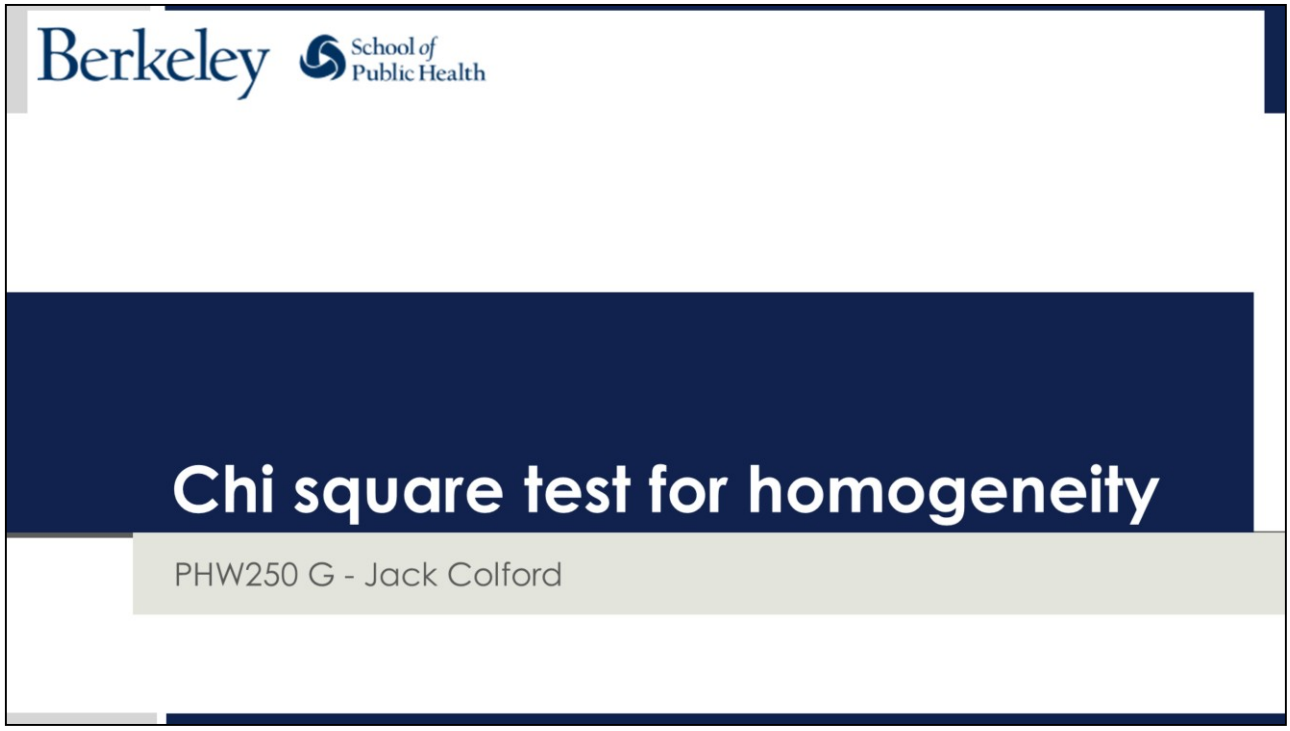
In order to detect statistical interaction or effect measure modification, in this course, we'll use three different methods. We can assess the homogeneity effects across levels of a potential modifier. We can test for statistical interaction using a chi square test of homogeneity. And we can assess whether the observed effect of two exposures differ from the expected effect of two exposures. So modern causal approaches to assessing effect modification use this method, this method number three.

Summary of key points

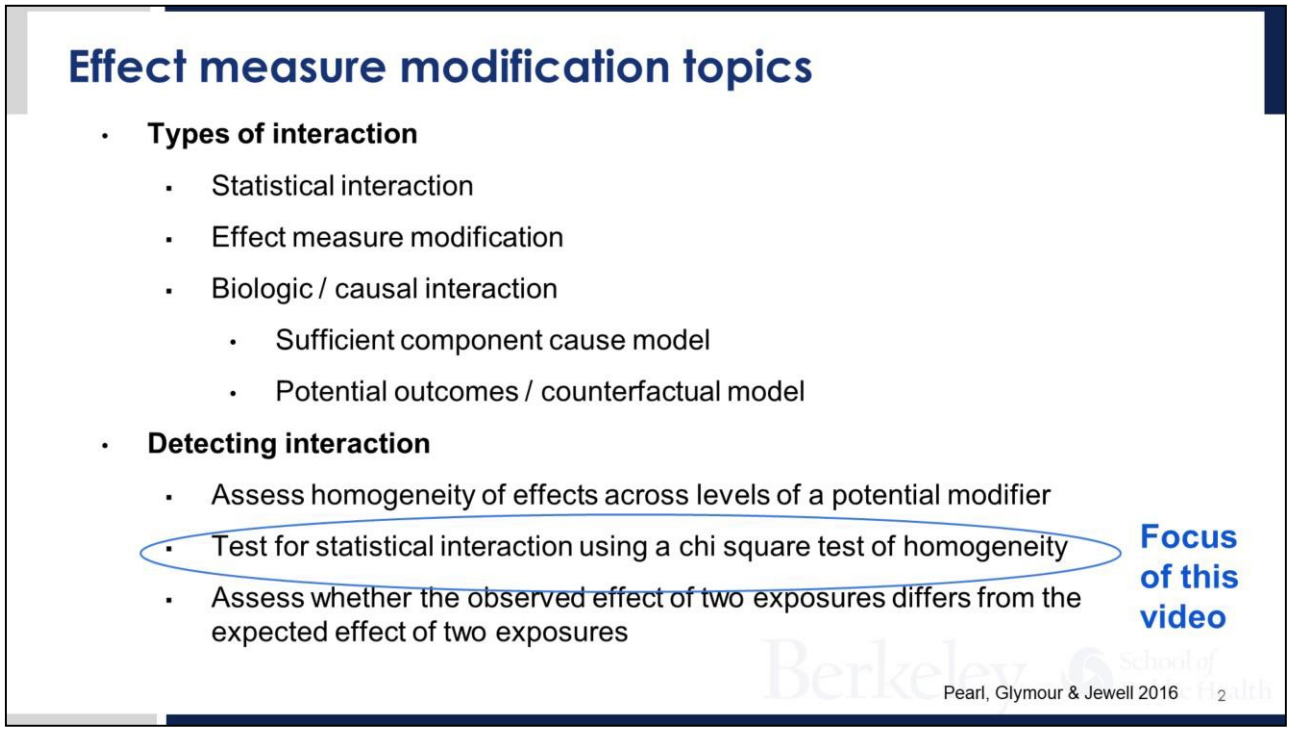
- Random error, confounding, bias, and other factors besides true biologic / causal interaction may be responsible for observed heterogeneity of effects.

10

So in summary, random error, confounding, bias, and other factors besides true biologic or causal interaction, may be responsible for the observed heterogeneity of affects.

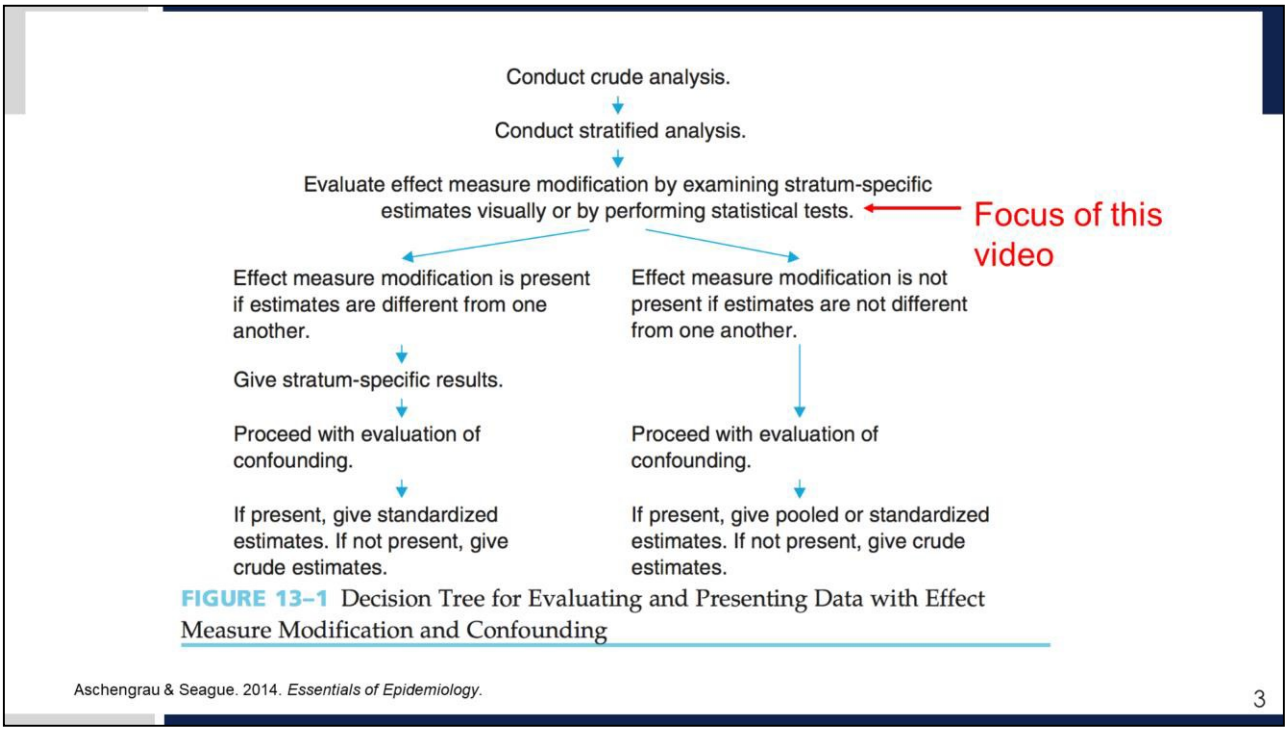


JACK COLFORD: Our next topic is the chi-square test for homogeneity.



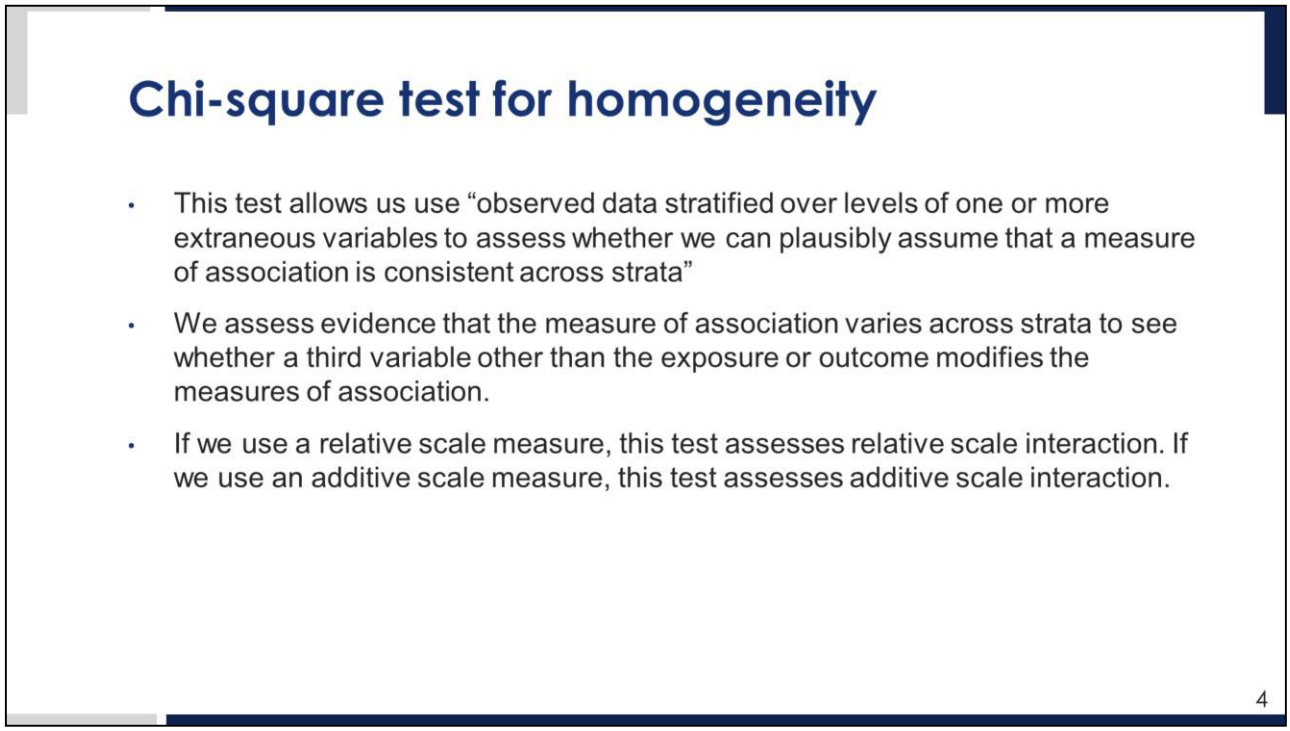
We've been talking about effect modification or effect measure modification, and we started off by discussing how there are different types of interaction, including statistical interaction, effect measure modification, and biologic or causal interaction. There are several ways to detect whether interaction is present in data, and they include assessing whether the homogeneity of effects exist across levels of a potential modifier.

And by that, I mean when I stratify my data by some other variable-- for example, if I stratify my data on smoking and lung cancer by age, one stratum is old, one stratum is young, then I evaluate whether age is an effect modifier of the relationship between smoking and lung cancer. I can also test for statistical interaction between the strata using a chi-square test of homogeneity. That will be the focus of this video. And finally, I can assess whether the observed effect of two exposures differs from the expected effect of two exposures.



This flow chart is very useful and is something you'll become quite accustomed to producing on your own or flowing through on your own as you start to work with the sorts of analyzes. But just to start at the top of the flow chart, where we'll be working today, when an epidemiologist proceeds to analyze data, the first step, of course, is to conduct a crude analysis. So say I'm examining the relationship between smoking and lung cancer. I would look at smokers versus nonsmokers, and look at the relative risk difference for lung cancer, as we've always done in the course so far.

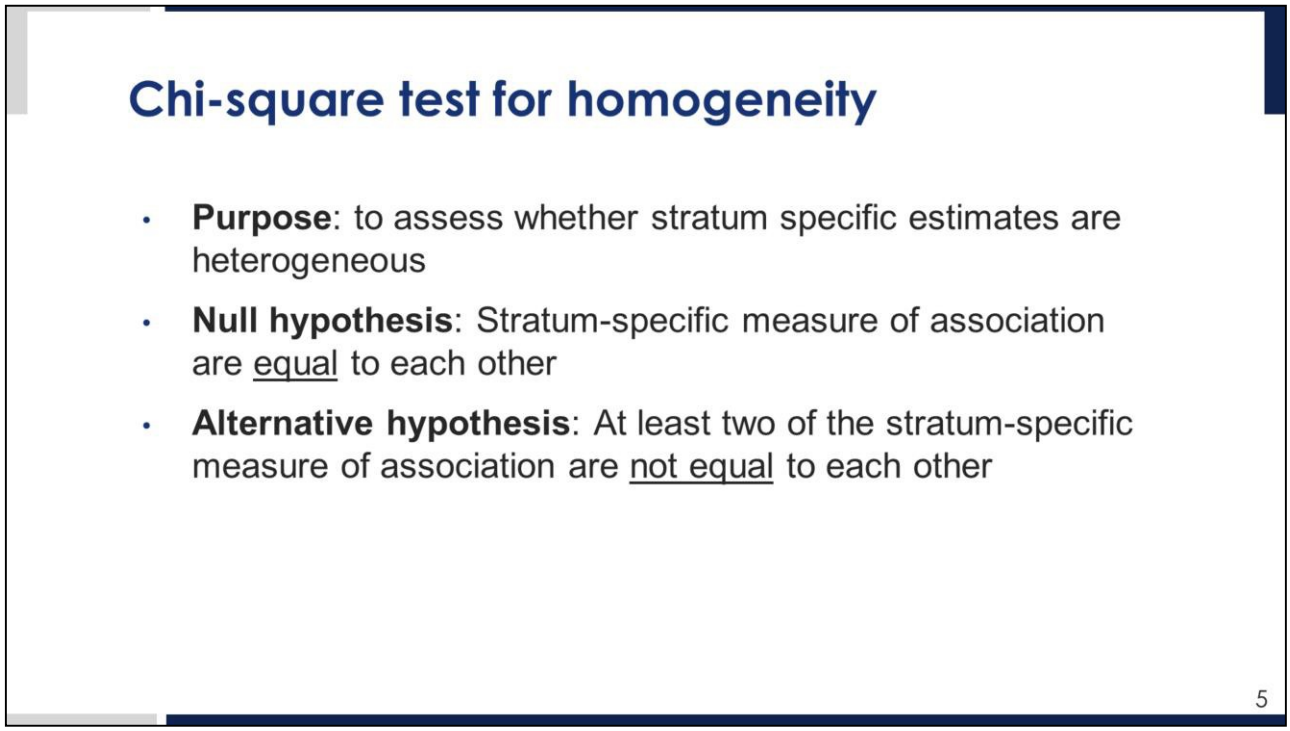
The next step would be to conduct a stratified analysis where I stratify the data by a third variable of interest. So for example, if I'm interested in examining the effects of age on a relationship between smoking and lung cancer, I would stratify my data into the old and the young if I were picking two levels, and then I would re-analyze the relationship between smoking and lung cancer among the old and smoking and lung cancer among the young. This step will be the focus of this video.



In words, the chi-square test allows us to use our observed data stratified over levels of one or more extraneous variables to assess whether we can plausibly assume that a measure of association is consistent across the strata. So simplifying that, that means when I stratify my smoker lung cancer data into the old and the young, I look at the risk in the old and I look at the risk in the young, and I see whether there's any homogeneity or difference, which would be heterogeneity, between the two different strata.

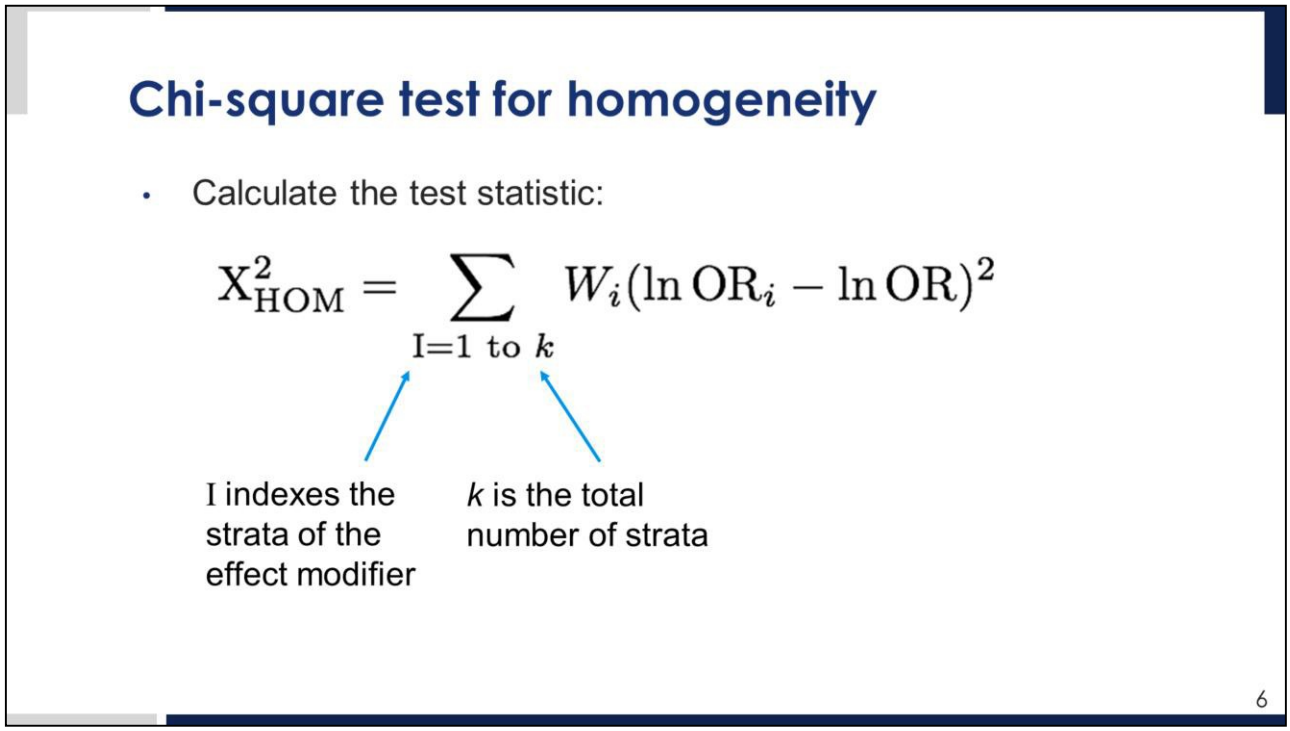
We assess the evidence that the measure of association varies across strata to see whether a third variable other than the exposure or outcome modifies the measure of association. So if I see that the relative risk is 4.0 in the old smokers and 2.0 in the young smokers, I need a test to tell whether 4.0 is different than 2.0. If we use a relative scale measure. This test of homogeneity assesses relative scale interaction. If we use an additive scale measure, this test assesses additive scale interaction.

Recall that the relative scale refers to relative risks that we work with, such as the odds ratio, the cumulative incidence ratio, the incidence density ratio, whereas the additive scale refers to absolute risks, such as the risk difference.



The purpose of the chi-square test for homogeneity is to assess whether the stratum-specific estimates are heterogeneous or homogeneous. The null hypothesis is that the stratum-specific measures of association are equal to each other, that is that they're homogeneous.

The alternative hypothesis is that at least two of the stratum-specific measures of association are not equal to each other. So again, if I look at smokers who are old and compare them to smokers who are young, those are my two strata, young and old. If those two numbers are statistically equal to each other, that's homogeneity. If they're different from each other, that's heterogeneity. My null hypothesis with this statistical test is that the strata are equal to each other.



This is the formula for the chi-square test for homogeneity. It may be a little imposing when you first see it, but let's talk it through, and we'll use it a bit during this video. So the x squared is really meant to represent the Greek character chi, and hom refers to homogeneity. So this is telling us to calculate the chi-square value for homogeneity, we do the following, the right side of the equation.

Well, we are going to sum over several different levels. And the levels refer to the levels of strata that we have. So if I've stratified into old and young, I have two strata. So that i underneath the grand sum that goes from 1 to k is called the index. So we index the strata from 1 to k, where k is the number of stratum, the highest stratum. So if we have two, we go from one to two. I would go from one to two.

Then we are at each stratum. So for the old and for the young separately, because we have two strata, we would do the following calculation. For each stratum, we would calculate a weight for that stratum. And we'll go through that in more detail in a moment. And then in each stratum, we would do the following. We would calculate an odds ratio for that stratum, take the logarithm of it, and then subtract from it the adjusted odds ratio, and take the logarithm of that, and square that quantity. Now, what is that adjusted log odds ratio? We're going to describe that in a little bit, because that's a summary odds ratio that's averaged across all the strata. It's not the crude odds ratio. Rather, it's a statistically averaged odds ratio across the strata.

Chi-square test for homogeneity

- Calculate the test statistic:

$$X^2_{HOM} = \sum_{I=1 \text{ to } k} W_i(\ln OR_i - \ln OR)^2$$

$$W_i = [1/a_i + 1/b_i + 1/c_i + 1/d_i]^{-1}$$

W_i is a weight that is inversely proportional to the variance of ln OR_i

7

So to calculate this chi-square statistic, we of course need this weight for each strata. So that's wi. This weight is a weight that's inversely proportional to the variance of the log of the odds ratio for that stratum. Now, you can't intuit this by looking at the formula, but the formula for the weight of each stratum is estimated by this a quantity, which is the sum of the reciprocal of all four cells for the two-by-two table for that stratum.

Again, I realize in saying this in words that this might seem confusing at first, but as you start to work with the problems and see the solutions, you'll see how really simple this is to do. But anyway, the weight for each stratum, the old and the young, if that were what I were doing, would be the sum of one over cell a plus 1 over cell b plus one over cell c plus one over cell d. Then I would do that again for the second stratum, the young.

Chi-square test for homogeneity

- Calculate the test statistic:

$$X^2_{HOM} = \sum_{I=1 \text{ to } k} W_i(\ln OR_i - \ln OR)^2$$

$$W_i = [1/a_i + 1/b_i + 1/c_i + 1/d_i]^{-1}$$

OR_i is the stratum-specific odds ratio

8

In the formula, the log of the odds ratio sub i is the stratum-specific odds ratio. So for the old, I would do the odds ratio, which would be a times d divided by b times c.

Chi-square test for homogeneity

- Calculate the test statistic:

$$X^2_{HOM} = \sum_{I=1 \text{ to } k} W_i(\ln OR_i - \ln OR)^2$$

$$W_i = [1/a_i + 1/b_i + 1/c_i + 1/d_i]^{-1}$$

$$\ln OR = \frac{\sum_{I=1 \text{ to } k} w_i \ln OR_i}{\sum_{I=1 \text{ to } k} w_i}$$

Can be:

1) the Mantel-Haenszel OR or

2) a weighted OR
(focus of this video)

9

And I also need this other term, which is the log of the adjusted odds ratio. Now, the adjusted odds ratio can be calculated in multiple different ways. It can be calculated as what's called a Mantel-Haenszel odds ratio, which we won't discuss today, or a weighted odds ratio, which will be the focus of this video.

So the weighted odds ratio that we'll use today is given by the formula to the left, where this log of the odds ratio average odds ratio, or weighted odds ratio, is given by this grand sum, again, going from one to two, because there are two strata in our example here, times each stratum-specific weight. That's the same weight that we've used before. Times the log of the odds ratio for that stratum, divided by the sum of all the weights. And that's why this is called a weighted average.

Chi-square test for homogeneity

- Calculate the test statistic:

$$X^2_{HOM} = \sum_{I=1 \text{ to } k} W_i(\ln OR_i - \ln OR)^2$$

- The **test-statistic** follows a chi-square distribution.
- Obtain a **p-value** for the test using a chi-square table or statistical software, such as R.
  - R command: `dchisq(#,df=#)`
  - df = degrees of freedom = k-1

10

Next we would calculate the test statistic. So after plugging in all these numbers for these various pieces in the different strata, we know that this chi-square for homogeneity test statistic follows a chi-square distribution. Hopefully you remember that from your biostatistics introductory courses. So we would get a p value for the test using a chi-square table or a statistical software such as R. The R command for this would be d chi-square, and then in parentheses number, comma, and then df equals number, which stands for degrees of freedom. And that's one less than the number of strata. So if there are two strata, the degrees of freedom would be 2 minus 1, or 1. So we simply look this up in a table, or more commonly plug it into R, and have R give us the p value for what we determined.

Example

TABLE 13–3 Data from a Hypothetical Case–Control Study of DDE Exposure and Breast Cancer Stratified According to Lactation History

	Never breastfed		*Breastfed*	
DDE level	*Cases*	*Controls*	*Cases*	*Controls*
High	140	300	360	300
Low	550	2,600	950	800
Total	690	2,900	1,310	1,100
	Stratum-specific odds ratio = 2.2		Stratum-specific odds ratio = 1.0	

Null hypothesis: The OR for DDE level and breast cancer among women who breastfed and the OR for women who never breastfed are equal.

Alternative hypothesis: The OR for DDE level and breast cancer among women who breastfed and the OR for women who never breastfed are different.

Aschengrau & Seague. 2014. *Essentials of Epidemiology*.

11

This is much easier to understand when we work through an example. Let's look at an example looking at the risk for DDE exposure causing breast cancer. And these investigators are examining whether breast feeding is an effect modifier of the relationship between DDE level and cancer. This is hypothetical breast cancer. So notice that we have two strata here on the left and the right. We have the never breastfed and we have the breastfed.

So each of these two strata has a two-by-two table, which is the usual a, b, c, d that we're used to seeing. On the never breastfed, we have cell a is 140, cell b is 300, cell c is 550, cell d is 2,600, and so forth. If you calculate the odds ratio for the never breastfed children-- you should be able to do this easily-- that's an odds ratio of 2.2. Then calculate the odds ratio for the breastfed children, and that's a stratum-specific odds ratio of 1.0. Stratum-specific, of course, just means for each stratum.

Well, the odds ratio for DDE level and breast cancer among women who breastfed and the odds ratio for women who never breastfed are equal would be a statement of the null hypothesis. So I think you understand that even if these aren't numerically the same, it's possible that they're statistically the same, and that's what the null hypothesis is stating.

If we do our chi-square test and we reject the null hypothesis, then we would conclude the opposite or the alternative hypothesis, which is the odds ratio for the DDE level and breast cancer among women who breastfed and the odds ratio for

women who never breastfed are different. So in simple terms, we're asking, is 2.2 different than 1.0 statistically, yes or no?

TABLE 13–3 Data from a Hypothetical Case–Control Study of DDE Exposure and Breast Cancer Stratified According to Lactation History

	Never breastfed i=1		*Breastfed* i=2	
DDE level	*Cases*	*Controls*	*Cases*	*Controls*
High	140	300	360	300
Low	550	2,600	950	800
Total	690	2,900	1,310	1,100
	Stratum-specific odds ratio = 2.2		Stratum-specific odds ratio = 1.0	

$$X^2_{HOM} = \sum_{I=1 \text{ to } k} W_i(\ln OR_i - \ln OR)^2 \qquad W_i = [1/a_i + 1/b_i + 1/c_i + 1/d_i]^{-1}$$

$W_1 = (1/140 + 1/300 + 1/550 + 1/2600)^{-1} = 78.87$

$W_2 = (1/360 + 1/300 + 1/950 + 1/800)^{-1} = 118.85$

Aschengrau & Seague. 2014. *Essentials of Epidemiology*.

12

Here are the data from the table put into the formula we need to use to make this calculation. So notice that the two strata are indexed as i equals 1 and i equals 2. The never breastfed as i equals 1. The breastfed is i equals 2. What do we need to do? We need to sum for each stratum the weight of the table. So the weight for the first stratum, 1 over a plus 1 over b plus 1 over c plus 1 over d would be 1 over 140 plus 1 over 300 plus 1 over 550 plus 1 over 2,600. And that adds up to 78.87. And we would do the same thing for the second stratum, the breastfed. 1 over 360 plus 1 over 300 plus 1 over 950 plus 1 over 800 equals 118.85.

TABLE 13–3 Data from a Hypothetical Case–Control Study of DDE Exposure and Breast Cancer Stratified According to Lactation History

	Never breastfed i=1		*Breastfed* i=2	
DDE level	*Cases*	*Controls*	*Cases*	*Controls*
High	140	300	360	300
Low	550	2,600	950	800
Total	690	2,900	1,310	1,100
	Stratum-specific odds ratio = 2.2		Stratum-specific odds ratio = 1.0	

$$X^2_{HOM} = \sum_{I=1 \text{ to } k} W_i(\ln OR_i - \ln OR)^2$$

$\ln OR_1 = \ln (140 \times 2600 / 300 \times 550) = 0.791$

$\ln OR_2 = \ln (360 \times 800 / 950 \times 300) = 0.010$

Aschengrau & Seague. 2014. *Essentials of Epidemiology*.

13

The next piece of information we need is the odds ratio for each stratum. So the odds ratio for the first stratum is calculated. And actually we're going to take the logarithm, because the formula is the log of the odds ratio. So the logarithm for the first stratum is the logarithm of 140 times 2,600 divided by 300 times 550. That is 0.791. So we're keeping the first stratum, the never breastfed in blue. The second stratum odds ratio is the log of a times d over b times c, or 360 times 800 over 950 times 300. So that's 0.010.

TABLE 13–3 Data from a Hypothetical Case–Control Study of DDE Exposure and Breast Cancer Stratified According to Lactation History

	Never breastfed i=1		*Breastfed* i=2	
DDE level	*Cases*	*Controls*	*Cases*	*Controls*
High	140	300	360	300
Low	550	2,600	950	800
Total	690	2,900	1,310	1,100
	Stratum-specific odds ratio = 2.2		Stratum-specific odds ratio = 1.0	

$$X^2_{HOM} = \sum_{I=1 \text{ to } k} W_i(\ln OR_i - \ln OR)^2 \qquad \ln OR = \frac{\sum_{I=1 \text{ to } k} w_i \ln OR_i}{\sum_{I=1 \text{ to } k} w_i}$$

$W_1 = 78.87$
$W_2 = 118.85$
$\ln OR_1 = 0.791$
$\ln OR_2 = 0.010$

$\ln OR = [(78.87 \times 0.791)+(118.85 \times 0.010)] / (78.87+118.85) = 0.322$

Aschengrau & Seague. 2014. *Essentials of Epidemiology*.

14

And finally, the last piece of the puzzle we need is to calculate this adjusted odds ratio. And that formula is given here using the weights and the logarithm of each stratum. The formula is given on the right side of the equation here as well. So if you put in all the numbers here, you get a value of 0.322 for the logarithm of the adjusted odds ratio.

TABLE 13–3 Data from a Hypothetical Case–Control Study of DDE Exposure and Breast Cancer Stratified According to Lactation History

	Never breastfed i=1		*Breastfed* i=2	
DDE level	*Cases*	*Controls*	*Cases*	*Controls*
High	140	300	360	300
Low	550	2,600	950	800
Total	690	2,900	1,310	1,100
	Stratum-specific odds ratio = 2.2		Stratum-specific odds ratio = 1.0	

$$X^2_{HOM} = \sum_{I=1 \text{ to } k} W_i(\ln OR_i - \ln OR)^2 \qquad \ln OR = \frac{\sum_{I=1 \text{ to } k} w_i \ln OR_i}{\sum_{I=1 \text{ to } k} w_i}$$

X_2^{HOM} = 78.87 x (0.791-0.322)^2 + 118.85 x (0.010-0.322)^2 = 28.918

W_1 = 78.87
W_2 = 118.85
ln OR_1 = 0.791
ln OR_2 = 0.010
ln OR = 0.322

Aschengrau & Seague. 2014. *Essentials of Epidemiology.*

15

So putting in, substituting all the numbers we have for the formula for the chi-square of homogeneity, we have a chi-square value of 28.918.

$$X^2_{HOM} = \sum_{I=1 \text{ to } k} W_i(\ln OR_i - \ln OR)^2$$

X_2^{HOM} = 28.918

Degrees of freedom (df) = 2-1 = 1
Get the p-value in R: **dchisq**(28.918, **df**=1) = <0.001

Conclusion: We can reject the null hypothesis that the ORs are the same across strata of breastfeeding status. Effect modification on the relative scale is present.

- For the chi-square test of homogeneity, we use a p-value cutoff for statistical significance of 0.2
  - (i.e., p-value <0.2 is statistically significant).
- This is because the statistical test of homogeneity has limited statistical power
- We err on the side of concluding that interaction is present

16

Next we need to look up how statistically significant that is. So we have a chi-square value for homogeneity of 28.918. The degrees of freedom for this chi-square value are 2 minus 1 because there were two strata. So that's one degree of freedom.

We get the p-value from R using the command d chi-square, 28.918 with one degree of freedom. That has a p-value of less than 0.001. Therefore, we can reject the null hypothesis that the odds ratios are the same across strata of breastfeeding status. Effect modification on the relative scale then is present, because we've rejected the null hypothesis.

For the chi-square test of homogeneity, we use a p-value cutoff for statistical significance of 0.20. So any p-value less than 2 is statistically significant for this test. This is because the statistical test of homogeneity has limited statistical power, so we use a higher p-value than you might be used to seeing in a clinical trial, for example, for a test of significance, which would be 0.05. We err on the side of concluding that interaction is present.

Formula for a risk ratio

$$X^2_{HOM} = \sum_{I=1 \text{ to } k} W_i(\ln OR_i - \ln OR)^2$$

This is the formula we just learned.

17

To this point, we've calculated the chi-square test for homogeneity using an odds ratio. We started there because that's a very simple formula, but there are formulae for other relative measures that we might want to use, or absolute measures, like the risk difference.

Formula for a risk ratio

$$X^2_{HOM} = \sum_{I=1 \text{ to } k} W_i(\ln OR_i - \ln OR)^2$$

$$X^2_{HOM} = \sum_{i=1}^{k} \frac{(\ln OR_i - \ln OR)^2}{W_i}$$

← It can be rewritten like this.

18

We can rewrite the chi-square test for homogeneity by some algebraic re-arrangement. And you see that formula here. Now the weight is now appearing in the denominator.

Formula for a risk ratio

$$X^2_{HOM} = \sum_{I=1 \text{ to } k} W_i(\ln OR_i - \ln OR)^2$$

$$X^2_{HOM} = \sum_{i=1}^{k} \frac{(\ln OR_i - \ln OR)^2}{W_i}$$

$$X^2_{HOM} = \sum_{i=1}^{k} \frac{(\ln OR_i - \ln OR)^2}{var(\ln(OR_i))}$$

The weights in the denominator of the formula are actually the variance of the ln(OR) in each stratum.

19

And the weights in the denominator of the formula are actually the variance of the log of the odds ratio in each stratum. That's not intuitive, but I can just tell you that that is true.

Formula for a risk ratio

$$X^2_{HOM} = \sum_{I=1 \text{ to } k} W_i(\ln OR_i - \ln OR)^2$$

$$X^2_{HOM} = \sum_{i=1}^{k} \frac{(\ln OR_i - \ln OR)^2}{W_i}$$

$$X^2_{HOM} = \sum_{i=1}^{k} \frac{(\ln OR_i - \ln OR)^2}{\text{var}(\ln(OR_i))}$$

$$X^2_{HOM} = \sum_{i=1}^{k} \frac{(\ln RR_i - \ln RR)^2}{\text{var}(\ln(RR_i))}$$

When we write the formula this way, we can swap in the RR (CIR or IDR) for the OR.

20

And so the formula for a risk ratio can be given when we write the formula this way in the bottom of this slide. We swap in the relative risk, which could be either cumulative incidence ratio or incidence density ratio for the odds ratio. So just rewriting it like this, we have this formulation.

Formula for a risk difference

$$X^2_{HOM} = \sum_{I=1 \text{ to } k} W_i(\ln OR_i - \ln OR)^2$$

$$X^2_{HOM} = \sum_{i=1}^{k} \frac{(\ln OR_i - \ln OR)^2}{W_i}$$

$$X^2_{HOM} = \sum_{i=1}^{k} \frac{(\ln OR_i - \ln OR)^2}{var(\ln(OR_i))}$$

$$X^2_{HOM} = \sum_{i=1}^{k} \frac{(RD_i - RD)^2}{var((RD_i))}$$

When we write the formula this way, we can swap in the RD for the OR. If we use the RD, we don't need to take the log

21

And similarly, for the risk difference, we have the formulation at the bottom here. If we use the risk difference, because we're working on an absolute scale, not a relative scale, we don't need to take the logarithm.

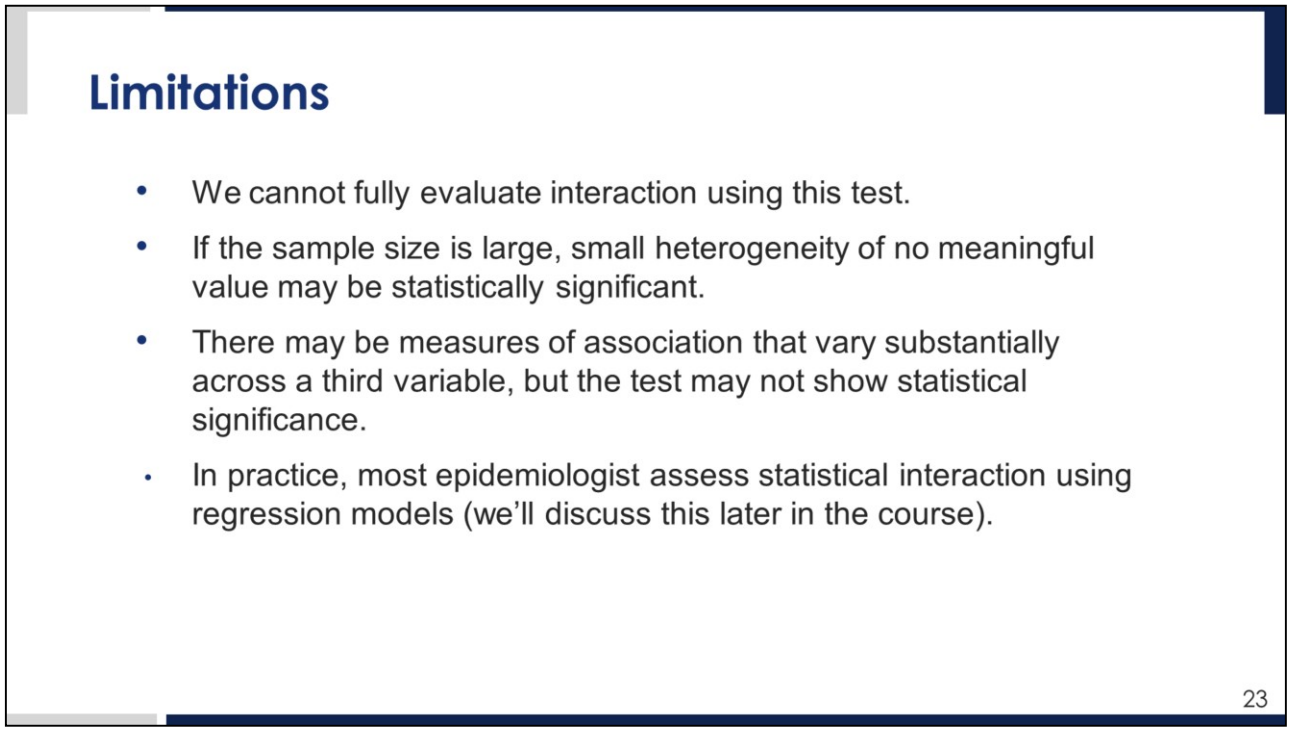
Table of formulas

General formulation: $X^2_{HOM} = \sum_{i=1}^{k} \frac{(MA_i - MA)^2}{var((MA_i))}$

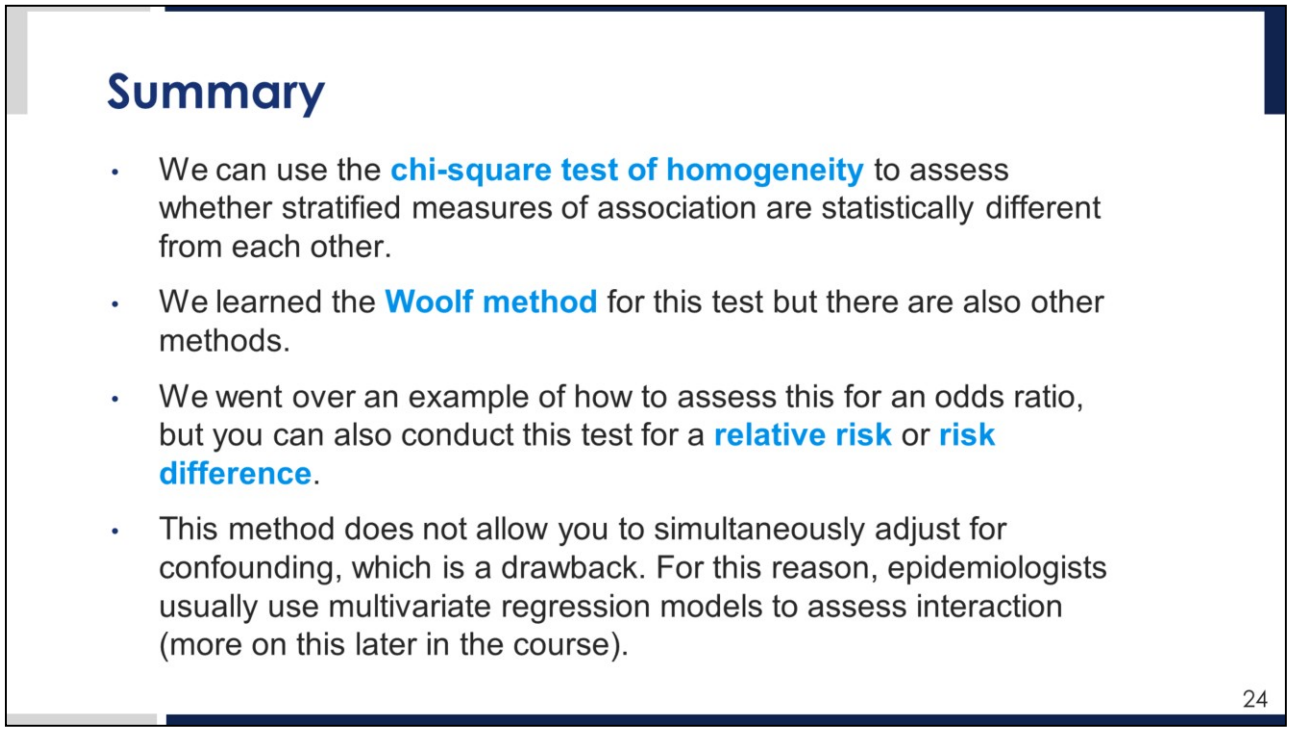
Measure of association (MA)	MA_i	$var(MA_i)$
(ln) Odds Ratio	$\ln\left(\frac{a_i \times d_i}{b_i \times c_i}\right)$	$\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$
(ln) Relative risk	$\ln\left(\frac{a_i/(a_i + b_i)}{c_i/(c_i + d_i)}\right)$	$\frac{b_i}{a_i(a_i + b_i)} + \frac{d_i}{c_i(c_i + d_i)}$
Risk difference	$\frac{a_i}{(a_i + b_i)} - \frac{c_i}{(c_i + d_i)}$	$\frac{a_i b_i}{(a_i + b_i)^3} + \frac{c_i d_i}{(c_i + d_i)^3}$

22

Here's a table of all the relevant formulas for what we've been doing for the chi-square test of homogeneity. In generic format, you can see how it's written at the top, where ma refers to the measure of association at each level, and the denominator is the variance of the measure of association at each level. And then how to get the measure of association at each level is given by the cells a, b, c, and d, each of the boxes for the odds ratio, the relative risk, and the risk difference.



There are some limitations. We can't fully evaluate interaction using this test. If the sample size is large, small heterogeneity of no meaningful value may be statistically significant. There may be measures of association that vary substantially across a third variable, but the test may not show statistical significance. In practice, most epidemiologists assessed statistical interaction using regression models, and we'll discuss these later in the course.



So to summarize, we can use the chi-square test of homogeneity to assess whether stratified measures of association are statistically different from each other. We learned the Wolff method for this test, but there are also other methods. We went over an example of how to assess this for an odds ratio, but you can also conduct this test for a relative risk or risk difference, and I showed you those formulae. This method does not allow you to simultaneously adjust for confounding, which is a drawback. For this reason, epidemiologists usually use multivariate regression models to assess interaction. We'll talk more about this later in the course.

CHAPTER 10

Interaction

Epidemiological studies over the past century have revealed many of the common primary risk factors for most major chronic diseases. A fundamental example was the elucidation of cigarette smoking as a risk factor for lung cancer. Nevertheless, for many diseases the mechanism, a biological understanding of how such risk factors lead to disease, remains elusive. Further, in the case of smoking, we are only too aware of individuals who do not develop lung cancer after a lifetime of heavy smoking and those who develop the disease with no apparent exposure to known risks. While these phenomena might be partially explained by the presence of as yet unidentified exposures, the possibility that there exist variables, perhaps genetic, that act synergistically or antagonistically with smoking is especially tantalizing. For example, is there a factor—some form of "immunity"—that protects certain smokers from cancer? Understanding and clarifying the roles of synergistic and antagonistic factors may be of particular value when developing a biological model for disease development or structuring targeted disease screening programs and interventions. For these reasons, questions surrounding synergism and antagonism are of considerable biological as well as epidemiological interest. At first glance, these issues seem to be accessible through the idea of statistical interaction, introduced in Chapter 8. However, in this chapter we discuss that while the ideas of synergism and antagonism appear self-evident, their relationship to statistical interaction is not as straightforward as we might have hoped.

The concept of interaction is focused on the idea that the level of one factor, say, C, influences the action of another factor, E, on the outcome, D. *Webster's Dictionary* defines synergism as the "cooperative action of discrete agencies such that the total effect is *greater* than the sum of the two effects taken independently" (emphasis added). Antagonism has a parallel definition, with "lesser" replacing "greater." Applying these definitions to determine whether two factors of interest, E and C, are synergistic or antagonistic in their association with D, we face two issues that the definition fails to specify:

1. How do we measure the *effect* of a factor or factors?
2. How do we *sum* the effects of individual factors?

The assessment of synergy depends critically on how we address these two questions. Two standard approaches, multiplicative and additive interaction, are often used, each based on different effect measures and different summing tactics.

Before these are described, it is helpful to introduce some additional notation to describe the risks associated at various combinations of the factors C and E. For simplicity, we restrict attention to the situation where both factors C and E are binary.

Now let the risks depend on the levels of C and E according to the following notation:

$$P_{11} = P(D|E\&C)\ P_{10} = P(D|E\&\bar{C})$$
$$P_{01} = P(D|\bar{E}\&C)\ P_{00} = P(D|\bar{E}\&\bar{C}).$$

Considering the group of individuals with neither E nor C as a baseline or reference group leads to the definitions of three Relative Risks by dividing the remaining population into three groups, depending on exposure to one or other factor or to both. Specifically, let

$$RR_{11} = \frac{P_{11}}{P_{00}}$$
$$RR_{10} = \frac{P_{10}}{P_{00}}$$
$$RR_{01} = \frac{P_{01}}{P_{00}}.$$

Thus, for example, RR_{10} measures the Relative Risk for those individuals only exposed to E (and not C), against those in the reference group who are exposed to neither factor. We can similarly define OR_{11}, OR_{10}, OR_{01}, and ER_{11}, ER_{10}, ER_{01}. Throughout this chapter, we assume that there is no confounding of the risks at various levels of E and C by other factors, so that all of the noted effect measures have a causal interpretation.

10.1 Multiplicative and additive interaction

In this section the notation above is used, along with our understanding of the measures of association from Chapter 4, to develop two fundamentally different approaches to both defining and combining effects, so we can apply the definitions of synergy and antagonism.

10.1.1 Multiplicative interaction

Relative Risk will first be used as the measure of the effect of a factor, so that RR_{10} describes the effect of E separate from C; RR_{01} describes the effect of C separate from E; and RR_{11} the total effect of both E and C together. With regard to "summing" the effects of individual factors, it is natural to consider multiplication of the separate Relative Risks (equivalent to summing the log Relative Risks). This leads to the definition that E and C *do not interact multiplicatively* if $RR_{11} = RR_{10} \times RR_{01}$.

What does this definition of interaction have to do with the definition in Chapter 8.1.3 in terms of the homogeneity of a measure of association of the E–D relationship across strata of C? To answer this, note that the above definition is equivalent to

$$\frac{P_{11}}{P_{00}} = \frac{P_{10}}{P_{00}} \times \frac{P_{01}}{P_{00}},$$

which, after multiplying both sides by $\frac{P_{00}}{P_{01}}$, yields

$$\frac{P_{11}}{P_{01}} = \frac{P_{10}}{P_{00}}.$$

Note that the left-hand side of this relationship is just the Relative Risk for the D–E relationship among individuals who all have characteristic C, that is, RR for D associated with E, *in the* C *stratum*. Similarly, the right-hand side is the Relative Risk for the D–E relationship among individuals who do not have characteristic C, that is, RR associated with E in the $\bar{C}$ stratum. The condition of no multiplicative interaction is thus the same as requiring that the Relative Risk associated with E remains constant over levels, or strata, of C.

Analogously the Odds Ratio can be used as the measure of effect, defining the absence of (multiplicative) interaction as $OR_{11} = OR_{10} \times OR_{01}$. Similar algebra shows that this, in turn, is equivalent to the Odds Ratio for D associated with E, being constant over levels of C. The two conditions are, of course, approximately the same in a rare disease setting and we will not further distinguish between these two formulations of multiplicative interaction. Note that for either the RR or OR the definition of multiplicative interaction is symmetric in the roles of C and E, in that C interacts with the relationship between E and D if and only if E interacts with the C–D association.

A classic example regarding the possible presence of interaction concerns the separate and combined effects of smoking and asbestos exposure on the incidence of lung cancer. In approximate terms, it is known that lung cancer incidence is ten times greater for smokers than nonsmokers in a population without exposure to asbestos. On the other hand, asbestos exposure raises the risk of lung cancer in nonsmokers about fivefold. Finally, for individuals who both smoke and have prior asbestos exposure, the risk is roughly 50 times greater than for those who have neither risk factor. Using the notation at the beginning of this chapter, we have $RR_{10} = 10$, $RR_{01} = 5$, and $RR_{11} = 50$, with E and C representing smoking and asbestos exposure, respectively. Since $RR_{11} = RR_{10} \times RR_{01}$, we conclude that there is no multiplicative interaction between smoking and asbestos exposure with regard to their effects on lung cancer incidence.

10.1.2 Additive interaction

There is, of course, nothing sacrosanct about using the Relative Risk as a measure of effect nor of the notion that separate factor Relative Risks should multiply as a benchmark to compare with the Relative Risk for both factors taken together. We could still use the Relative Risk but, rather than multiplying, add deviations of RR from the null value 1 in order to assess synergism. This leads to the definition that E *and* C *do not interact additively* if $(RR_{11} - 1) = (RR_{10} - 1) + (RR_{01} - 1)$. This is equivalent to saying that

$$(P_{11} - P_{00}) = (P_{10} - P_{00}) + (P_{01} - P_{00}),$$

that is,

$$ER_{11} = ER_{10} + ER_{01}. \tag{10.1}$$

So, an equivalent approach to additive interaction is the use of Excess Risk as a measure of effect, with summing the Excess Risks the appropriate way to combine individual effects as a reference for considering the joint effect of the factors.

Note that by adding $(P_{00} - P_{01})$ to both sides of Equation 10.1, we see that the condition for no additive interaction is equivalent to

$$P_{11} - P_{01} = P_{10} - P_{00},$$

that is, the Excess Risk associated with E in the C stratum is the same as the Excess Risk from E in the $\bar{C}$ stratum. Hence, the absence of additive interaction is the same as requiring that the Excess Risk arising from E is homogeneous across the C strata. Additive interaction is also symmetric in the roles of C and E.

With the example of Section 10.1.1 regarding smoking and asbestos exposure and their association with lung cancer incidence, we now see quite a different picture with regard to interaction. With $RR_{01} = 5$, $RR_{10} = 10$, and $RR_{11} = 50$, we have $RR_{11} - 1 = 49$, which is considerably larger than $(RR_{10} - 1) + (RR_{01} - 1) = 9 + 4 = 13$. We can conclude that there is substantial additive interaction, synergistically, with regard to the effects of smoking and asbestos exposure, despite the absence of multiplicative interaction.

For two separate risk factors, both with Relative Risks greater than one, it is possible to move from additive or multiplicative antagonism, to no additive interaction and multiplicative antagonism, to additive synergism and multiplicative antagonism, to additive synergism and no multiplicative interaction, and finally to both additive and multiplicative synergism, depending on the sizes of the Relative Risks for individuals exposed to both factors as compared with those exposed to only one. While this is often confusing at first, remember that the definitions of additive and multiplicative interaction refer only to patterns in the Relative Risks for the population subgroups experiencing various combinations of exposures and do not necessarily have any biological meaning. We discuss the implications of the difference between these two definitions of interaction in Sections 10.2 and 10.5.

10.2 Interaction and counterfactuals

We now turn to counterfactuals, introduced in Section 8.1.1, to see if causal ideas can shed light on the appropriate form of interaction that we should look for. We continue with the simplest possible case of two binary risk factors E and C, and we illustrate our discussion in terms of our example where the outcome is CHD and the two variables are behavior type (E) and a dichotomized version of body weight (C). The situation is now more complex than described in Table 8.1, since with two factors there are four possible experimental counterfactuals corresponding to the four possible combinations of the two binary factors. For each of these counterfactual exposure combinations there are two possible outcomes, D or $\bar{D}$, so that in all there are $2^4 = 16$ possible counterfactual outcome patterns. These are listed in Table 10.1,

Table 10.1 *Distribution of possible CHD responses (D) to behavior type (E) and binary measure of body weight (C)*

Group	Type A and High Weight ($E\&C$)	Type A and Low Weight ($E\&\bar{C}$)	Type B and High Weight ($\bar{E}\&C$)	Type B and Low Weight ($\bar{E}\&\bar{C}$)	Number
1	D	D	D	D	Np_1
2	D	D	D	$\bar{D}$	Np_2
3	D	D	$\bar{D}$	D	Np_3
4	D	D	$\bar{D}$	$\bar{D}$	Np_4
5	D	$\bar{D}$	D	D	Np_5
6	D	$\bar{D}$	D	$\bar{D}$	Np_6
7	D	$\bar{D}$	$\bar{D}$	D	Np_7
8	D	$\bar{D}$	$\bar{D}$	$\bar{D}$	Np_8
9	$\bar{D}$	D	D	D	Np_9
10	$\bar{D}$	D	D	$\bar{D}$	Np_{10}
11	$\bar{D}$	D	$\bar{D}$	D	Np_{11}
12	$\bar{D}$	D	$\bar{D}$	$\bar{D}$	Np_{12}
13	$\bar{D}$	$\bar{D}$	D	D	Np_{13}
14	$\bar{D}$	$\bar{D}$	D	$\bar{D}$	Np_{14}
15	$\bar{D}$	$\bar{D}$	$\bar{D}$	D	Np_{15}
16	$\bar{D}$	$\bar{D}$	$\bar{D}$	$\bar{D}$	Np_{16}

together with the population distribution of the 16 groups that share identical patterns. For example, Group 1 individuals develop CHD regardless of their combination of behavior type and body weight, and the population proportion of Group 1 individuals is p_1. Similarly, Group 4 individuals always get CHD when they are Type A but never if they are Type B, regardless of their weight.

Note that some of the counterfactual patterns for a given group exhibit synergistic or antagonistic characteristics. For example, Group 8 individuals only contract CHD when they are *both* Type A and of high weight and not if they have only one or neither risk. For such individuals, each factor is only a risk in the presence of the other. On the other hand, as we already noted, body weight is irrelevant in determining the onset of CHD for Group 4 individuals who do not exhibit synergism or antagonism between factors. Careful consideration of each type reveals that Groups 2, 3, 5, 7–10, 12, and 14–15 show synergistic or antagonistic properties Groups 1, 4, 6, 11, 13, and 16 do not.

We can compute, in terms of the counterfactual proportions, the proportion of Ds in the population if everyone were Type A and high weight, denoted by $P_{11(causal)}$:

$$P_{11(causal)} = p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 + p_8.$$

Similarly, the probability of D if everyone is Type A and low weight is

$$P_{10(causal)} = p_1 + p_2 + p_3 + p_4 + p_9 + p_{10} + p_{11} + p_{12};$$

if everyone is Type B and high weight,

$$P_{01(causal)} = p_1 + p_2 + p_5 + p_6 + p_9 + p_{10} + p_{13} + p_{14};$$

and, finally, if everyone is Type B and low weight,

$$P_{00(causal)} = p_1 + p_3 + p_5 + p_7 + p_9 + p_{11} + p_{13} + p_{15}.$$

In general, these causal risks can be contrasted using multiplicative or additive interaction, as we did with the observable risks, P_{11}, P_{10}, P_{01}, P_{00}, in Section 10.1, although this does not provide much insight. However, one valuable deduction can be gleaned directly. Suppose that there are *no* individuals of any of the groups that display synergism and antagonism as discussed above, that is, $p_2 = p_3 = p_5 = p_7 = p_8 = p_9 = p_{10} = p_{12} = p_{14} = p_{15} = 0$. Then, we can immediately see that

$$\begin{aligned} ER_{11(causal)} &= P_{11(causal)} - P_{00(causal)} = (p_1 + p_4 + p_6) - (p_1 + p_{11} + p_{13}) \\ &= (p_4 - p_{13}) + (p_6 - p_{11}) \\ &= (P_{10(causal)} - P_{00(causal)}) + (P_{01(causal)} - P_{00(causal)}) \\ &= ER_{10(causal)} + ER_{01(causal)}. \end{aligned} \tag{10.2}$$

The difference between Equations 10.1 and 10.2 is that the former refers to observed Excess Risks whereas the latter refers to causal Excess Risks. However, assuming the randomization assumption conditional on both E and C (equivalently, that there is no additional confounding after stratification by E and C), the observable and causal measures are the same. We can thus conclude in the absence of confounding by other factors that if we detect additive interaction in a population (Equation 10.1 does not hold) then additive interaction occurs amongst the causal risks (Equation 10.2 does not hold), and therefore at least some members of the population show synergistic or antagonistic behavior with regard to their four counterfactuals. Our back-of-the-envelope assessment of smoking and asbestos exposure as risk factors for lung cancer in the last section thus shows that some fraction of the population possesses antagonistic/synergistic counterfactual patterns, although how many and of which group we cannot say. Despite the weakness of this statement, it suggests that the presence or absence of additive interaction is the right benchmark to claim some form of biological interaction between the two risk factors. We return to this point again in Section 10.5. However, be careful not to equate the absence of additive interaction with a lack of biological synergism or antagonism; it is easy to see that Equation 10.2 can still hold even when some of the synergistic/antagonistic groups exist in the population, that is, some of the proportions p_2, p_3, p_5, p_7, p_8, p_9, p_{10}, p_{12}, p_{14}, p_{15} are nonzero.

10.3 Test of consistency of association across strata

As argued in Chapter 9.1, it is helpful to examine the consistency of effect measures across strata before launching into single summary procedures. Such assessment highlights population subgroups where we believe there to be different impacts of the exposure, E, on disease D. Since we usually deal with population samples, this is a good point to introduce sampling variation into our comparison of various effect measures, in part so that we are not distracted by apparent strata differences in the Relative Risk, say, that are easily explained by mere chance variation. That is, we describe how to use observed data stratified over levels of one or more extraneous

variables to assess whether we can plausibly assume that a measure of association associating E with D is consistent across strata. In other words, we assess evidence suggesting that the measure of association of choice may vary across strata, indicating that the stratifying variables modify the effect of E on D. Depending on the particular effect measure used, these techniques help to assess the presence of multiplicative or additive interaction.

As in Chapter 9, we assume that the data on exposure and outcome have been stratified into I strata based on common levels of a set of extraneous variables. We again use the notation of Table 9.1 for the data in the ith stratum. An example of such stratified data was given in Table 9.2 for the Western Collaborative Group Study. In Section 10.3.3, we reexamine this data on behavior type and CHD by body weight to consider whether it is plausible that the observed estimated measures of association, given in Table 9.2, all arise from the same stratum-specific value.

10.3.1 The Woolf method

For the moment, use the Odds Ratio as the measure of association of interest, so that consistency across strata refers to absence of multiplicative interaction. The null hypothesis under investigation is then H_0: $OR_1 = OR_2 = \cdots = OR_I$, where, as before, OR_i is the Odds Ratio for the E–D association in the ith stratum for $i = 1, \ldots, I$. The alternative hypothesis is that there is at least one stratum-specific Odds Ratio that differs from another one. Figure 10.1 illustrates a typical null hypothesis.

Note that the null hypothesis has one degree of freedom since it does not specify the particular constant value of the assumed common Odds Ratio. One approach to testing interaction is to pick a reasonable estimate of a common Odds Ratio, and then look at the variation of the individual stratum Odds Ratio estimates from this common value and determine whether it appears compatible with random variation. Because of skewness in the sampling distribution of the Odds Ratio estimates, we again prefer to work on the logarithm scale, using estimates of $\log OR_i$. We have

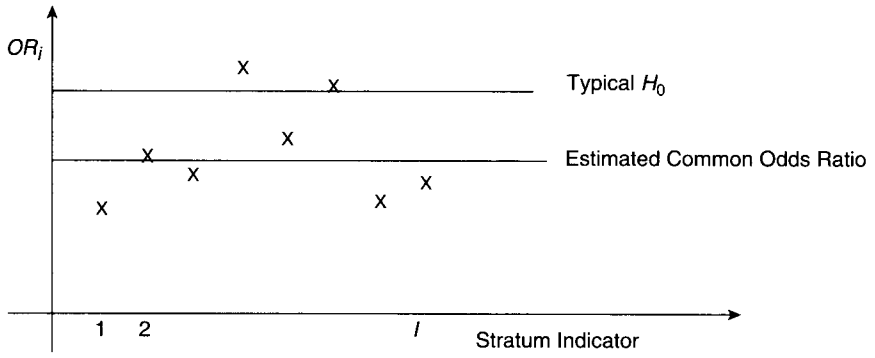


Figure 10.1 *Schematic describing null hypothesis in test for multiplicative interaction.*

already derived an estimate of a plausible "average" of the set of stratum-specific log Odds Ratios, namely, the Woolf estimator (Equation 9.2). We use this as our best guess of a common Odds Ratio. The deviation of the individual stratum-specific Odds Ratio from this "average" value can then be measured by $(\log \widehat{OR}_i - \log \widehat{OR}_W)^2$. These deviations can be added across the strata, each deviation weighted by the reciprocal of the variance of the specific log Odds Ratio estimate, so that we give more precise estimates of $\log OR_i$ a greater role in our assessment of variation. This yields the test statistic

$$\chi_H^2 = \sum_{i=1}^{I} w_i (\log \widehat{OR}_i - \log \widehat{OR}_W)^2,$$

where $\log \widehat{OR}_W$ and w_i are given by Equations 9.2 and 9.3, respectively.

Under the null hypothesis and with sufficiently large samples, χ_H^2 approximately follows a χ^2 sampling distribution with $I - 1$ degrees of freedom. The number of degrees of freedom arises from the fact that the stratum Odds Ratios in Figure 10.1 are all free to vary, in I degrees of freedom, whereas deviations from $\log \widehat{OR}_W$ can only vary with $I - 1$ degrees of freedom, as their weighted average must equal zero (since the weighted average of the $\log \widehat{OR}_i$s equals $\log \widehat{OR}_W$).

This test of the homogeneity of stratum Odds Ratios is easily adapted to cases where the Relative Risk or Excess Risk is being used to describe the E–D relationship. The generic form of the test statistic of homogeneity of a specific measure of association is given by

$$\chi_H^2 = \sum_{i=1}^{I} \frac{[\widehat{MA}_i - Avg(MA)]^2}{Var(\widehat{MA}_i)} = \sum_{i=1}^{I} w_i [\widehat{MA}_i - Avg(MA)]^2, \qquad (10.3)$$

where

$$Avg(MA) = \frac{\sum_{i=1}^{I} w_i \widehat{MA}_I}{\sum_{i=1}^{I} w_i}.$$

Table 10.2 provides formulas for calculating $\widehat{MA}_i$ and the associated weights w_i, for each of the Odds Ratio, Relative Risk, and Excess Risk. As with the Woolf method for estimating a common measure of association over strata, this technique

Table 10.2 *Components of χ^2 test for homogeneity for various measures of association*

MA	$\widehat{MA}_i$	$\widehat{Var}(\widehat{MA}_i) = w_i^{-1}$
(log) Odds Ratio	$\log\left(\frac{(a_i+\frac{1}{2})(d_i+\frac{1}{2})}{(b_i+\frac{1}{2})(c_i+\frac{1}{2})}\right)$	$\frac{1}{a_i+\frac{1}{2}}+\frac{1}{b_i+\frac{1}{2}}+\frac{1}{c_i+\frac{1}{2}}+\frac{1}{d_i+\frac{1}{2}}$
(log) Relative Risk	$\log\left(\frac{a_i/(a_i+b_i)}{c_i/(c_i+d_i)}\right)$	$\frac{b_i}{a_i(a_i+b_i)}+\frac{d_i}{c_i(c_i+d_i)}$
Excess Risk	$\frac{a_i}{a_i+b_i}-\frac{c_i}{c_i+d_i}$	$\frac{a_i b_i}{(a_i+b_i)^3}+\frac{c_i d_i}{(c_i+d_i)^3}$

for investigating interaction does not work well when there are many strata with few observations in each stratum. The computation and application of the Woolf test of homogeneity is illustrated for consistency in Section 10.3.3.

10.3.2 Alternative tests of homogeneity

The Woolf test for homogeneity, based on Equation 10.3, focuses directly on a specific measure of association. An alternative approach proposed by Breslow and Day (1980) instead examines variation in the cell counts themselves, assuming a fixed and common association across strata. To illustrate this method, we focus on the Odds Ratio and examine the "a_i" cells, using the notation of Table 9.1 (although we can equivalently use the b, c, or d cells). Based on the data of Table 9.2 from the Western Collaborative Group Study, the observed a_is (namely, 22, 21, 29, 47, and 59) vary, in part because of changes in the sample sizes and marginal totals from stratum to stratum, as well as because the Odds Ratio may also vary by body weight. So first we must evaluate what we expect the values of each of these a_is to be, *assuming that there is a common Odds Ratio* that describes the association between D and E in each stratum. But what is this expectation? For convenience, Table 10.3 revisits the 2 × 2 table from Table 6.1, with simplified notation.

Assuming that the Odds Ratio underlying this table is OR and that the marginal totals are fixed, then for large n, $A^* = E(a)$ is that value that, when substituted for a in Table 10.3, yields a 2 × 2 table with exactly these same marginals, n_E, $n_{\bar{E}}$, n_D, $n_{\bar{D}}$, and the assumed Odds Ratio OR. Table 10.4 gives the large sample expected values, B^*, C^*, D^* for the b, c, and d cells that, along with A^*, yield the observed marginals.

Table 10.3 *Simple notation for general* 2 × 2 *table*

		Disease		
		D	not D	
Exposure	E	a	$b = n_E - a$	n_E
	not E	$c = n_D - a$	$d = n_{\bar{E}} - n_D + a$	$n_{\bar{E}}$
		n_D	$n_{\bar{D}}$	n

Table 10.4 *Calculating the expectation of A in a general* 2 × 2 *table*

		Disease		
		D	not D	
Exposure	E	A^*	$B^* = n_E - A^*$	n_E
	not E	$C^* = n_D - A^*$	$D^* = n_{\bar{E}} - n_D + A^*$	$n_{\bar{E}}$
		n_D	$n_{\bar{D}}$	n

To calculate $A^* = E(a)$ we simply need to examine the equation

$$\frac{A^*D^*}{B^*C^*} = \frac{A^*(n_{\bar{E}} - n_D + A^*)}{(n_E - A^*)(n_D - A^*)} = OR, \tag{10.4}$$

which gives the required Odds Ratio. This is just a quadratic equation in A^* and easily solved. The values of B^*, C^*, and D^* are then computed by subtracting the value of A^* from the relevant marginals as in Table 10.4. This formulation of the expected value of a when n is large is based on the *noncentral hypergeometric distribution* and extends the (exact) result for $E(a)$ for the special case $OR = 1$ (that is, when E and D are independent), discussed in Chapter 6.4.1 and used in the Cochran–Mantel–Haenszel test.

This then gives the expected value of the a cell, assuming an Odds Ratio fixed at OR. But how big is the variation of a, given the sample size, marginals, and so on? The variance of a, assuming the Odds Ratio is OR, is given by

$$V^* = Var(a|OR) = \left[\frac{1}{A^*} + \frac{1}{B^*} + \frac{1}{C^*} + \frac{1}{D^*}\right]^{-1} \tag{10.5}$$

when n is large and with A^*, B^*, C^*, and D^* given by the calculations above.

We are now in the position to give an alternative test statistic for homogeneity of the Odds Ratio, based on the observed variation of the a_i terms across the strata. Specifically, the test statistic we are aiming for is

$$\chi^2_H(a) = \sum_{i=1}^{I} \frac{(a_i - A_i^*)^2}{V_i^*},$$

where A_i^* and V_i^* denote the expectation and variation of a_i for the ith stratum, assuming a common Odds Ratio, OR, for each stratum. But, of course, A_i^* and V_i^*, given by Equations 10.4 and 10.5, depend on the unknown OR. Instead we substitute an estimate of this assumed common Odds Ratio using, for instance, the Woolf or Mantel–Haenszel estimators described in Chapter 9.2.1 and 9.2.2, and with this estimate compute the A_i^* and V_i^* terms that appear in the formula for $\chi^2_H(a)$. Like χ^2_H, and for the same reasons, $\chi^2_H(a)$ approximately follows a χ^2 sampling distribution with $I - 1$ degrees of freedom, assuming the null hypothesis of a consistent Odds Ratio across strata. Again, this approach is not effective for several strata, all with small sample sizes. In such cases, an exact test (Zelen, 1971) can be used.

10.3.3 Example—the Western Collaborative Group Study: part 5

Return to the Western Collaborative Group Study and the data of Table 9.2, repeated in Table 10.5 for convenience. We first look at evidence of variation of the Odds Ratio for CHD associated with behavior type across the five body weight strata. Here we are checking for multiplicative interaction since the Odds Ratio is the effect measure. Table 9.5 provides the necessary components to calculate the appropriate test statistic; the necessary information has been abstracted in Table 10.6. In particular,

$$\begin{aligned}\chi^2_H &= 6.807(0.949 - 0.817)^2 + \cdots + 13.606(1.070 - 0.817)^2 \\ &= 3.981,\end{aligned}$$

using the fact that the Woolf estimate is given by $\log \widehat{OR}_W = 0.817$.

Table 10.5 *Coronary heart disease and behavior type, stratified by body weight*

Body Weight (lb)			CHD Event Yes	CHD Event No	$\widehat{OR}$
≤150	Behavior type	Type A	22	253	2.652
		Type B	10	305	
150$^+$–160	Behavior type	Type A	21	235	2.413
		Type B	10	270	
160$^+$–170	Behavior type	Type A	29	297	1.381
		Type B	21	297	
170$^+$–180	Behavior type	Type A	47	248	2.524
		Type B	19	253	
>180	Behavior type	Type A	59	378	2.966
		Type B	19	361	

Table 10.6 *Calculations for the Woolf test of homogeneity of the odds ratio, based on data from Table 10.5*

Body Weight (lb)	$\log(\widehat{OR}_i)$	w_i
≤150	0.949	6.807
150$^+$–160	0.855	6.680
160$^+$–170	0.316	11.477
170$^+$–180	0.910	12.453
>180	1.070	13.606

In this case, there are five body weight strata; so under the null hypothesis that the Odds Ratios associated with behavior type are constant across these strata, the test statistic χ^2_H should have a χ^2 sampling distribution with 4 degrees of freedom. With this as a benchmark, the observed test statistic χ^2_H yields a p-value of 0.41. Thus there is little convincing evidence that the Odds Ratios for CHD associated with behavior type are modified by body weight.

Analogous calculations can be carried out to examine homogeneity of both the Relative Risk and Excess Risk. Specifically, for the Relative Risk the computation of χ^2_H follows from the calculations of Table 9.7 and is given by

$$\begin{aligned}\chi^2_H &= 7.213(0.924 - 0.763)^2 + \cdots + 15.465(0.993 - 0.763)^2 \\ &= 3.948,\end{aligned}$$

also yielding a p-value of 0.41 in comparison with a $\chi^2_{(4)}$ distribution. For the Excess Risk, we have, using the results of Table 9.9,

$$\begin{aligned}\chi^2_H &= 2738(0.0483 - 0.0563)^2 + \cdots + 2549(0.0850 - 0.0563)^2 \\ &= 6.623.\end{aligned}$$

Table 10.7 *Breslow–Day test statistic calculations for the data in Table 10.5, assuming the common odds ratio* = $\widehat{OR}_W = 2.264$

Body Weight (lb)	a_i	A_i^*	B_i^*	C_i^*	D_i^*	V_i^*
≤150	22	20.937	254.063	11.063	303.937	6.879
150$^+$–160	21	20.587	235.413	10.413	269.587	6.555
160$^+$–170	29	34.300	291.701	15.700	302.300	10.042
170$^+$–180	47	45.650	249.350	20.350	251.650	12.653
>180	59	55.193	381.807	22.807	357.193	14.840

Here the associated p-value is 0.16, suggesting some evidence that the variation in Excess Risks across the body weight strata may not be compatible with a common Excess Risk for behavior type at each body weight. We raised this possibility earlier in Chapter 9.4.1 following a qualitative examination of the stratum Excess Risks.

Finally, we compute the Breslow–Day test statistic for homogeneity of the Odds Ratio, described in Section 10.3.2. We first need to compute the expected values, A_i^*, for the a_i cells in each stratum, as given by Equation 10.4. For example, suppose we use the Woolf estimate, $\widehat{OR} = 2.264$, of an assumed common Odds Ratio, calculated in Chapter 9.2.3; for the stratum with body weight ≤150 lb of Table 10.5, Equation 10.4 then becomes $A^*(283 + A^*)/(275 - A^*)(32 - A^*) = 2.264$. The solution of this quadratic equation that gives positive values for each of A^*, B^*, C^*, and D^* in Table 10.4 is $A^* = 20.937$, which then yields the values $B^* = 254.063$, $C^* = 11.063$, and $D^* = 303.937$. In turn, this gives $V^* = [20.937^{-1} + \cdots + 303.937^{-1}]^{-1} = 6.879$, from Equation 10.5. Solutions for A_i^*, B_i^*, C_i^*, D_i^* and the associated values for V_i^* for each stratum are given in Table 10.7.

Thus,

$$\chi^2_H(a) = \frac{(22 - 20.937)^2}{6.879} + \cdots + \frac{(59 - 55.193)^2}{14.840}$$
$$= 4.108,$$

yielding a p-value of 0.39 from a table of the $\chi^2_{(4)}$ distribution. Here, the Breslow–Day test gives almost equivalent results to the Woolf test for consistency of the Odds Ratio, calculated at the beginning of this section.

10.3.4 The power of the test for homogeneity

Suppose for a moment that we reassign the strata in Table 10.5 to a different hypothetical covariate, labeled X for convenience, as in Table 10.8. Note that we have kept the data exactly as it was in the body weight strata, just changed the labels of the strata. If we now carried out the χ^2 test for homogeneity of the Odds Ratio across these strata of X we would obtain exactly the same result, $\chi^2_H = 3.981$ and p-value = 0.41, as calculated in the previous subsection. This is because the test of homogeneity pays no attention to any underlying structure in the definition of the

Table 10.8 *Coronary heart disease and behavior type, stratified by hypothetical variable X*

			CHD event			
X			Yes	No	$\widehat{OR}$	$\widehat{RR}$
$X = 1$	Behavior type	Type A	29	297	1.381	1.347
		Type B	21	297		
$X = 2$	Behavior type	Type A	21	235	2.413	2.297
		Type B	10	270		
$X = 3$	Behavior type	Type A	47	248	2.524	2.281
		Type B	19	253		
$X = 4$	Behavior type	Type A	22	253	2.652	2.520
		Type B	10	305		
$X = 5$	Behavior type	Type A	59	378	2.966	2.700
		Type B	19	361		

strata. On the other hand, with the data of Table 10.8, we might be considerably more reluctant than for body weight to say that there is no evidence of effect modification by X, due to the obvious pattern of an increasing Odds Ratio for behavior type as X increases.

It is important to note that the χ^2 test for homogeneity looks for all possible departures from the null hypothesis. As discussed in our treatment of the Cochran–Mantel–Haenszel test, the consequence of extremely broad alternative hypotheses is a lack of power in the resulting test, since a fixed significance level is "used up" in protecting the procedure from falsely rejecting the null in comparison to such a wide variety of alternatives. There are two immediate ways around this difficulty. First, we can increase power by simply raising the nominal significance level from the arbitrary level, 0.05, to 0.2, say, below which point we consider there to be signs of nonconsistency. While this appears to be an artificial adjustment, it often produces the desired effect of drawing attention to important variations that might be missed by inappropriately fixating on the standard 0.05 level of significance.

Second, as in the Cochran–Mantel–Haenszel situation, we can develop a more targeted test statistic by restricting the alternatives of interest. This can be achieved, for example, by designing a test of homogeneity where the alternative hypothesis is that the measures of association increase (or decrease) corresponding to a natural order in the strata labels. With the data of Table 10.5, this alternative would focus on evidence that the Odds Ratio for CHD associated with behavior type increases or decreases as body weight increases. We will defer the development of such targeted tests for interaction until Chapter 14, where they arise naturally in the context of regression models. Note that the gain in such an approach is the increase in power in detecting a specific kind of departure from the null hypothesis of homogeneity (such as a trend in the measures of association); the disadvantage is that such a targeted test will have very little, if any, power against other kinds of alternatives.

10.4 Example of extreme interaction

We end this chapter by giving a practical example of extreme interaction, a case where the E–D association is in one direction at one level of C and in the other direction at the other level. While this was illustrated with hypothetical data in Table 8.6, it is valuable to look at some real data.

Infection with human cytomegalovirus (CMV) is a primary concern in renal transplants. The need for immunosuppression to prevent organ rejection leaves a transplant recipient vulnerable to reactivation of a prior asymptomatic infection or to a new infection from a CMV-positive but asymptomatic donor. (In the general U.S. population, CMV infection is extremely common but rarely causes symptoms; approximately 80% of Americans test positive for CMV by the age of 60.) On the other hand, prior exposure to infection in the recipient may lead to a partial immunity that will be of value in resisting CMV after a transplant. Thus, the nature of the risk associated with either a recipient testing CMV-positive or the organ donor testing CMV-positive, prior to transplant, is unclear. In unpublished work, Stafford (1988) pooled data from five studies and looked at the Relative Risk for a CMV infection after transplant associated with these two risk factors.

Table 10.9 gives the Relative Risks for posttransplant CMV disease, associated with receiving a kidney from a CMV-positive donor, first for the stratum where the recipient patient is already CMV-positive ($RR = 1.3$) and then for the stratum where the recipient is CMV-negative ($RR = 15.1$). A CMV-negative patient therefore experiences a huge increase in risk when receiving the organ from a CMV-positive donor as against a CMV-negative donor, presumably in large part due to the effects of immunosuppressant drugs. On the other hand, there is little elevation in risk when the patient is already CMV-positive prior to the transplant, perhaps due to some form of developed immunity. Similarly, the Relative Risk associated with being CMV-positive prior to transplant is very different depending on whether the donor is CMV-positive ($RR = 0.5$) or CMV-negative ($RR = 6.1$). The latter high Relative Risk is again perhaps the result of immune suppression with resulting reactivation of CMV. By either consideration of the roles of the two risk factors, the donor's and recipient's prior CMV status, the data present a striking interactive effect.

This discussion is not necessarily advocating stratification on a donor's CMV status in assessing the causal effect of a recipient's CMV status on posttransplant disease. In fact, a possible causal graph relating the three variables is illustrated in Figure 10.2.

Table 10.9 *Relative risks for CMV disease associated with organ donor and recipient's prior CMV status*

Risk Factor is CMV+ Donor		Risk Factor is CMV+ Recipient	
Stratum	*RR*	Stratum	*RR*
Recipient CMV+	1.3	Donor CMV+	0.5
Recipient CMV−	15.1	Donor CMV−	6.1

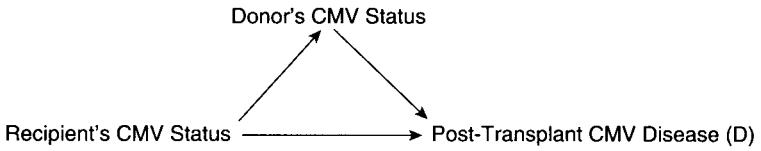


Figure 10.2 *A simple directed causal graph linking a donor and recipient's predisease CMV-status to posttransplant CMV disease in renal transplants.*

This figure assumes that the selected donor organ is influenced by the characteristics of the recipient, rather than the other way around. In this case, both pathways from the recipient's CMV status to posttransplant disease represent causal effects, so that stratification is unnecessary to remove confounding. On the other hand, the data of Table 10.9 show that a single measure of the effect of the recipient's CMV status conceals huge variation across the donor's status, with important practical implications.

10.5 Comments and further reading

We already know that in a given population it is possible for two risk factors to interact multiplicatively but not additively, and vice versa. Further, Siemiatycki and Thomas (1981) describe a range of multistage models for carcinogenesis, all leading to the same population risk patterns for exposure combinations, that have very different interpretations regarding the joint action of risk factors. Thompson (1991) makes a similar point. As a result, it is difficult to definitively conclude from epidemiological data that a specific form of biological synergism (or antagonism) exists. Without further understanding of a true biological model or mechanism describing in detail how factors act, together and alone, to produce the disease outcome, it is misleading to focus on a specified "interaction scale," additive, multiplicative, or any other alternative.

By using counterfactuals, we have seen that additive interaction points to synergistic/antagonistic behavior in at least some fraction of the population. A similar conclusion is reached by Rothman (1976), based on the conceptual idea of sets of "sufficient" causes or pathways for the disease outcome, with the subsequent interpretation that two factors do not biologically interact if they do not appear together in at least one of these pathways. Further, the additive scale is usually the scale of choice when either contemplating the contribution of risk factors to the total disease burden or designing effective interventions.

While counterfactuals provide the basis of causal inference when examining the total effects of one or more risk factors, we have seen that they only slightly illuminate subtleties of joint actions. This inability to definitively identify specific forms of joint actions of risk factors is an additional weakness of nonexperimental studies in determining causal effects.

On a positive note, we stress that statistical modeling of interactive effects is still a key component in exploring the contribution of two or more risk factors to explaining

disease outcomes. First, ignoring interaction on a particular scale may generate a misleading assessment of the size of an assumed common measure of association for one or both factors, even in the absence of confounding. Second, detection of interaction may give insight into specific conditions where one exposure is particularly strongly related to disease. Careful consideration of the various kinds of interaction may also lead to a simple or parsimonious description of the association of a series of factors with disease. Finally, highlighting the nature of observed interactions can suggest targeted screening programs and interventions.

All of the above points are particularly important in cases of extreme interaction, where the direction of the effect of one factor changes over levels of the other factor. For example, the data in Table 10.9 indicate that in renal transplants, a CMV-positive donor raises the risk of a posttransplant CMV infection only if the recipient is CMV-negative. This strongly suggests that CMV-positive donors, if used at all, be matched with CMV-positive recipients only. In another setting, Colditz et al. (1996) examine the impact of bearing children on the risk of breast cancer and how this association is modified by a family history of breast cancer in close female relatives. They report a decrease in the risk for breast cancer of about 20% in women who have ever given birth, as compared to nulliparous women, absent any family history. On the other hand, they observed an increase of risk of about 40% for the same comparison when there is family history. These findings suggest that breast cancer screening assessments should occur more frequently amongst parous women with a family history of breast cancer (McKnight, 1998).

10.6 Problems

Question 10.1

Determine if the following statements are true or false:

1. For C to interact multiplicatively with the association between E and D, it is required that C and E are not independent.
2. It is impossible to have data reflecting both additive and multiplicative interaction between two factors at the same time.

Question 10.2

Suppose a moderately sized case-control study of the relationship between a binary risk factor, E, and a disease, D, provides the following results:

- Pooled Odds Ratio estimate is 2.02 with associated 95% confidence interval of (1.25, 3.15).
- After stratification on another factor C, the Odds Ratio estimate when C is present is 1.10 with associated 95% confidence interval of (0.55, 2.70).
- After stratification on another factor C, the Odds Ratio estimate when C is absent is 1.01 with associated 95% confidence interval of (0.72, 1.65).

Discuss whether the following statements are likely to be true or false and give a brief reason:

1. The χ^2 test for independence of E and D in the pooled table yields a p-value < 0.05.
2. The χ^2 test for independence of E and D among those individuals with C yields a p-value < 0.05.
3. The χ^2 test for homogeneity of the two Odds Ratios in the C and $\bar{C}$ strata yields a p-value < 0.05.
4. The Cochran–Mantel–Haenszel χ^2 test for independence of E and D yields a p-value < 0.05.

Question 10.3

Refer to the data of Perez-Padilla et al. (2001) in Question 9.1, given in Table 9.13. Using both the Woolf and Breslow–Day test for interaction, examine the empirical evidence of multiplicative interaction of income and indoor air pollution. Given your results and interpretation, which Odds Ratio would you present for the data of Table 9.13: (1) the pooled Odds Ratio, (2) the Odds Ratio, adjusted for income, or (3) the stratum-specific Odds Ratios?

Question 10.4

In Perez-Padilla et al. (2001), the authors also stratified by smoking the data of Table 6.7 on association between indoor air pollution and tuberculosis. The stratified data are shown in Table 10.10, with smoking information coded as "never" and "past or present." The authors were particularly concerned that smoking might modify the effect of indoor air pollution on tuberculosis. In other words, they were looking for (multiplicative) interaction between smoking and indoor air pollution.

Based on the stratified data, what is the Odds Ratio, associated with biomass fuel exposure, for the never smokers stratum? What is the Odds Ratio for the past or present smokers stratum? Using the Mantel–Haenszel method, estimate the Odds Ratio associated with biomass fuel exposure after adjusting for smoking. Provide your assessment of the role of smoking as a confounding variable.

Table 10.10 *Case-control data on biomass fuel exposure and tuberculosis, stratified by smoking*

Smoking Status			Tuberculosis	
			Cases	Controls
Never smokers	Biomass fuel exposure	Yes	33	17
		No	186	411
Past or present smokers	Biomass fuel exposure	Yes	17	4
		No	52	113

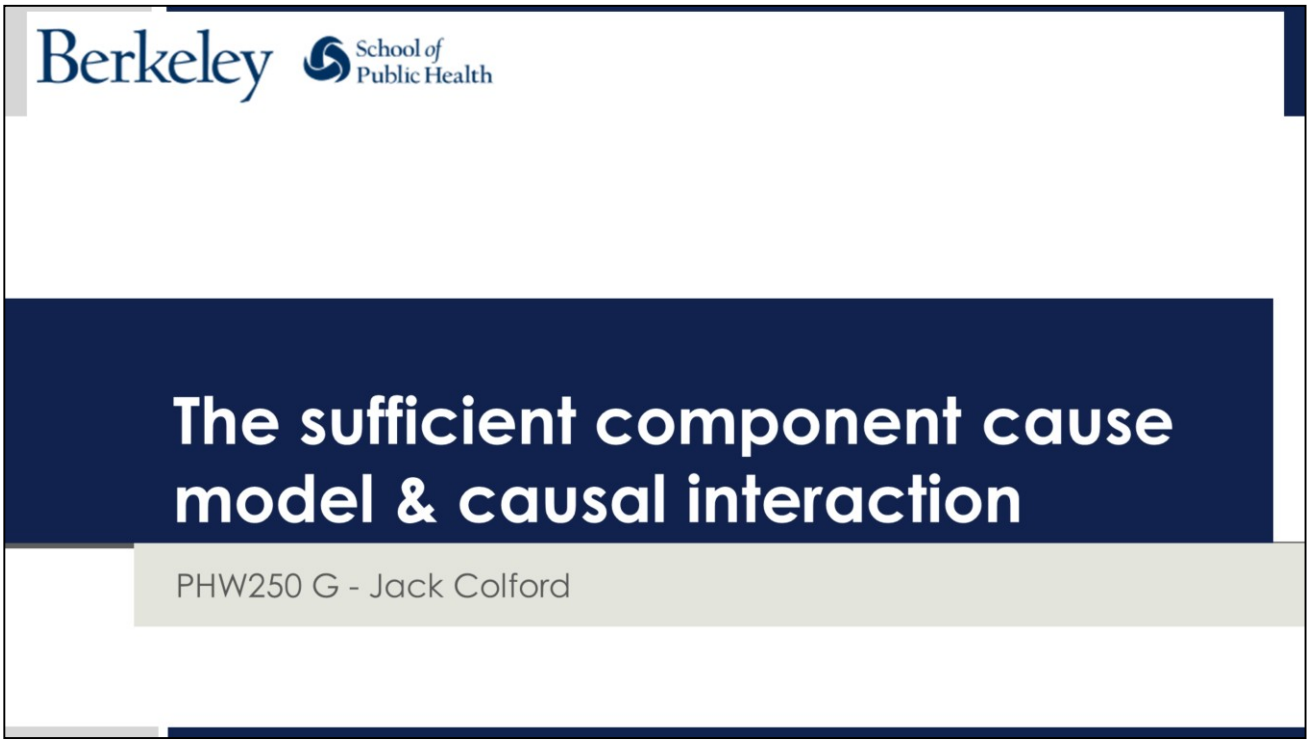
Using the Breslow–Day method, now examine the evidence for multiplicative interaction between smoking and indoor pollution. Given all your computations, which Odds Ratio would you report for the data of Table 10.10: (1) the pooled Odds Ratio, (2) the Odds Ratio, adjusting for smoking, or (3) the stratum-specific Odds Ratios?

Question 10.5

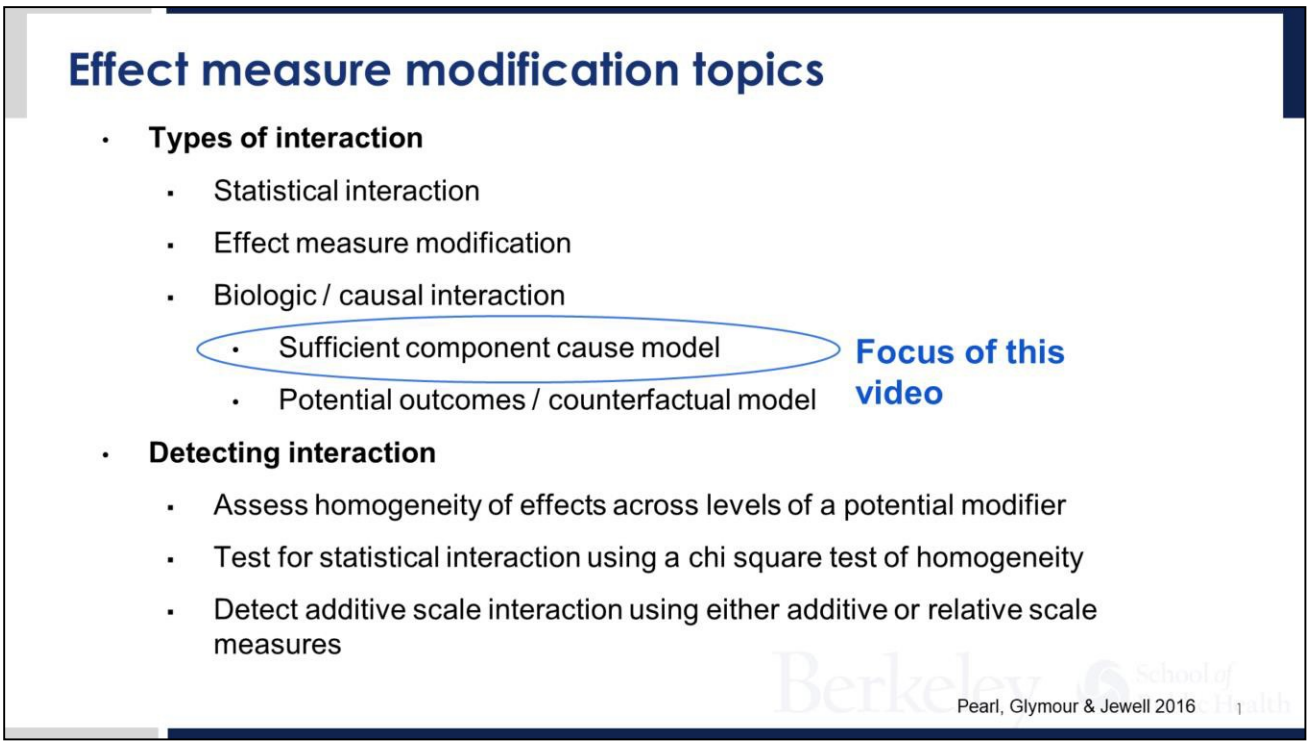
Refer again to the data set *oesoph*, found by going to http://www.crcpress.com/e_products/downloads/. Using both a binary measure of alcohol consumption (Question 7.3) and a binary measure of age (Question 9.2), examine the evidence for multiplicative interaction of age on the relationship between alcohol consumption and the incidence of esophageal cancer.

Question 10.6

Using the stratified data on the *Titanic* passengers of Question 9.4, examine the evidence for multiplicative interaction between sex and ticket class, using the Woolf method, with the Relative Risk as the basic measure of association.



JACK COLFORD: Our next topic is the sufficient component cause model and causal interaction.

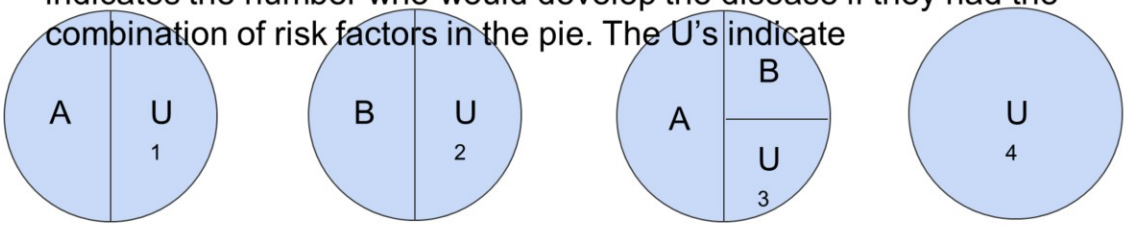


Just to review, we've previously discussed statistical interaction, effect measure modification, biologic or causal interaction as types of interaction. I've mentioned that there are a couple of different sub-topics to explore under biologic and causal interaction. And the focus of this video will be specifically on the sufficient component model of biologic causal interaction.

Just to review, we've talked about how we can assess homogeneity effects across levels of a potential modifier in order to detect interaction. We can also test for statistical interaction using a chi-square test of homogeneity. And we can detect additive scale interaction using either additive or relative scale measures.

Calculate the prevalence

Example: Population of 1,000 people. The number below each pie indicates the number who would develop the disease if they had the combination of risk factors in the pie. The U's indicate



100 People | 50 People | 50 People | 100 People

Estimate the prevalence if:

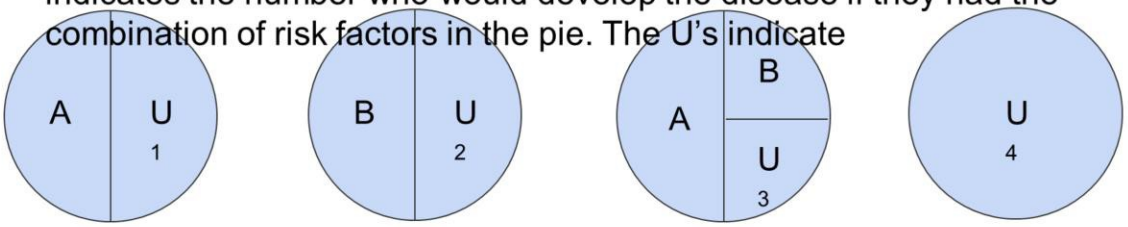
1) A and B were absent.
2) A was present and B was absent.
3) B was present and A was absent.
4) A and B were present.

2

Let's start with an example. In this example, we have a population of 1,000 people. The number below each pie indicates the number who would develop the disease if they had the combination of risk factors in the pie. The u's indicate the unknown causes. So estimate the prevalence of disease in the following situations-- number one, if A and B were absent, number two, A was present and B was absent, number three, B is present and A is absent, and number four, A and B are present.

Calculate the prevalence

Example: Population of 1,000 people. The number below each pie indicates the number who would develop the disease if they had the combination of risk factors in the pie. The U's indicate



100 People | 50 People | 50 People | 100 People

Estimate the prevalence if:

1) A and B were absent. **100/1000**
2) A was present and B was absent. **100/1000+100/1000 = 200/1000**
3) B was present and A was absent. **50/1000+100/1000 = 150/1000**
4) A and B were present. **100/1000+50/1000+50/1000+100/1000 = 300/1000**

3

Well, for the first example, when A and B are absent, we would have 100 people with disease out of the 1,000 population. In number two, when A is present and B is absent, we would have 100 or that when A is present and 100 for when B was absent. So that would be 100 over 1,000 plus 100 over 1,000 equals 200 over 1,000. In the third situation, when B is present and A is absent, we have 50 over 1,000 plus 100 over 1,000 for a total of 150 over 1,000. And finally if A and B are both present, we have 100 over 1,000 plus 50 over 1,000 plus 50 over 1,000 plus 100 over 1,000 for a total of 300 over 1,000 as our prevalence.

Assess effect modification: additive scale

Prevalence if:

1) A and B were absent. **100/1000**
2) A was present and B was absent. **100/1000+100/1000 = 200/1000**
3) B was present and A was absent. **50/1000+100/1000 = 150/1000**
4) A and B were present. **100/1000+50/1000+50/1000+100/1000 = 300/1000**

Using the prevalence estimates above, calculate prevalence differences for:

1) A without B
2) B without A
3) The expected combination of A & B
4) The observed combination of A & B

Let's assess effect modification on an additive scale. So this is just repeating in the top box the prevalence that we calculated in each of the different situations for A and B. Using the present prevalence estimates above, let's calculate the prevalence differences for A without B, B without A, the expected combination of A and B, and the observed combination of A and B.

Assess effect modification: additive scale

Prevalence if:

1) A and B were absent. **100/1000**
2) A was present and B was absent. **100/1000+100/1000 = 200/1000**
3) B was present and A was absent. **50/1000+100/1000 = 150/1000**
4) A and B were present. **100/1000+50/1000+50/1000+100/1000 = 300/1000**

Using the prevalence estimates above, calculate prevalence differences for:

1) A without B **= (200/1000) - (100/1000) = 100/1000**
2) B without A **= (150/1000) - (100/1000) = 50/1000**
3) The expected combination of A & B **= 100/1000 + 50/1000 = 150/1000**
4) The observed combination of A & B **= (300/1000) - (100/1000) = 200/1000**

5

Here are the calculations. For A without B present, we have 200 minus 100 over 1,000, which is 100 over 1,000. For B without A, we have 150 minus 100 over 1,000. So that's 50 over 1,000. The expected combination of A and B would be 100 over 1,000 plus 50 over 1,000 giving us 150 over 1,000. And the observed combination of A and B would be 300 over 1,000 minus 100 over 1,000 for a calculation of 200 over 1,000 prevalence.

Assess effect modification: additive scale

Prevalence if:

1) A and B were absent. **100/1000**
2) A was present and B was absent. **100/1000+100/1000 = 200/1000**
3) B was present and A was absent. **50/1000+100/1000 = 150/1000**
4) A and B were present. **100/1000+50/1000+50/1000+100/1000 = 300/1000**

Using the prevalence estimates above, calculate prevalence differences for:

1) A without B **= (200/1000) - (100/1000) = 100/1000**
2) B without A **= (150/1000) - (100/1000) = 50/1000**
3) The expected combination of A & B **= 100/1000 + 50/1000 = 150/1000**
4) The observed combination of A & B **= (300/1000) - (100/1000) = 200/1000**

Because the expected prevalence difference differs from the observed difference, we conclude effect modification was present on the additive scale.

6

So what we see is the observed and the expected prevalence is different from each other. So we conclude that effect modification was present on the additive scale.

Assess effect modification: relative scale

Prevalence if:

1) A and B were absent. **100/1000**
2) A was present and B was absent. **100/1000+100/1000 = 200/1000**
3) B was present and A was absent. **50/1000+100/1000 = 150/1000**
4) A and B were present. **100/1000+50/1000+50/1000+100/1000 = 300/1000**

Using the prevalence estimates above, calculate prevalence ratios for:

1) A without B
2) B without A
3) The expected combination of A & B
4) The observed combination of A & B

Now using the same prevalence estimates above, let's calculate prevalence ratios for the same situations.

Assess effect modification: relative scale

Prevalence if:

1) A and B were absent. **100/1000**
2) A was present and B was absent. **100/1000+100/1000 = 200/1000**
3) B was present and A was absent. **50/1000+100/1000 = 150/1000**
4) A and B were present. **100/1000+50/1000+50/1000+100/1000 = 300/1000**

Using the prevalence estimates above, calculate prevalence ratios for:

1) A without B **= (200/1000) / (100/1000) = 2**
2) B without A **= (150/1000) / (100/1000) = 1.5**
3) The expected combination of A & B **= 2 * 1.5 = 3**
4) The observed combination of A & B **= (300/1000) / (100/1000) = 300/1000**

8

Here are the calculations. For A without B, we have 200 over 1,000 divided by 100 for over 1,000 for a prevalence risk ratio of 2. For B without A, the estimate is 1.5. The expected combination of A and B would be 2 times 1.5 or 3. The observed combination of A and b is 300 over 1,000 divided by 100 or over 1,000. So that's 3.

Assess effect modification: relative scale

Prevalence if:

1) A and B were absent. **100/1000**
2) A was present and B was absent. **100/1000+100/1000 = 200/1000**
3) B was present and A was absent. **50/1000+100/1000 = 150/1000**
4) A and B were present. **100/1000+50/1000+50/1000+100/1000 = 300/1000**

Using the prevalence estimates above, calculate prevalence ratios for:

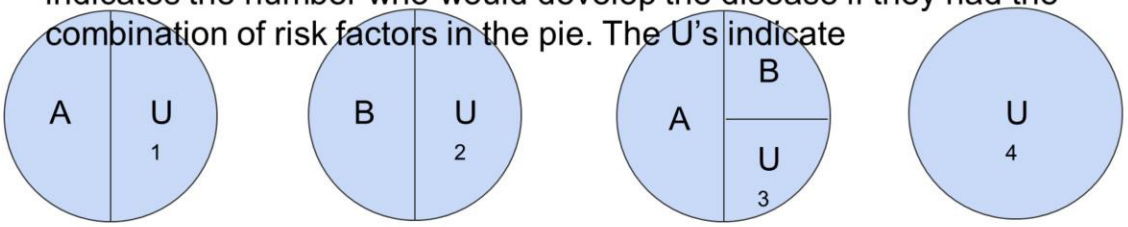
1) A without B **= (200/1000) / (100/1000) = 2**
2) B without A **= (150/1000) / (100/1000) = 1.5**
3) The expected combination of A & B **= 2 * 1.5 = 3**
4) The observed combination of A & B **= (300/1000) / (100/1000) = 300/1000**

Because the expected and observed prevalence ratios were the same, we conclude effect modification was absent on the relative scale.

Here the expected and observed prevalence ratios were the same. So we conclude effect modification was absent on the relative scale.

Calculate the prevalence

Example: Population of 1,000 people. The number below each pie indicates the number who would develop the disease if they had the combination of risk factors in the pie. The U's indicate



100 People | 50 People | 50 People | 100 People

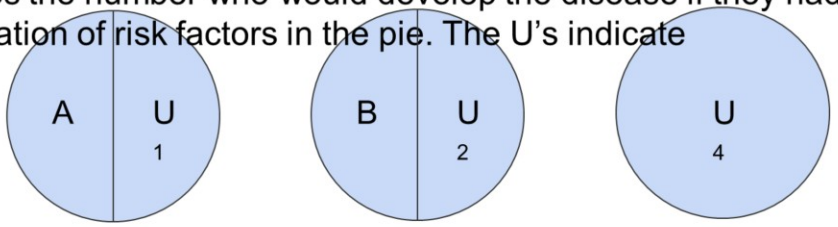
Which pie is responsible for effect modification? If we remove the third pie and repeat this exercise, what happens?

10

Here's our example again. Which pie is responsible for effect modification? If we remove the third pie and repeat the exercise, what happens?

Calculate the prevalence

Example: Population of 1,000 people. The number below each pie indicates the number who would develop the disease if they had the combination of risk factors in the pie. The U's indicate



100 People | 50 People | 100 People

Estimate the prevalence if:

1) A and B were absent. **100/1000**
2) A was present and B was absent. **100/1000+100/1000 = 200/1000**
3) B was present and A was absent. **50/1000+100/1000 = 150/1000**
4) A and B were present. **100/1000+50/1000+100/1000 = 250/1000**

11

Here are our calculations with the third pie removed.

And again, we have a total population of 1,000 people. So if A and B are absent, we would have 100 over 1,000 as a prevalence. If A was present and B was absent, we'd have 200 over 1,000. If B was present and A was absent, we have 150 over 1,000. And if A and B are present, we'd have 250 over 1,000.

Assess effect modification: additive scale

Prevalence if:

1)	A and B were absent.	100/1000
2)	A was present and B was absent.	100/1000+100/1000 = 200/1000
3)	B was present and A was absent.	50/1000+100/1000 = 150/1000
4)	A and B were present.	100/1000+50/1000+100/1000 = 250/1000

Using the prevalence estimates above, calculate prevalence differences for:

1) A without B
2) B without A
3) The expected combination of A & B
4) The observed combination of A & B

12

So using these prevalence estimates that we just calculate, let's calculate prevalence differences for each situation.

Assess effect modification: additive scale

Prevalence if:

1) A and B were absent. **100/1000**
2) A was present and B was absent. **100/1000+100/1000 = 200/1000**
3) B was present and A was absent. **50/1000+100/1000 = 150/1000**
4) A and B were present. **100/1000+50/1000+100/1000 = 250/1000**

Using the prevalence estimates above, calculate prevalence differences for:

1) A without B **= (200/1000) - (100/1000) = 100/1000**
2) B without A **= (150/1000) - (100/1000) = 50/1000**
3) The expected combination of A & B **= 100/1000 + 50/1000 = 150/1000**
4) The observed combination of A & B **= (250/1000) - (100/1000) = 150/1000**

Because the expected and observed prevalence differences were the same, we conclude effect modification was absent on the additive scale.

13

For A without B, we have 100 over 1,000. For B without A, we have 50 over 1,000. The expected combination of A and B would be 150 over 1,000. The observed combination of A and B would be 250 minus 1000 over 1,000 or 150 over 1,000. So here because the expected and observed prevalences were the same, we conclude that effect modification was absent on the additive scale in this example.

Assess effect modification: relative scale

Prevalence if:

1)	A and B were absent.	100/1000
2)	A was present and B was absent.	100/1000+100/1000 = 200/1000
3)	B was present and A was absent.	50/1000+100/1000 = 150/1000
4)	A and B were present.	100/1000+50/1000+100/1000 = 250/1000

Using the prevalence estimates above, calculate prevalence ratios for:

1) A without B
2) B without A
3) The expected combination of A & B
4) The observed combination of A & B

And finally, let's calculate the prevalence ratios for A without B, B without A, the expected combination of A and B, and the observed combination of A and B again.

Assess effect modification: relative scale

Prevalence if:

1) A and B were absent. **100/1000**
2) A was present and B was absent. **100/1000+100/1000 = 200/1000**
3) B was present and A was absent. **50/1000+100/1000 = 150/1000**
4) A and B were present. **100/1000+50/1000+100/1000 = 250/1000**

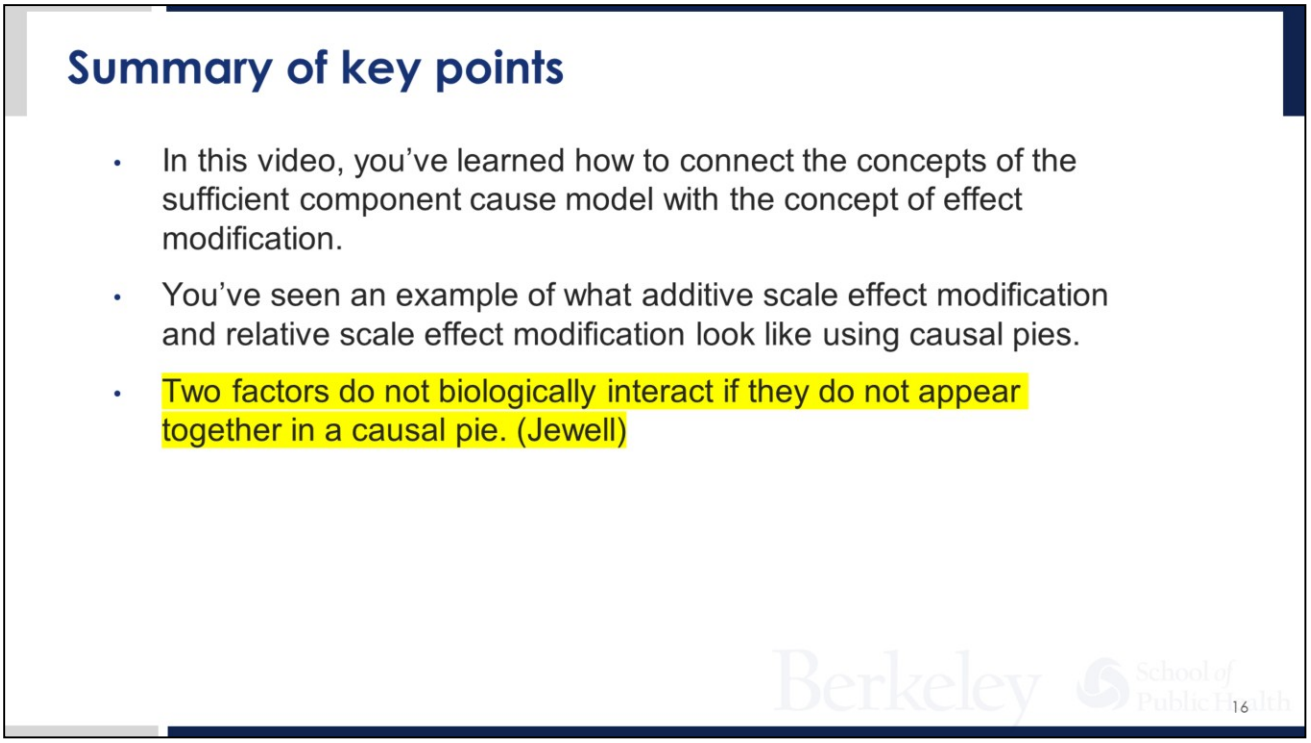
Using the prevalence estimates above, calculate prevalence ratios for:

1) A without B **= (200/1000) / (100/1000) = 2**
2) B without A **= (150/1000) / (100/1000) = 1.5**
3) The expected combination of A & B **= 2 * 1.5 = 3**
4) The observed combination of A & B **= (250/1000) / (100/1000) = 2.5**

Because the expected and observed prevalence ratios differ, we conclude effect modification was present on the relative scale.

15

Here are the calculations again. 2 is our estimate for A without B. 1.5 is our prevalence ratio estimate for B without A. The expected combination of A and B is 2 times 1.5 or 3. The observed combination of and B is 2.5. Because the expected and observed prevalence ratios are different here, we conclude that effect modification was present on the relative scale, because they are different from each other.



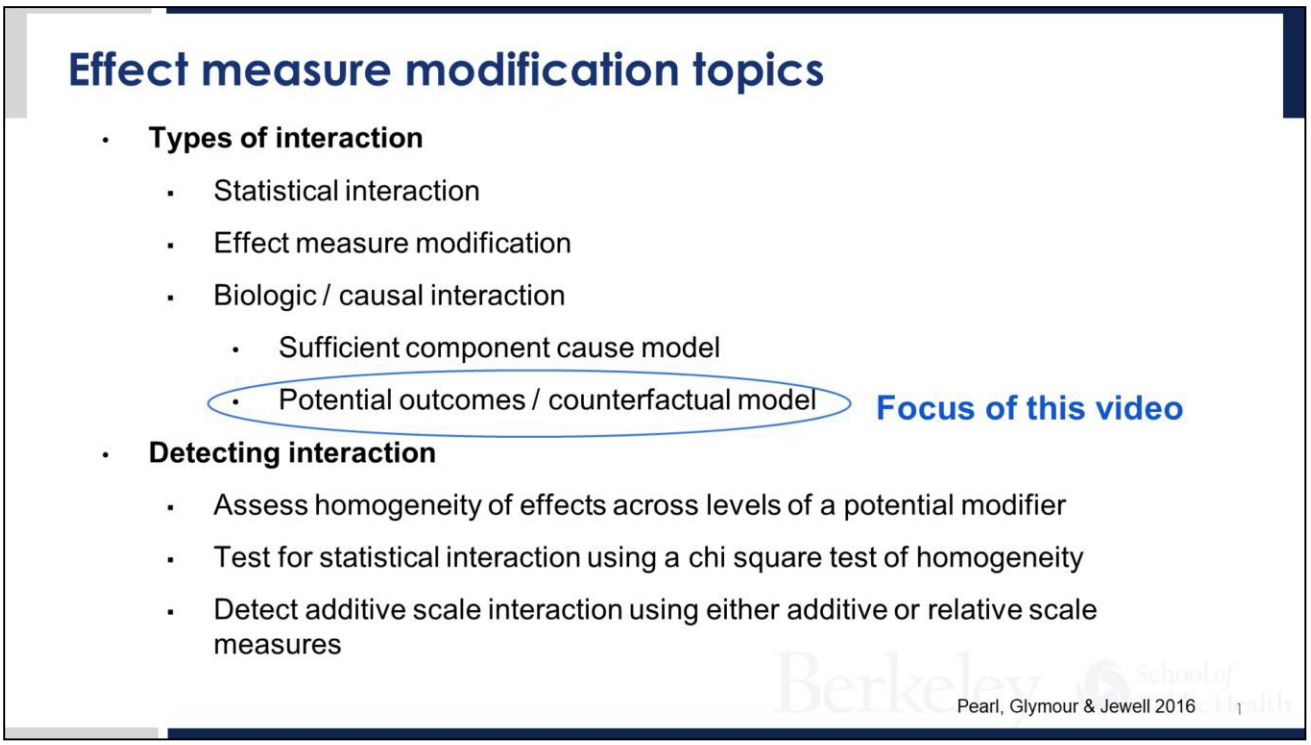
So in this video, you've seen some concrete examples of how to connect the concepts of the sufficient component cause model with the concept of effect modification. You've seen an example of what additive scale effect modification and relative scale effect modification look like when using causal pies. Two factors do not biologically interact if they do not appear together in a causal pie. And you can review that in the Jewell textbook.

Berkeley School of Public Health

The potential outcomes model & causal interaction

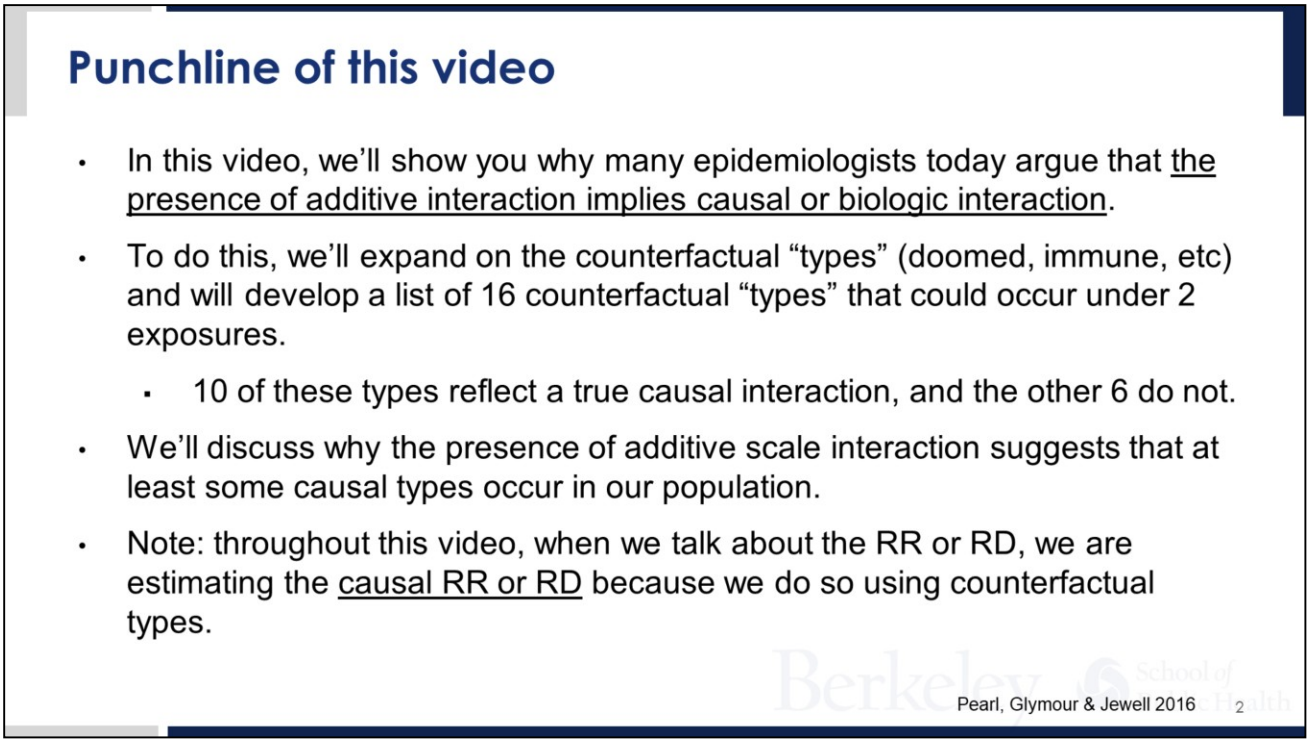
PHW250 B – Andrew Mertens

Andrew Mertens: In this video, I'll talk about the potential outcomes model and how it relates to the concept of causal interaction.



Here is the same slide you've seen before listing the different topics we've covered on effect measure modification. And so the focus of this video, again, is on potential outcomes in the counterfactual model, and how this model can be used to think about biologic or causal interaction.

So this is different from thinking about effect measure modification or statistical interaction. Really, what we're focusing on in this video is, is there truly, in a biological sense or even in terms of physics, out in the real world, is there an interaction between two variables, and how can we think about that within this counterfactual framework.



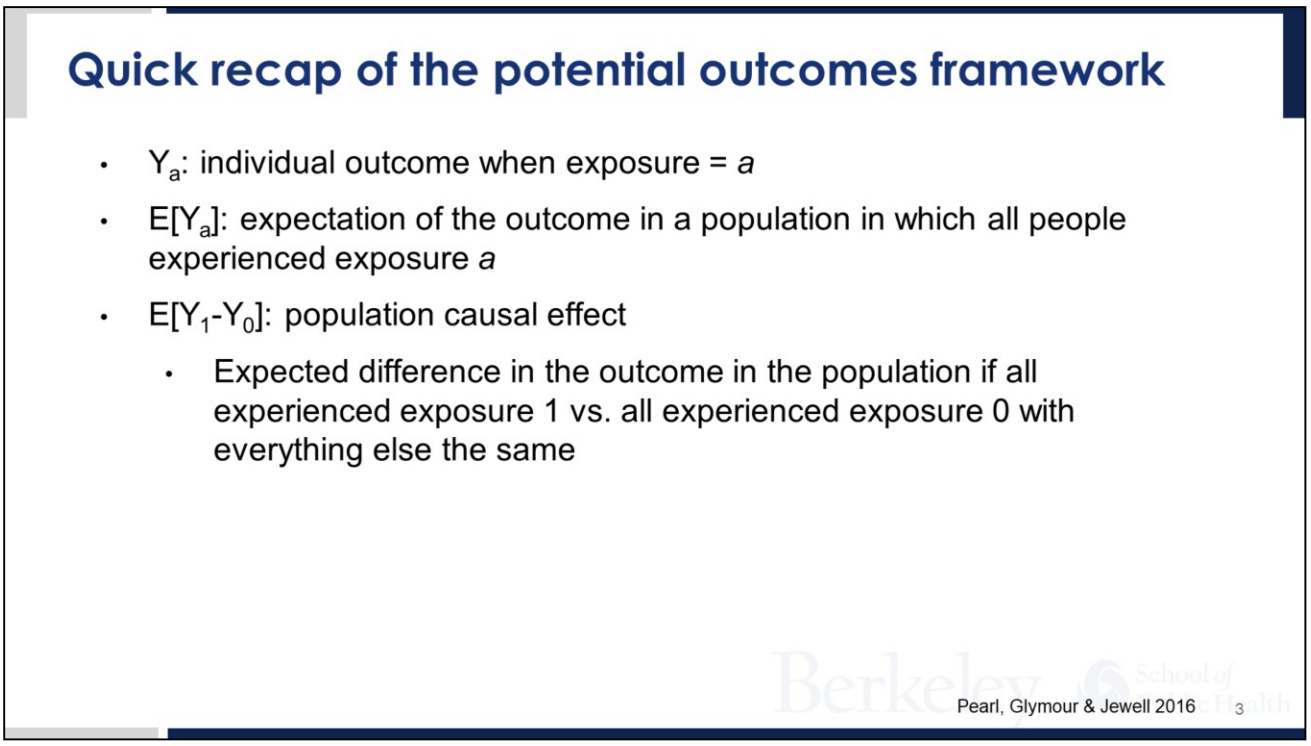
I'm going to walk through what might feel like a complicated process to make some points about biologic interaction within this model. So this slide is pretty wordy because I just want to make sure that it's crystal clear what the punchline of this video is. I'm going to show you why many epidemiologists today argue that the presence of additive interaction-- so that's interaction or effect measure modification on the additive scale, not the relative scale-- implies causal or biologic interaction. So that's the point that I'm going to reach by the end of the video. And in order to get to that point, to make that argument, I'm going to expand on the concept of counterfactual types.

We saw this earlier in the course when we thought about the people who will get the disease no matter what their exposure status, and those people would be doomed. The people who will never get the disease regardless of their exposure status, we called them the immune. And then there were the other two types, the people who would get the disease when they were exposed, and the people who would not get the disease when they were exposed. So those were the four counterfactual types we talked about previously.

And in this video, I'm going to develop the list with 16 counterfactual types that could occur when we have two different exposures. Now, out of these 16 types, 10 reflect a true causal interaction between our two exposures, and the other six don't. I'm going to work through some logic to discuss why the presence of additive scale interactions suggest that at least some of these 10 causal types are present in the population

when we see additive scale interaction.

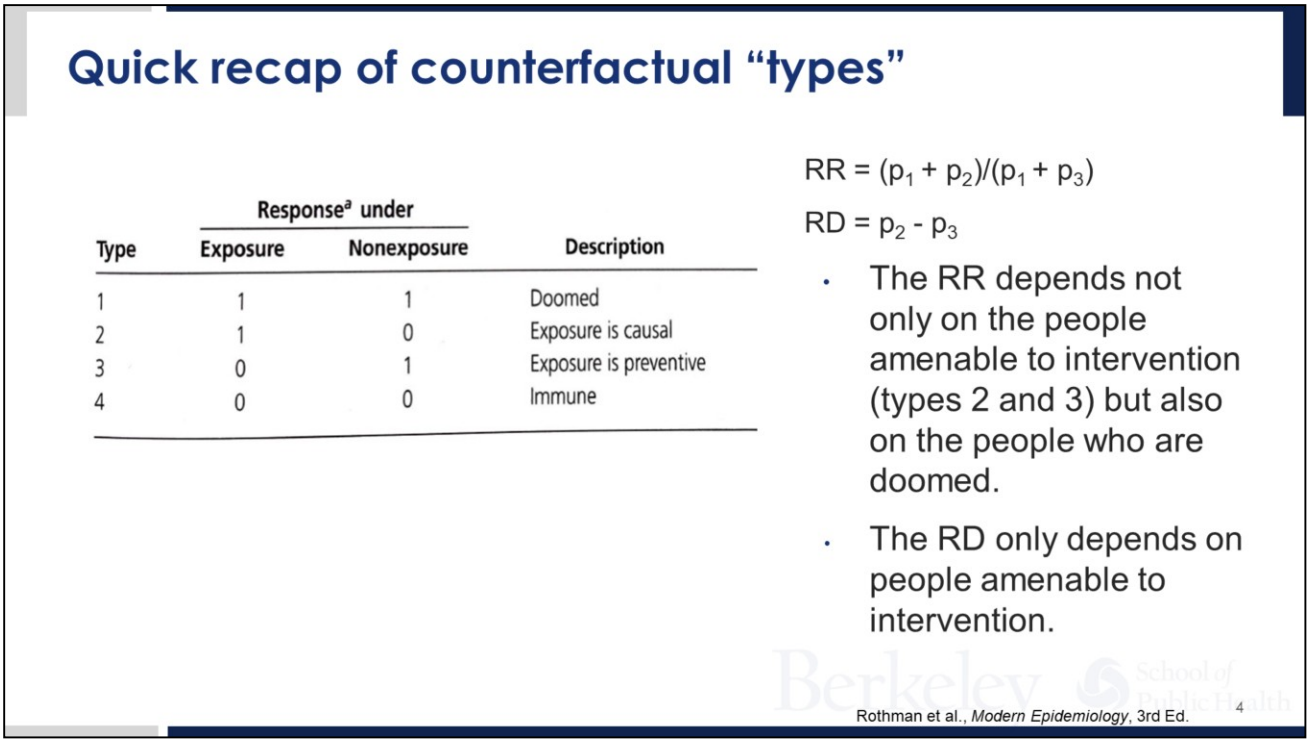
I also just want to make a note that, throughout the video, I'll be referring to the relative risk or the risk difference. And in this video, I'm focusing on a causal relative risk or risk difference. So this is the true, unbiased RR or RD. And that's because we are using this counterfactual model to come up with formulas for the RR or RD. So this is different from saying we're estimating the relative risk. This is really saying, what's the true relative risk or the true risk difference in the population.



Quickly, to recap the potential outcomes framework and its notation, we can use Y to denote the outcome. So if someone had, for example, lung cancer, Y could denote their lung cancer status. Y equals 1 could indicate that they do have lung cancer, and Y equals 0 could indicate that they don't. And then we can use Y sub a to indicate the individual outcome when an exposure is equal to a level A. So we could define a capital A as the random variable for our exposure, and a specific value of our exposure as lowercase a, as noted here. So Y sub a is the individual potential outcome for someone with an exposure equal to A.

E with brackets and then Y sub a is the expectation of that outcome in a population in which all people experience the exposure A. So this is a population-level measure, as opposed to the first bullet point, which is an individual-level measure. And then we can define the population causal effect. So here, we have E bracket Y1 minus Y0. So here, the 1 is equivalent to saying A equals 1, or the person had the exposure, or everyone in the population had the exposure, and Y sub 0 indicates that no one in the population had the exposure.

When we define this population and causal effect, in its essence, it's a counterfactual concept because it's not possible for everyone to experience the exposure and for everyone to not experience the exposure at the same time. So that's the notation we'd use for a causal risk difference. And if we wanted to define a causal relative risk or causal risk ratio, it would be the same, except we would divide instead of take the difference.



Now let's quickly recap the four counterfactual types we talked about earlier in the course. In this table, the rows indicate different types, so types one through four, and then the exposure and non-exposure, or unexposed, columns indicate the potential outcome when the person is exposed or not exposed. So for example, for type one in row one, this kind of person, if they're exposed, their value here is 1, which means they will have the outcome. And if they're not exposed, the value here is also 1, so they will also have the outcome, regardless of if they are exposed or not. And that's why we call them doomed.

In the second row, when they're exposed, they have the outcome. When they're unexposed, they don't have the outcome. We call that causal. The exposure is causal for this type of person. Type three is the opposite of type two, so we call that exposure-preventive, because they only get the disease when they're not exposed. And then type four, they never get the disease, regardless of their exposure status. And so we call this person immune.

And then we previously went through the process of showing how we could calculate the relative risk and the risk difference using these counterfactual types. And so P1 would be the probability of type one in the population, P2 would be the probability of type two in the population. and so on. And to calculate the relative risk, we then take P1 plus P2 divided by P1 plus P3. And that's because these are the two types of people where we have our exposure equal to 1 in the numerator. And then in the denominator, we divide by our two types of people who have none exposure equal to

1, meaning they're unexposed. So that's P1 and P3.

And then we can do the same thing to calculate the risk difference. And this simplifies to P2 minus P3. So I'm going over this very quickly because you can go back to our previous video to review this in detail if that's of interest, but the take-home message from this before was that the relative risk depends not only on people who we can affect through our exposure or through an intervention, which would be types two and three, but also on type one people, who are doomed. On the other hand, the risk difference only depends on people who are amenable to intervention, types two and three. And so the risk difference is isolating the people in our population who we actually can affect through causal actions, such as removing a harmful exposure.

So this set of steps and this logic has been used to make the case that absolute scale or additive scale measures of association are better for making causal statements than relative scale measures of association. Now, not everybody agrees with this, but this is the line of reasoning I'm going to go through in this video so that you're familiar with this not only in the context of measures of association, but also when thinking about effect modification.

OK, so here are our four causal types. And that's with one exposure.

Counterfactual types with two exposures

- Counterfactual "types" can be expanded to involve two exposures (X and Z)
- Each row represents the counterfactual outcomes that would be observed for a "type" of person given each potential combination of exposure to two factors, X and Z
- We imagine that our study population can be categorized into these types but we cannot know the proportion of each type because counterfactuals are unobserved
- Counterfactual outcomes – 1=disease, 0=no disease

	Outcome (Risk) Y when Exposure Combination Is			
Type	$X = 1$ $Z = 1$	$X = 0$ $Z = 1$	$X = 1$ $Z = 0$	$X = 0$ $Z = 0$
1	1	1	1	1
2*	1	1	1	0
3*	1	1	0	1
4	1	1	0	0
5*	1	0	1	1
6	1	0	1	0
7*	1	0	0	1
8*	1	0	0	0
9*	0	1	1	1
10*	0	1	1	0
11	0	1	0	1
12*	0	1	0	0
13	0	0	1	1
14*	0	0	1	0
15*	0	0	0	1
16	0	0	0	0

*Defined as interaction response type in present discussion (types

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

5

And now, the table on the right shows us all 16 causal types when we factor in a second exposure. So this table is in Rothman, and x equals 1 indicates that the first exposure is present, and z equals 1 indicates that the second exposure is present. And so we can see here that now we have four different columns indicating the different combinations that we might see of the two exposures. And then in each cell of this table, the 1 or the 0 indicate the potential outcome for that person under these different combinations of exposures.

So we're going to imagine that our study population can be categorized into these 16 different types, but we actually can't know the true proportion of each type in the population because these are counterfactual concepts, and so they're not observed. They're not something that we can really measure. But we're going to use that notation, again, of P1 through P16 to think theoretically about how these different types will play into our relative risk and risk ratio in an effect modification setting.

Counterfactual types with two exposures

- Types with **no causal interaction**
- At least one factor never has an effect, so there can be no interaction
- Examples:
  - Type 1: get the disease regardless of values of X and Z (doomed)
  - Type 6: disease is caused by exposure to X=1, Z has no effect
  - Type 4: only get the disease when Z=1

Type	Outcome (Risk) Y when Exposure Combination Is: $X=1, Z=1$	$X=0, Z=1$	$X=1, Z=0$	$X=0, Z=0$
1	1	1	1	1
2*	1	1	1	0
3*	1	1	0	1
4	1	1	0	0
5*	1	0	1	1
6	1	0	1	0
7*	1	0	0	1
8*	1	0	0	0
9*	0	1	1	1
10*	0	1	1	0
11	0	1	0	1
12*	0	1	0	0
13	0	0	1	1
14*	0	0	1	0
15*	0	0	0	1
16	0	0	0	0

*Defined as interaction response type in present discussion (types

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

Now, in the Type column, there is an asterisk to indicate whether that causal type suggests that a causal interaction is present. So type two, three, five, seven, eight, nine, 10, and 12, 14, and 15 are all causal types that indicate some interaction is present. And the types without the asterisk suggest there's no causal interaction.

When we're thinking about the types with no causal interaction, these are the types for which at least one factor never has an effect, which means there can't be any interaction. So if one exposure never has an effect, then it doesn't matter what the other exposure's value is. And this is a way of saying there is no interaction, no causal interaction. Let's go through some examples.

So type one in row one, this person gets the disease no matter what the values of x and z are. So they're doomed. That one's pretty straightforward. Because they're doomed regardless of their exposure status, there's no causal interaction present. Now let's look at type six.

So in the first cell for type six, we see that, if both exposures are present, they have the disease. If z is present but x is absent, they don't have the disease. And then if we flip those, if x is present and z is absent, they do have the disease. And then if both z and x are absent, they don't have the disease.

So this is a bit complicated to look at, but if we zoom in here, we can see that disease is caused by exposure to x. And this is true because there's a 1 in the first and the

third column. But z has no effect. And we can see this because, in the absence of x, there's never the potential outcome equal to 1. Right? So the person never gets the disease in the absence of x, so z has no effect. Because z has no effect, there's no causal interaction.

Now let's look at type four. Type four people only get the disease when z is equal to 1. So in the two right-hand columns, the third and fourth column, when z is equal to 0, they don't get the disease. So this is sort of the backwards of what we saw for type six. And that's why it's another example of no causal interaction.

Counterfactual types with two exposures

- Types **with causal interaction**
- Don't know what the effect of X will be without knowing the value of Z (and vice versa)
- Examples:
  - Type 8: get the disease only when X and Z are present
  - Type 5: X=0 with Z=1 blocks disease

Type	Outcome (Risk) Y when Exposure Combination Is: $X=1, Z=1$	$X=0, Z=1$	$X=1, Z=0$	$X=0, Z=0$
1	1	1	1	1
2*	1	1	1	0
3*	1	1	0	1
4	1	1	0	0
5*	1	0	1	1
6	1	0	1	0
7*	1	0	0	1
8*	1	0	0	0
9*	0	1	1	1
10*	0	1	1	0
11	0	1	0	1
12*	0	1	0	0
13	0	0	1	1
14*	0	0	1	0
15*	0	0	0	1
16	0	0	0	0

*Defined as interaction response type in present discussion (types

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

7

Now let's talk about the types with causal interaction. So one way of thinking about this is that causal interaction is present if we don't know what the effect of x would be without knowing the value of z, or vice versa. So that's just another way of saying that potential outcome depends on both the exposure status of x and the exposure status of z.

To go over an example, let's first look at type eight. Type eight people only get the disease with both x and z are present. So this is a pretty extreme example of causal interaction. If we see only x and no z or only z and no x, they don't get the disease, and if neither are present, they don't get the disease. But if both are present, they do. So their potential outcome depends on both exposures, not just on one.

Now let's look at type five people. So in type five people, they always get the disease, except when x is 0 and z is 1. And so that combination of x is 0 and z is 1 blocks disease. But there's no other pattern that we can discern where one exposure is operating on its own. In other words, the potential outcome depends on the status of both exposures for type five people.

What we are about to do:

Take home point: When EM is present on the additive scale, causal effect modification is implied (if you can make a set of other assumptions)

To show this, we will:

1. Get the RD for Z (RD_{01}) assuming no causal types
2. Get the RD for X (RD_{10}) assuming no causal types
3. Get the RD for X and Z (RD_{11}) assuming no causal types
4. Show that $RD_{11} = RD_{01} + RD_{10}$ when there are no causal types in the population (i.e. the RD is homogeneous)
5. By this logic, we can infer that if we observe $RD_{11} \neq RD_{01} + RD_{10}$, there are at least some causal types in our population (i.e. the RD is not homogeneous)

Pearl, Glymour & Jewell 2016 8

So again, I'm going to forecast for you what we're about to do. So these are the steps I'm about to go through. I'm then going to show you these steps one by one. And then I'll come back to remind us what we just did. So the take-home point is, when effect modification is present on the additive scale, causal effect modification is implied. If you can make a set of other assumptions, which we'll come back to. And so here are the steps we're going to take to make this point.

First, we're going to get the risk difference for z on its own in the absence of x, assuming there are no causal types in the population. Then we'll get the risk difference for x without z, assuming no causal types in the population. Then we'll get the risk difference for both exposures together, again assuming no causal types in the population. Then we'll show that the risk difference for both x and z is equal to the sum of the risk difference for x plus the risk difference for z on their own when there are no causal types in the population.

And we've done this a number of times now in different settings in this course. This is how we typically would go about showing that the risk difference is homogenous. In other words, that there's no effect modification when the observed risk difference, or any measure of association of two exposures is the same as the sum of the individual risk differences for each exposure on its own.

So based on this logic, after having gone through these steps and showing that the risk difference is homogeneous when there are no causal types in the population, we

will infer that, if we observe that the risk difference for both x and z is different from the sum of the risk difference for x and the risk difference for z, then there must be at least some causal types in our population. OK? So that's the logic.

It's a little bit complicated, but it's basically combining things we've done in two different settings. So what I'm about to do combines the way that we've shown you for evaluating effect modification in a non-causal context, and it combines that approach with this concept of people who are doomed, people who are immune, and all the other causal types, to put those together to think about how we can detect causal interaction.

Counterfactual types with two exposures

- So far we have discussed individual counterfactual risks
- Now, let's consider the average risk of Y in a study population
  - Causal risks of outcome for combinations of exposures X and Z
  - Notation: R_{XZ}
- Sum types that have value 1 down each column:
  - R_{11} = p1+p2+p3+p4+p5+p6+p7+p8
  - R_{01} = p1+p2+p3+p4+p9+p10+p11+p12
  - R_{10} = p1+p2+p5+p6+p9+p10+p13+p14
  - R_{00} = p1+p3+p5+p7+p9+p11+p13+p15

	Outcome (Risk) Y when Exposure Combination Is			
Type	$X = 1$ $Z = 1$	$X = 0$ $Z = 1$	$X = 1$ $Z = 0$	$X = 0$ $Z = 0$
1	1	1	1	1
2*	1	1	1	0
3*	1	1	0	1
4	1	1	0	0
5*	1	0	1	1
6	1	0	1	0
7*	1	0	0	1
8*	1	0	0	0
9*	0	1	1	1
10*	0	1	1	0
11	0	1	0	1
12*	0	1	0	0
13	0	0	1	1
14*	0	0	1	0
15*	0	0	0	1
16	0	0	0	0

*Defined as interaction response type in present discussion (types

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

So far, we've been talking about individual counterfactual risks. Now, in this slide, we're moving towards talking about the average risk of an outcome Y in a study population. And so we're going to use this notation, R sub xz. So R sub 1 1 is the risk when the exposure-- when x is 1 and z is 1. And then P1 is the probability of a causal type one in the population. So this is getting at this concept of average risk across the study population.

So how do we get R 1 1? Well, it's equal to the sum of P1, P2, P3, four, five, six, seven, and eight. And how do we know that? Well, if we look at the first column, where x equals 1 and z equals 1, those are the causal types with a potential outcome equal to 1.

Now for R sub 01, this is the risk when X is absent and Z is present. So that's column two. So we're going to go down column two and sum up the probability of all the causal types that have a potential outcome equal to one. And then we can do the same for R10. So that's the third column.

So in the third column, the types with potential outcome equal to one are p1, p2, p5, p6, 9, 10, 13, and 14. And then finally, R00 corresponds to the last column in the table. OK. So these are four different risks. And now what we're going to do is take these formulas that we have to find at the bottom of the slide and use them to calculate the risk difference and the relative risk.

Counterfactual types with two exposures

- **Calculate the RD for Z if no X assuming there are no interacting causal types**
  - R_{11} = p1+p4+p6
  - R_{01} = p1+p4+p11
  - R_{10} = p1+p6+p13
  - R_{00} = p1+p11+p13
- $RD_{01} = R_{01} - R_{00}$
- RD_{01} = (p1+p4+p11) – (p1+p11+p13)
- RD_{01} = p4-p13
- In words: This is the causal effect of Z if there are no preventive types

	Outcome (Risk) Y when Exposure Combination Is			
Type	$X = 1$ $Z = 1$	$X = 0$ $Z = 1$	$X = 1$ $Z = 0$	$X = 0$ $Z = 0$
1	1	1	1	1
2*	1	1	1	0
3*	1	1	0	1
4	1	1	0	0
5*	1	0	1	1
6	1	0	1	0
7*	1	0	0	1
8*	1	0	0	0
9*	0	1	1	1
10*	0	1	1	0
11	0	1	0	1
12*	0	1	0	0
13	0	0	1	1
14*	0	0	1	0
15*	0	0	0	1
16	0	0	0	0

*Defined as interaction response type in present discussion (types

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

First we're going to calculate the risk difference for Z if there's no X-- so if X is absent and Z is present, assuming that there are no interacting causal types in the population. So if there are no interacting causal types in the population, we're only interested in the rows in this table that have a red box around them. So this simplifies our formulas.

On the last slide, our formulas for risk were quite long. But now we only include the rows with no interacting causal types. For example, R11 is only p1 plus p4 plus p6. Those are the rows where there's a potential outcome equal to one and it is not an interacting causal type. We've basically reduced the number of causal types contributing to the sum for each risk.

Now taking these risks, we'll calculate the risk difference when X is absent and Z is present. So that's equal to the risk when X is absent and Z is present minus the risk when both are absent. So if we take R01 from up here-- that's p1 plus p4 plus p11- and then we subtract R00 from up here-- that's p1 plus p11 plus p13-- simplifying, we can cancel out p1, and we can cancel out p11. And that gives us p4 minus p13. So in words, what does this mean? This is the causal effect of Z if there's no preventive types in the population.

Counterfactual types with two exposures

- **Calculate the RD for X if no Z assuming there are no interacting causal types**
  - R_{11} = p1+p4+p6
  - R_{01} = p1+p4+p11
  - R_{10} = p1+p6+p13
  - R_{00} = p1+p11+p13
- $RD_{10} = R_{10} - R_{00}$
- RD_{10} = (p1+p6+p13) – (p1+p11+p13)
- RD_{10} = p6-p11
- In words: This is the causal effect of X if there are no preventive types

Type	Outcome (Risk) Y when Exposure Combination is: $X=1, Z=1$	$X=0, Z=1$	$X=1, Z=0$	$X=0, Z=0$
1	1	1	1	1
2*	1	1	1	0
3*	1	1	0	1
4	1	1	0	0
5*	1	0	1	1
6	1	0	1	0
7*	1	0	0	1
8*	1	0	0	0
9*	0	1	1	1
10*	0	1	1	0
11	0	1	0	1
12*	0	1	0	0
13	0	0	1	1
14*	0	0	1	0
15*	0	0	0	1
16	0	0	0	0

*Defined as interaction response type in present discussion (types

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

11

Now we're going to do the same thing for X. We'll calculate the risk difference for X if Z is absent, assuming there's no interacting causal types. So it's very similar to what we just did. But now it's RD sub 10. So we take R10 minus R00. So R10 is p1 plus p6 plus p13. And R00 is p1 plus p11 plus p13. And then we can cancel out p1 and p13. And that gives us p6 minus p11. So this is, in words, the causal effect of X if there's no preventive types in the population.

What do we mean when we say there are no preventive types in the population? Well, if we look at the two causal types for RD10, we have p6 and p11. And remember we're focused on X being present and Z being absent here. So if we look at p6, type 6, when X is present and Z is absent, the person has the disease. And if we look at p11 or type 11, when X is absent and Z is present, they have the disease.

So for X, we would call type 11 a preventive type for X, because not having the exposure X causes them to have the disease. If we just remove p11, we're left with p6. And this type of person is someone who has the disease if X is present. So that's why, in words, this means the risk difference is the causal effect of exposure to X if there's no preventive types. If there are preventative types, this is the net causal difference in risk between exposed and unexposed counterfactual populations.

Counterfactual types with two exposures

- **Calculate the observed RD for X and Z assuming there are no interacting causal types**
  - R_{11} = p1+p4+p6
  - R_{01} = p1+p4+p11
  - R_{10} = p1+p6+p13
  - R_{00} = p1+p11+p13
- $RD_{11} = R_{11} - R_{00}$
- RD_{11} = (p1+p4+p6) – (p1+p11+p13)
- RD_{11} = (p4+p6) – (p11+p13)

	Outcome (Risk) Y when Exposure Combination Is			
Type	$X = 1$ $Z = 1$	$X = 0$ $Z = 1$	$X = 1$ $Z = 0$	$X = 0$ $Z = 0$
1	1	1	1	1
2*	1	1	1	0
3*	1	1	0	1
4	1	1	0	0
5*	1	0	1	1
6	1	0	1	0
7*	1	0	0	1
8*	1	0	0	0
9*	0	1	1	1
10*	0	1	1	0
11	0	1	0	1
12*	0	1	0	0
13	0	0	1	1
14*	0	0	1	0
15*	0	0	0	1
16	0	0	0	0

*Defined as interaction response type in present discussion (types

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

12

Now we're going to do the same thing, but get the risk difference for both exposures together. So when X and Z are present, what's the causal risk difference, assuming there's no interacting causal types? This formula is equal to R11 minus R00. R11 is p1 plus p4 plus p6. R00 is p1 plus p11 plus p13. p1 cancels out. And that leaves us with p4 plus p6 minus p11 11 plus p13.

Implications for assessment of additive scale interaction

RD_{01} = p4-p13

RD_{10} = p6-p11

RD_{11} = (p4+p6) – (p11+p13)

- **What do we conclude if $RD_{01} + RD_{10} = RD_{11}$?**
  - Either there are no interacting types, or they are canceling each other out (example – a type 3 and a type 14 would cancel)
- **What do we conclude if $RD_{01} + RD_{10} \neq RD_{11}$?**
  - There are at least some interacting types (if we can assume that the RD estimates are not confounded).

Pearl, Glymour & Jewell 2016 13

Now let's put our three formulas together. So we have the risk difference when X is absent and Z is present, the risk difference when X is present and Z is absent, and the risk difference for both of them being present. And what we can see quite quickly is that if we add up RD01 and RD10, it's equal to RD11 once we rearrange everything. We take p4, the first term in the first RD, plus p6, the first term in the second RD. And then we subtract the second term from both RDs. And that gives us RD11. So in other words, RD01 plus RD10 equals RD11.

And what do we conclude if this is true? So remember, we went through all of these steps using only the causal types that were not interacting types. And so given that we did that and we're seeing homogeneity of the risk difference-- so this is the measure of association on the additive scale-- this suggests that either there really are no interacting types, or they're canceling each other out in some way.

Now what do we mean when we say they're canceling each other out? Well, if we have two types that have essentially opposite patterns of potential outcomes, they can cancel each other. So for example, type 3 and type 14 could cancel each other. Let's go back a few slides and look at an example of this.

Counterfactual types with two exposures

- **Calculate the observed RD for X and Z assuming there are no interacting causal types**
  - R_{11} = p1+p4+p6
  - R_{01} = p1+p4+p11
  - R_{10} = p1+p6+p13
  - R_{00} = p1+p11+p13
- $RD_{11} = R_{11} - R_{00}$
- RD_{11} = (p1+p4+p6) – (p1+p11+p13)
- RD_{11} = (p4+p6) – (p11+p13)

	Outcome (Risk) Y when Exposure Combination Is			
Type	$X = 1$ $Z = 1$	$X = 0$ $Z = 1$	$X = 1$ $Z = 0$	$X = 0$ $Z = 0$
1	1	1	1	1
2*	1	1	1	0
3*	1	1	0	1
4	1	1	0	0
5*	1	0	1	1
6	1	0	1	0
7*	1	0	0	1
8*	1	0	0	0
9*	0	1	1	1
10*	0	1	1	0
11	0	1	0	1
12*	0	1	0	0
13	0	0	1	1
14*	0	0	1	0
15*	0	0	0	1
16	0	0	0	0

*Defined as interaction response type in present discussion (types

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

14

So type 3, we have a pattern of 1, 1, 0, 1 for our potential outcomes. And then for type 14, we have a pattern of 0, 0, 1, 0. So these are opposite patterns in how the potential outcome responds to an exposure combination. And they are causal types. But essentially, if both of these types have the same prevalence in the population, they'll wash each other out. We won't see an effect. That's an interaction effect due to these causal types.

Implications for assessment of additive scale interaction

RD_{01} = p4-p13

RD_{10} = p6-p11

RD_{11} = (p4+p6) – (p11+p13)

- **What do we conclude if $RD_{01} + RD_{10} = RD_{11}$?**
  - Either there are no interacting types, or they are canceling each other out (example – a type 3 and a type 14 would cancel)
- **What do we conclude if $RD_{01} + RD_{10} \neq RD_{11}$?**
  - There are at least some interacting types (if we can assume that the RD estimates are not confounded).

Pearl, Glymour & Jewell 2016 15

Now by this logic, what do we conclude if we find that the risk difference for each exposure on its own is not equal to the risk difference for both exposures together? Well, by this logic, we could say that there's at least some interacting types in the population. And that's, of course, under the assumption that our risk difference estimates are not biased or confounded.

So this is really the punch line if we have a lack of homogeneity on the additive scale. By going through this logic, we've basically tried to argue that there must be some interacting types in the population, which means that there must be some causal interaction in the population.

What we just showed

Take home point: When EM is present on the additive scale, causal effect modification is implied (if you can make a set of other assumptions)

To show this, we will:

1. Get the RD for Z (RD_{01}) assuming no causal types
2. Get the RD for X (RD_{10}) assuming no causal types
3. Get the RD for X and Z (RD_{11}) assuming no causal types
4. Show that $RD_{11} = RD_{01} + RD_{10}$ when there are no causal types in the population (i.e. the RD is homogeneous)
5. By this logic, we can infer that if we observe $RD_{11} \neq RD_{01} + RD_{10}$, there are at least some causal types in our population (i.e. the RD is not homogeneous)

Pearl, Glymour & Jewell 2016 16

To walk through again what we just showed, our take-home point is that when effect modification is present on the additive scale, causal effect modification is implied. And to show this, we went through the process of getting the risk difference for each exposure on its own, assuming there were no causal types, the risk difference for each exposure together, assuming there were no causal types.

And then we showed that the risk difference was homogeneous when there were no causal types. And by that logic, we inferred that when the risk difference is not homogeneous, there are at least some causal types in our population. In other words, there is some causal interaction at play.

Counterfactual types with two exposures

- **Calculate the observed RR for X and Z assuming there are no interacting causal types**
  - R_{11} = p1+p4+p6
  - R_{01} = p1+p4+p11
  - R_{10} = p1+p6+p13
  - R_{00} = p1+p11+p13
- RR_{01} = (p1+p4+p11) / (p1+p11+p13)
- RR_{10} = (p1+p6+p13) / (p1+p11+p13)
- RR_{11} = (p1+p4+p6) / (p1+p11+p13)
- $RR_{01} \times RR_{10} \neq RR_{11}$

Type	Outcome (Risk) Y when Exposure Combination Is: $X=1, Z=1$	$X=0, Z=1$	$X=1, Z=0$	$X=0, Z=0$
1	1	1	1	1
2*	1	1	1	0
3*	1	1	0	1
4	1	1	0	0
5*	1	0	1	1
6	1	0	1	0
7*	1	0	0	1
8*	1	0	0	0
9*	0	1	1	1
10*	0	1	1	0
11	0	1	0	1
12*	0	1	0	0
13	0	0	1	1
14*	0	0	1	0
15*	0	0	0	1
16	0	0	0	0

*Defined as interaction response type in present discussion (types

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

17

Now let's take a quick look at how this works for the relative risk. I'm not going to go through this quite as slowly. Here we have the same risks defined for X and Z at the top. And let's quickly get the relative risk, where Z in the absence of X. So here, that's p1 plus p4 plus p11 over p1 plus p11 plus p13. And then we can fill in the remainder of these equations using these risk definitions from up at the top. And so that's defined here in these three bullet points.

Now let's look at whether the observed and expected relative risk for both exposures together are equal to each other. And I'm not going to go through every single step on this slide. But basically, if you get out a pen and paper and try to do this for yourself-- so calculate RR01 times RR10-- you'll find that it's not equal to the observed RR11. So again, right here, the product of these two relative risks is the expected relative risk for both exposures together. And it's not equal to the observed relative risk when we have no interacting causal types.

Implications for assessment of relative scale interaction

RR_{01} = p4/p13

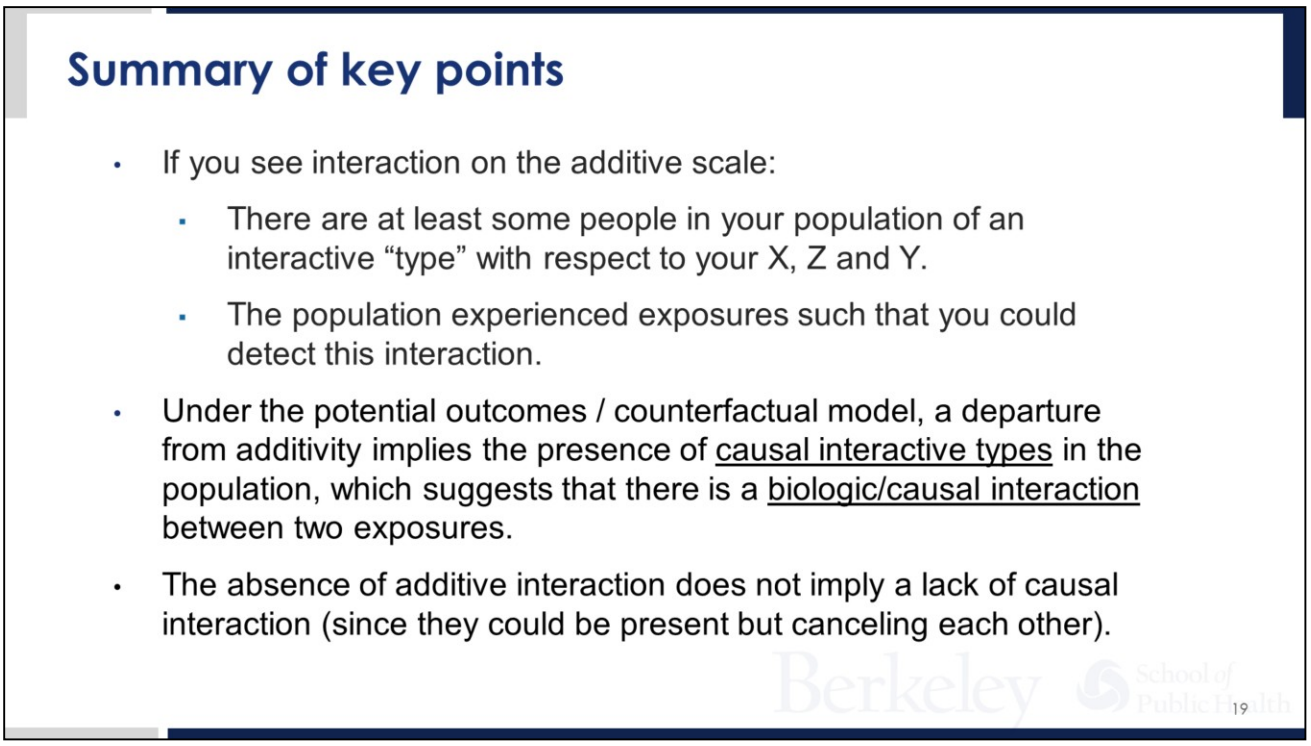
RR_{10} = p6/p11

RR_{11} = (p4/p6) x (p11/p13)

- $RR_{01} \times RR_{10} \neq RR_{11}$ when there are no interacting types.
- Cannot make statements about interaction as a causal concept based on multiplicative interaction
- Can make statements about effect modification only (measure of association modification)

Pearl, Glymour & Jewell 2016 18

So because this is true, we can't make statements about interaction as a causal concept based on multiplicative scale or relative scale of interaction. We can make statements about effect modification or effect measure modification, modification of our measure of association, which may be subject to confounding or bias, which is different from talking about causal interaction.



To summarize, if you see an interaction on the additive scale, then there are least some people in your population with an interactive type with respect to your particular exposures and outcomes of interest. And the population had to experience the exposures in a way that you could actually detect this interaction.

Under the potential outcome and counterfactual model, a departure from additivity-- meaning a lack of homogeneity-- on the additive scale implies the presence of causal interactive types in the population, suggesting that there is a biologic or causal interaction between your two exposures. Now if we don't see additive interaction in our data, that doesn't necessarily imply that there is no causal interaction, because there could be causal types that are canceling each other out.

So I've gone through these set of steps to walk through for you why people make the argument that it's better to assess interaction on the additive scale than on the relative scale when you're interested in causal interaction. And it's important to just briefly point out that not all epidemiologists fully agree with this. It's somewhat of a theoretical argument.

And so there's definitely different camps. There are people who want to see in papers that you report both the additive and relative scale of measures of interaction. And then there are others who stick with just relative scale interaction. And that's because there's a long tradition of using relative scale models in epidemiology.

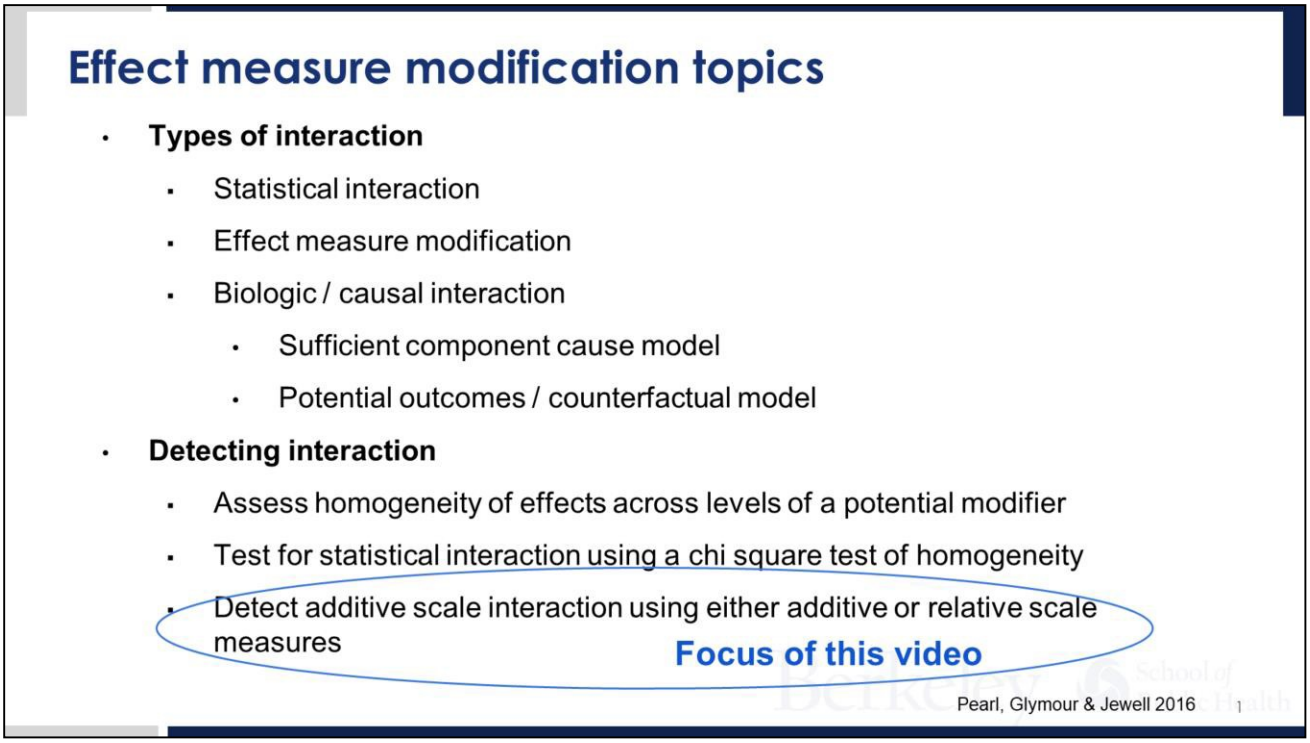
So that would be logistic regression models or other forms of regression, which we'll talk about later in the course, that produce a risk ratio or odds ratio. And it's much easier to assess interaction on the relative scale with those models. Now these days, there's lots of software options that allow you to more easily calculate the additive scale interaction measures. And we have a whole video dedicated to that coming up. So I'm just setting the stage for that future video.

Berkeley School of Public Health

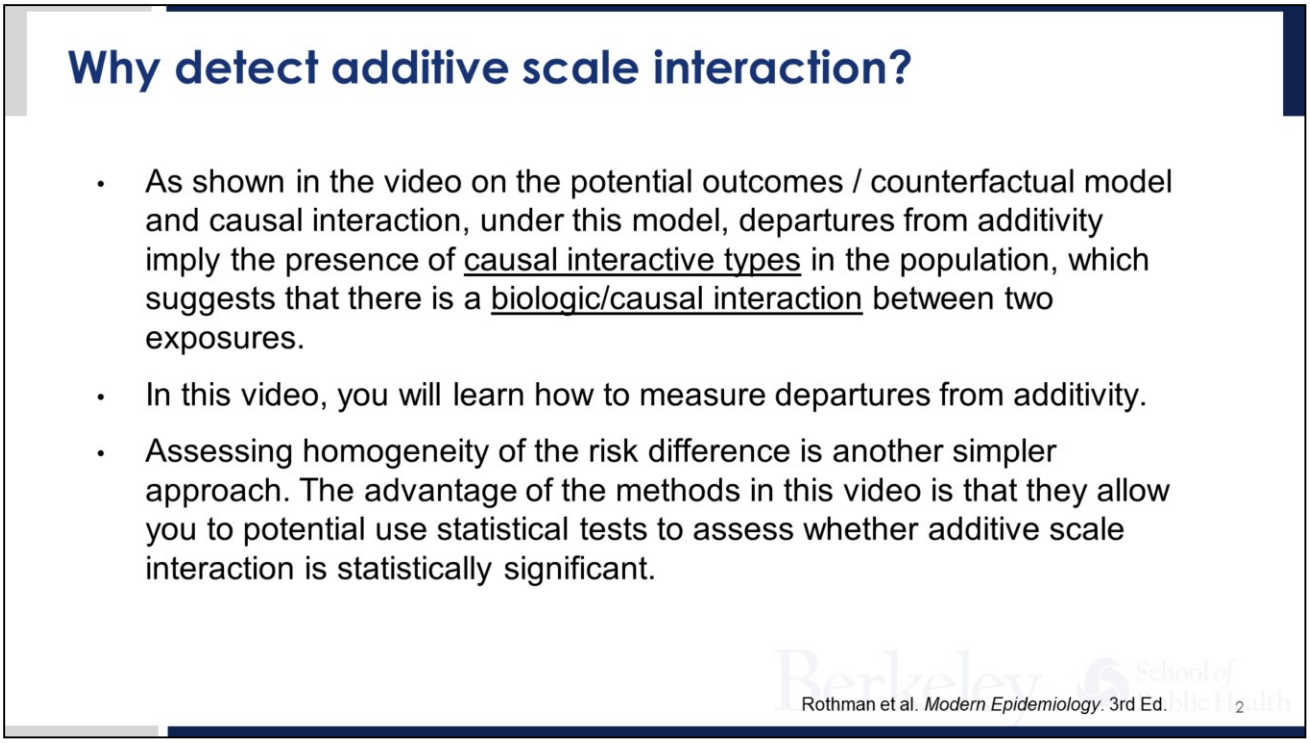
Detecting additive scale interaction

PHW250 B – Andrew Mertens

Andrew Mertens: This video talks about how to detect additivescale interaction.

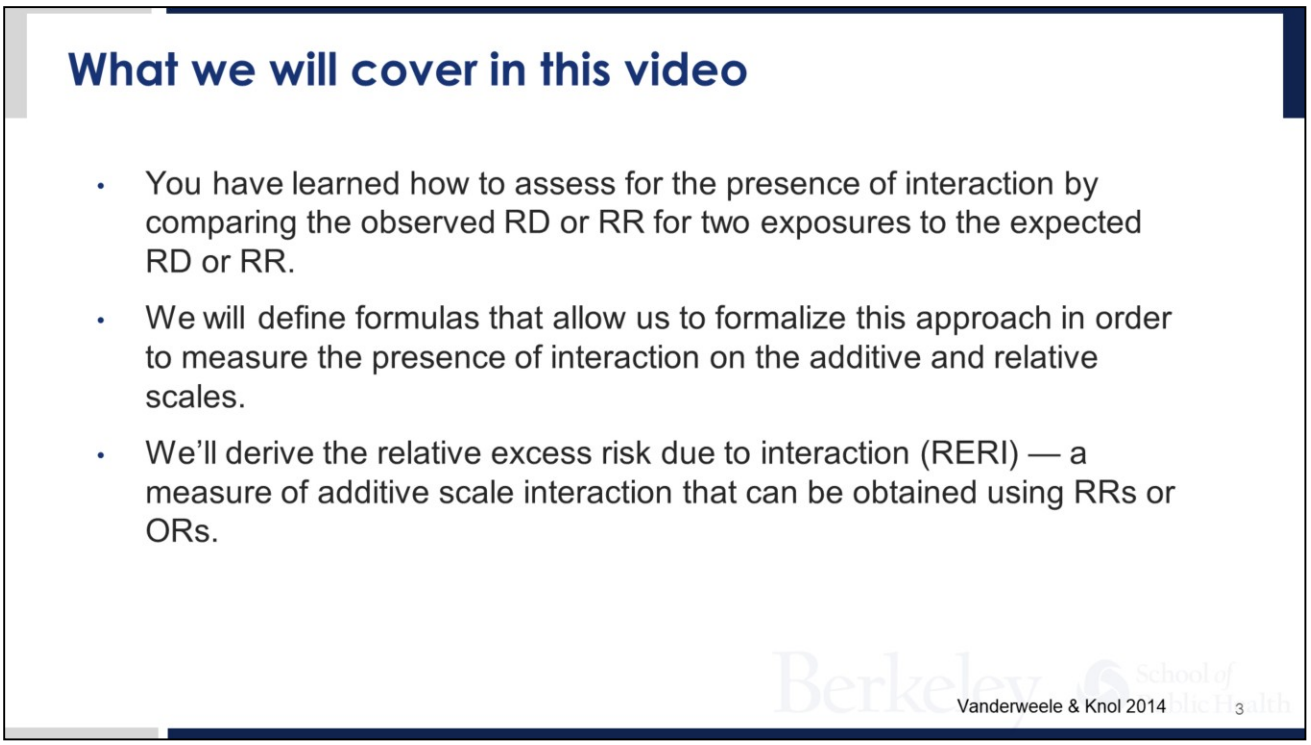


Here's our list of topics for effect measure modification and we've gone through different types of interaction. And in a prior video, I made the case that many epidemiologists believe that biologic or causal interaction are best detected on the additive scale, not on the relative scale. In this video, I'll be talking about how we can detect additive scale interaction. Whether we've used an additive scale measure of association or a relative scale measure of association.



To briefly recap what we've already discussed. When we talked about the potential outcomes and counterfactual model and how it relates to causal interaction, we showed that departures from additivity imply the presence of causal interactive types in the population, suggesting that there is a true biologic or causal interaction between two exposures. So as I just mentioned, our focus in this video is how to detect these departures from additivity.

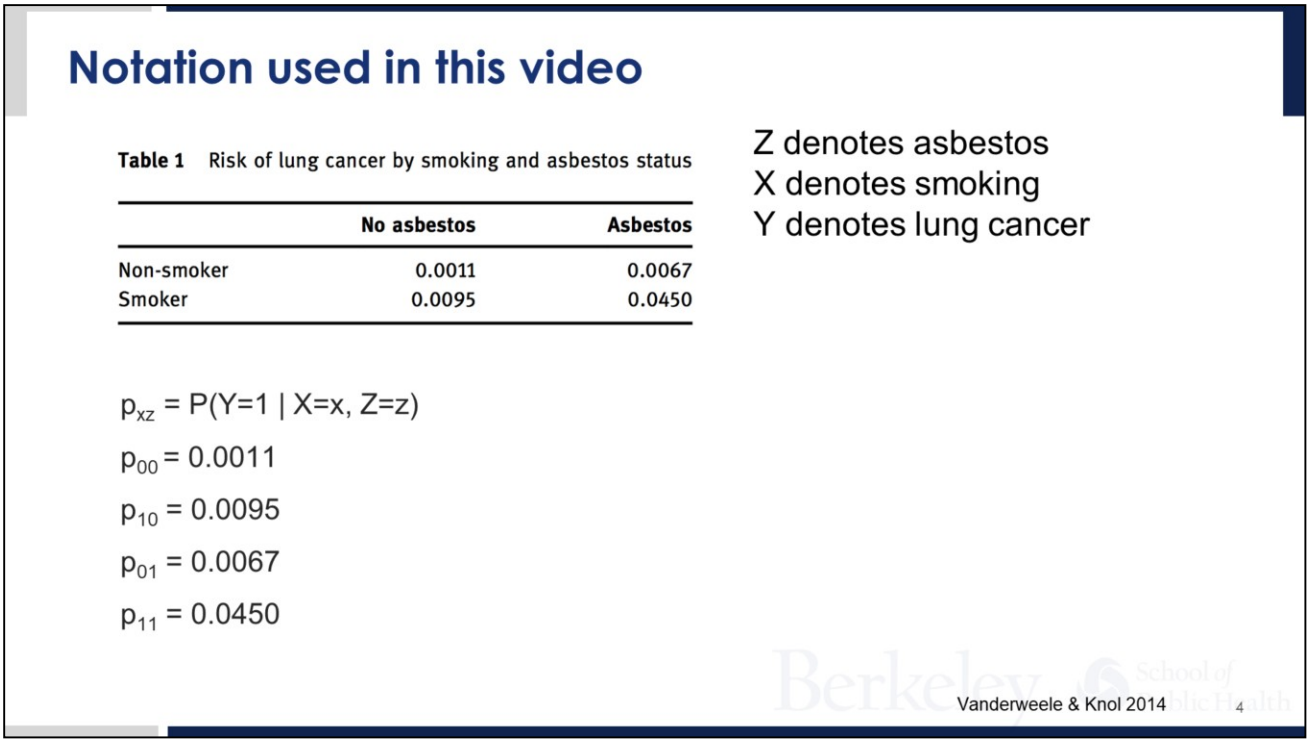
Now a simple approach to doing this if you've estimated a risk difference is just to assess homogeneity of that risk difference. The methods I'm introducing in this video have some advantages compared to that traditional approach of assessing homogeneity. Specifically, they allow you to assess the statistical significance of additive scale interaction. They also allow you to use a relative scale measure, for example a risk ratio you may see in a paper, to assess whether there is additive scale interaction. So these are relatively newer methods and that's why we'd like to introduce them to you in this course.



So you've learned how to assess the presence of interaction, comparing your observed risk difference or risk ratio for two exposures jointly to the expected risk difference or risk ratio that you obtain by, for example, adding the risk difference for their first exposure to the risk difference for the second exposure. This video will walk through formulas that formalize this essential approach to measure interaction on the additive scale. And we can also do so on the relative scale, but really our focus is going to be on the additive scale for the reasons earlier mentioned.

And then I'm going to derive something called the relative excess risk due to interaction. It's quite a term. RERI is what we'll call it. And this is a measure of additive scale interaction that we can use when we only have risk ratios or odds ratios.

Now why might we only have those things? Well, sometimes for statistical reasons we need to use a relative scale model when we're estimating our measure of association. And also you might just be reading a paper that's only reported relative scale measures and this would allow you to assess additive scale interaction in that situation as well.



Now to introduce the notation in this video, here is a two by two table with a very different layout from the one we usually use. So this is our interaction style two by two table. Our outcome is lung cancer. And our first exposure to smoking, denoted by X. Our second exposure is asbestosis, denoted by Z. And Y denotes lung cancer. And then the cells of the two by two table are probabilities.

So they're not counts. Usually we see A, B, C, D and those are counts. In this case, we have probabilities in the cells of our two by two table. And then I'm going to use the notation, P sub XZ. So P sub XZ is the probability that Y is 1, that lung cancer is present, conditional on the exposure X equal to a specific value of X and the exposure Z equal to a specific value of Z.

So P00 is the probability of lung cancer if the person is a nonsmoker and was not exposed to asbestosis. And P10 is the probability of lung cancer if the person was a smoker and was not exposed to asbestosis. So in that second one, P10, we can see right here in the two by two table that that value is 0.0095. So get familiar with this notation. We'll be using it throughout the video.

Comparing observed vs. expected RD

Table 1 Risk of lung cancer by smoking and asbestos status

	No asbestos	Asbestos
Non-smoker	0.0011	0.0067
Smoker	0.0095	0.0450

Observed RD for both smoking and asbestos

$= (p_{11} - p_{00})$

Expected RD for both smoking and asbestos

$= (p_{10} - p_{00}) + (p_{01} - p_{00})$

Vanderweele & Knol 2014 5

So here are the formulas for the observed and the expected risk difference. If we want to get the observed risk difference for both smoking and asbestos together, we need to take the probability of smoking and asbestos together minus the probability when neither are present. So that's P11 minus P00. And then the expected risk difference is the sum of two different risk differences for each exposure on its own.

So we could take first, the risk difference for smoking as P10. So that's the probability when the person is a smoker, but was not exposed to asbestos, minus the probability when they were not exposed to either. And then the second risk difference is the probability that someone is not a smoker, but was exposed to asbestos minus the probability when they were exposed to neither. So here are our formulas.

Comparing observed vs. expected RD

Table 1 Risk of lung cancer by smoking and asbestos status

	No asbestos	Asbestos
Non-smoker	0.0011	0.0067
Smoker	0.0095	0.0450

Observed RD for both smoking and asbestos

$= (p_{11} - p_{00})$

$= 0.0450 - 0.0011 = 0.0439$

Expected RD for both smoking and asbestos

$= (p_{10} - p_{00}) + (p_{01} - p_{00})$

$= (0.0095 - 0.0011) + (0.0067 - 0.0011) = 0.014$

And now we can plug the two by two cell values into these formulas. And what we get is for P11, 0.0450 minus P00, which is 0.0011. And that gives us an observed difference of 0.0439. And then for our expected risk difference, we can plug in the value for P10, which is 0.0095. And then plug in the value for P01, and that's 0.0067. And when we calculate this through, our expected risk difference is 0.014. So this so far should be review. This isn't anything new. But I just wanted to go over this to set the stage for what's coming next.

Defining a measure of interaction on the additive scale

Observed RD = $(p_{11} - p_{00})$

Expected RD = $(p_{10} - p_{00}) + (p_{01} - p_{00})$

To assess interaction on the additive scale, we could use the formula:

$(p_{11} - p_{00}) - [(p_{10} - p_{00}) + (p_{01} - p_{00})]$

Which can be re-written as:

$= p_{11} - p_{00} - p_{10} + p_{00} - p_{01} + p_{00}$

$= p_{11} - p_{10} - p_{01} + p_{00}$

So at the top of the slide, here are the formulas we just went over. And what if we could put these formulas together? So far in this course we've just separately calculated the observed risk difference and compared it to the expected risk difference. But what if we took the difference between the observed risk difference and the expected risk difference?

So if we do that, what we get is here, this formula in the center of the slide, P11 minus P00, that's the observed risk difference. And then we can subtract the expected risk difference here. The intuition behind this is that if we take the difference between the two, it will equal 0 if there's no difference. And that means that there's no interaction on the additive scale. If the value of this quantity is not equal to 0, it suggests that there is interaction on the additive scale.

So it's the same thing we've been doing. It's just putting it into a single formula. And there's advantages to that. The primary advantage being that it allows us to get a p value or a confidence interval for that estimate.

Now we can rewrite this to make it a little easier to read. So let's cancel out some of the repeated terms. So first, if we apply the minus right here through the rest of the terms, this can be rewritten as P11 minus P00 minus P10 plus P00 minus P01 plus P00. And then the P00s cancel, leaving us with P11 minus P10 minus P01 plus P00. This right here on the bottom left hand of the slide is the formula that we'll use to assess additive scale interaction.

Now it's important to keep in mind that in order to use this formula we actually have to have these four values. We might not have these four values, so I'm just foreshadowing what's coming soon. That's something for you to keep in mind as we move forward.

Defining a measure of interaction on the additive scale

If the **exposures are harmful** (increase risk of disease):

- If $p_{11} - p_{10} - p_{01} + p_{00} > 0$ interaction is positive or synergistic
- If $p_{11} - p_{10} - p_{01} + p_{00} = 0$ no interaction is present
- If $p_{11} - p_{10} - p_{01} + p_{00} < 0$ interaction is negative or antagonistic

Table 1 Risk of lung cancer by smoking and asbestos status

	No asbestos	Asbestos
Non-smoker	0.0011	0.0067
Smoker	0.0095	0.0450

$p_{11} - p_{10} - p_{01} + p_{00} = 0.045 - 0.0095 - 0.0067 + 0.0011 = 0.0299$

We conclude that there is positive interaction. The observed risk difference associated with both asbestos and smoking is greater than the expected risk difference for both.

Vanderweele & Knol 2014 8

Now if we have an exposure that is harmful, so an exposure that increases the risk of disease, like pollution, or a chemical exposure, then the following equalities are true. So if this value is greater than 0, we have a positive or a synergistic interaction. If it's equal to 0, there's no interaction on the additive scale. And if it's less than 0, we have a negative or an antagonistic interaction.

So let's go over our example. When we plug in the values from our two by two table into this formula, we have 0.045 minus 0.0095 minus 0.0067 plus 0.0011. That's equal to 0.0299, which is greater than 0. And so we conclude that there's positive interaction. In other words, the observed risk difference for both asbestosis and smoking is larger than the expected risk difference for each exposure on its own. So that means that if you're someone who smokes and you also are exposed to asbestos, you will have a greater risk than the risk total for if you were just a smoker or you were just exposed to asbestos.

Defining a measure of interaction on the additive scale

If the **exposures are harmful** (increase risk of disease):

- If $p_{11} - p_{10} - p_{01} + p_{00} > 0$ interaction is positive or "super-additive"
  - The public health consequence of an intervention on Z (e.g., asbestos) would be larger in the X=1 (e.g., smoker) group.
- If $p_{11} - p_{10} - p_{01} + p_{00} = 0$ no interaction is present
  - The public health consequence of an intervention on Z (e.g., asbestos) would be the same in both groups X=1 and X=0 (e.g., smokers and non-smokers).
- If $p_{11} - p_{10} - p_{01} + p_{00} < 0$ interaction is negative or "sub-additive"
  - The public health consequence of an intervention on Z (e.g., asbestos) would be larger in the X=0 (e.g., non-smoker) group.

Now let's go over each of these. If we have a positive or a super-additive interaction when this quantity is greater than 0, the public health consequence of our intervention on Z, asbestosis, would be larger in the smoker group. And that's because these two exposures are synergistically working together. So if we do something to reduce asbestosis, the impact will be greater among smokers than among nonsmokers.

If there's no interaction present, if this quantity equals 0, then any intervention on asbestosis would have the same impact on smokers and nonsmokers. And if this quantity is less than 0, we have a negative interaction or a sub-additive interaction. And the public health consequence of intervening on asbestosis would be larger among the nonsmoker group than among the smoker group.

Comparing observed vs. expected RR

Table 1 Risk of lung cancer by smoking and asbestos status

	No asbestos	Asbestos
Non-smoker	0.0011	0.0067
Smoker	0.0095	0.0450

Observed RR for both smoking and asbestos

$= (p_{11} / p_{00})$

Expected RR for both smoking and asbestos

$= (p_{10} / p_{00}) \times (p_{01} / p_{00})$

Now let's go through these steps for the relative risk. So the formula for the observed relative risk for both smoking and asbestos is P11 over P00. And then the expected relative risk for smoking and asbestos together is P10 over P00 times, because we're on the relative scale, P01 over P00.

Comparing observed vs. expected RR

Table 1 Risk of lung cancer by smoking and asbestos status

	No asbestos	Asbestos
Non-smoker	0.0011	0.0067
Smoker	0.0095	0.0450

Observed RR for both smoking and asbestos

$= (p_{11} / p_{00})$

$= 0.0450 / 0.0011 = 41$

Expected RR for both smoking and asbestos

$= (p_{10} / p_{00}) \times (p_{01} / p_{00})$

$= (0.0095 / 0.0011) \times (0.0067 / 0.0011) = 53$

And then if we plug this in. I won't go through it as slowly because we've done this before the risk difference. We get a value of 41 for our observed risk ratio and a value of 53 for our expected risk ratio.

Defining a measure of interaction on the relative scale

Observed RR = $(p_{11} / p_{00}) = RR_{11}$

Expected RR = $(p_{10} / p_{00}) \times (p_{01} / p_{00}) = RR_{10} \times RR_{01}$

To assess interaction on the relative scale, we could use the formula:

$(p_{11} / p_{00}) / [(p_{10} / p_{00}) \times (p_{01} / p_{00})] = RR_{11} / (RR_{10} \times RR_{01})$

Vanderweele & Knol 2014 12

Now I'll come back to those specific values in a moment. We can define a measure of interaction on the relative scale in much the same way as we did on the additive scale. Here are our formulas. And we can basically do something very similar, where we take the observed relative risk and we divide it by the expected relative risk.

So here P11 over P00 is the observed relative risk. And we can divide that by the expected relative risk right here. This simplifies to RR11 over RR10 times RR01.

Defining a measure of interaction on the relative scale

If the **exposures are harmful** (increase risk of disease):

- If $RR_{11} / (RR_{10} \times RR_{01}) > 1$ interaction is positive or "super-multiplicative"
- If $RR_{11} / (RR_{10} \times RR_{01}) = 1$ no interaction is present
- If $RR_{11} / (RR_{10} \times RR_{01}) < 1$ interaction is negative or "sub-multiplicative"

Table 1 Risk of lung cancer by smoking and asbestos status

	No asbestos	Asbestos
Non-smoker	0.0011	0.0067
Smoker	0.0095	0.0450

$RR_{11} / (RR_{10} \times RR_{01}) = (0.0450/0.0011) / [(0.0095/0.0011) \times (0.0067/0.0011)] = 0.78$

We conclude that there is negative interaction. The observed risk ratio associated with both asbestos and smoking is less than the expected risk ratio for both.

Vanderweele & Knol 2014 13

For exposures that are harmful, this quantity here, R11 over R10 times RR01 is greater than 1 when we have a positive interaction. It's equal to 1 when we have no interaction. And it's less than 1 when the interaction is negative.

So for our previous example, if we take the observed RR and we divide it by the expected RR, here is all the numbers filled in here. If you work through the math, that's equal to 0.78. Because this quantity is less than 1 and we have a harmful exposure, we conclude that there is negative interaction. The observed risk ratio associated with asbestos and smoking is less than the expected risk ratio for both.

So as we've talked about earlier in the course, this discrepancy between our conclusion about interaction is expected. It's common to see interaction on both scales or interaction on one scale or the other. Mathematically, if we see the absence of interaction on one scale, it does imply the presence on another scale.

Assessing additive interaction when only RRs are available

- In some cases, we may only be able to estimate an RR, or we may want to evaluate additive scale interaction in a study that only reported RRs.
- We can use the "relative excess risk due to interaction" (RERI) to do so.

We start with our formula for assessing additive scale interaction:
$p_{11} - p_{10} - p_{01} + p_{00}$

Then divide it through by p_{00}:

$(p_{11} / p_{00}) - (p_{10} / p_{00}) - (p_{01} / p_{00}) + (p_{00} / p_{00})$

$RERI = RR_{11} - RR_{10} - RR_{01} + 1$

Now we've gone through all these steps so far to come to the point that I'm about to make in this slide. Which is that often we only have relative scale measures available. And this will be explained more in our unit on statistical analyses for epidemiology. But for now, let's take it as a given that sometimes we only have relative scale measures.

When this is the case, in the past, it would have been quite difficult to assess additive scale interaction. But these newer methods allow us to evaluate additive scale interaction when we only have risk ratios or odds ratios. And so we can estimate something called the relative excess risk due to interaction, RERI, using risk ratios or odds ratios.

So let's start with this formula that we use to assess additive scale interaction right here. P11 minus P10 minus P01 plus P00. Now if we divide everything through by P00, here is what we get. Now notice that P11 over P00 is equivalent to the risk ratio for both exposures X and Z. And the second term, P10 over P00, is the risk ratio for the presence of X and the absence of Z. The next term is the risk ratio for the absence of X and the presence of Z. And then the last term is equal to 1.

So this is what we call the RERI. This formula right here. It's essentially the same formula we use to assess additive scale interaction, but we've just divided everything through by the probability of our disease when neither exposure is present.

Defining a measure of interaction on the additive scale

For **harmful exposures** that increase the risk of disease:

- If RERI = $RR_{11} - RR_{10} - RR_{01} + 1 > 0$ interaction is positive or "super-additive"
- If RERI = $RR_{11} - RR_{10} - RR_{01} + 1 = 0$ no interaction is present
- If RERI = $RR_{11} - RR_{10} - RR_{01} + 1 < 0$ interaction is negative or "sub-additive"

For **preventive exposures** that decrease the risk of disease:

- If RERI = $RR_{11} - RR_{10} - RR_{01} + 1 < 0$ interaction is positive or "super-additive"
- If RERI = $RR_{11} - RR_{10} - RR_{01} + 1 = 0$ no interaction is present
- If RERI = $RR_{11} - RR_{10} - RR_{01} + 1 > 0$ interaction is negative or "sub-additive"

The Vanderweele & Knol 2014 paper includes a derivation for the odds ratio as well.

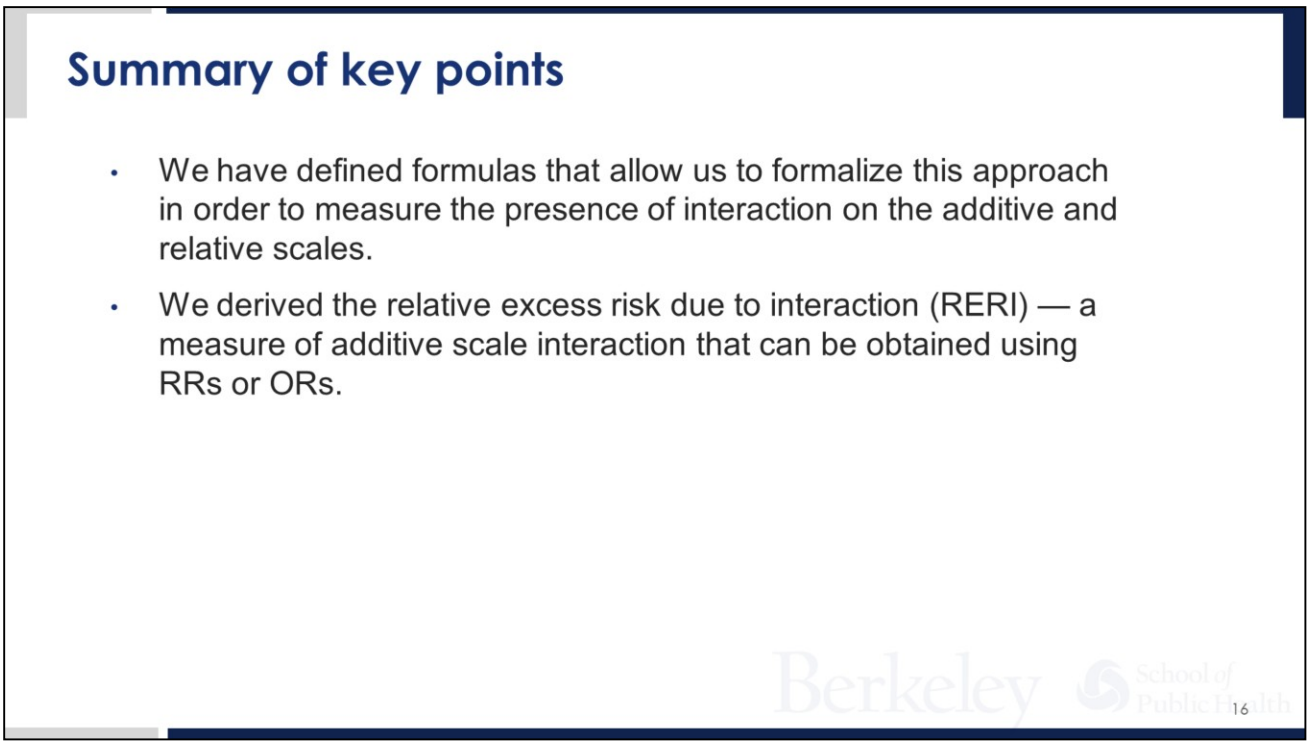
The paper also includes example code for estimating the RERI and obtaining standard errors for the RERI.

Vanderweele & Knol 2014 15

For harmful exposures, we can use these formulas to assess what type of interaction is present if any. So if the RERI is greater than 0, we say that the interaction is positive or super-additive. If the RERI equals 0, then no interaction is present. If the RERI is less than 0, then we have negative interaction.

And to be very clear, this is additive scale interaction. So we're using relative risks to assess additive scale interaction. And then we can use the same formula, just switching the signs when we have a preventive exposure. So VanderWeele and Knol have a nice paper called Tutorial on Interaction and it shows you the derivation for the odds ratio as well. We're just focusing on the relative risk here, but you can refer to that paper if you'd like to do this with an odds ratio.

The paper also has some great code that you can use to try to estimate the RERI. And also to obtain standard errors for the RERI, as well as the other measure of additive scale interaction earlier in this talk. So this would be useful if you want to obtain a confidence interval around your RERI.



To summarize. In this video, we have defined formulas that allow us to formalize our approach to measure the presence of interaction on the additive and relative scales. So previously, we had looked at the homogeneity of the risk difference or the relative risk. And fundamentally, this approach is doing the same thing. It's formalizing it by defining formulas.

And the main advantage of that is that it potentially allows us to detect additive scale interaction with relative risks or odds ratios. And it allows us to obtain a confidence interval around that estimate. And we derived the relative excess risk due to interaction, a measure of additive scale interaction. Now this is a relatively new type of measure. And so if you're looking at older epidemiology papers, you're unlikely to see this. But in the last few years, this has become increasingly common, and so that's why we wanted to introduce it to you in this course.

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.

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

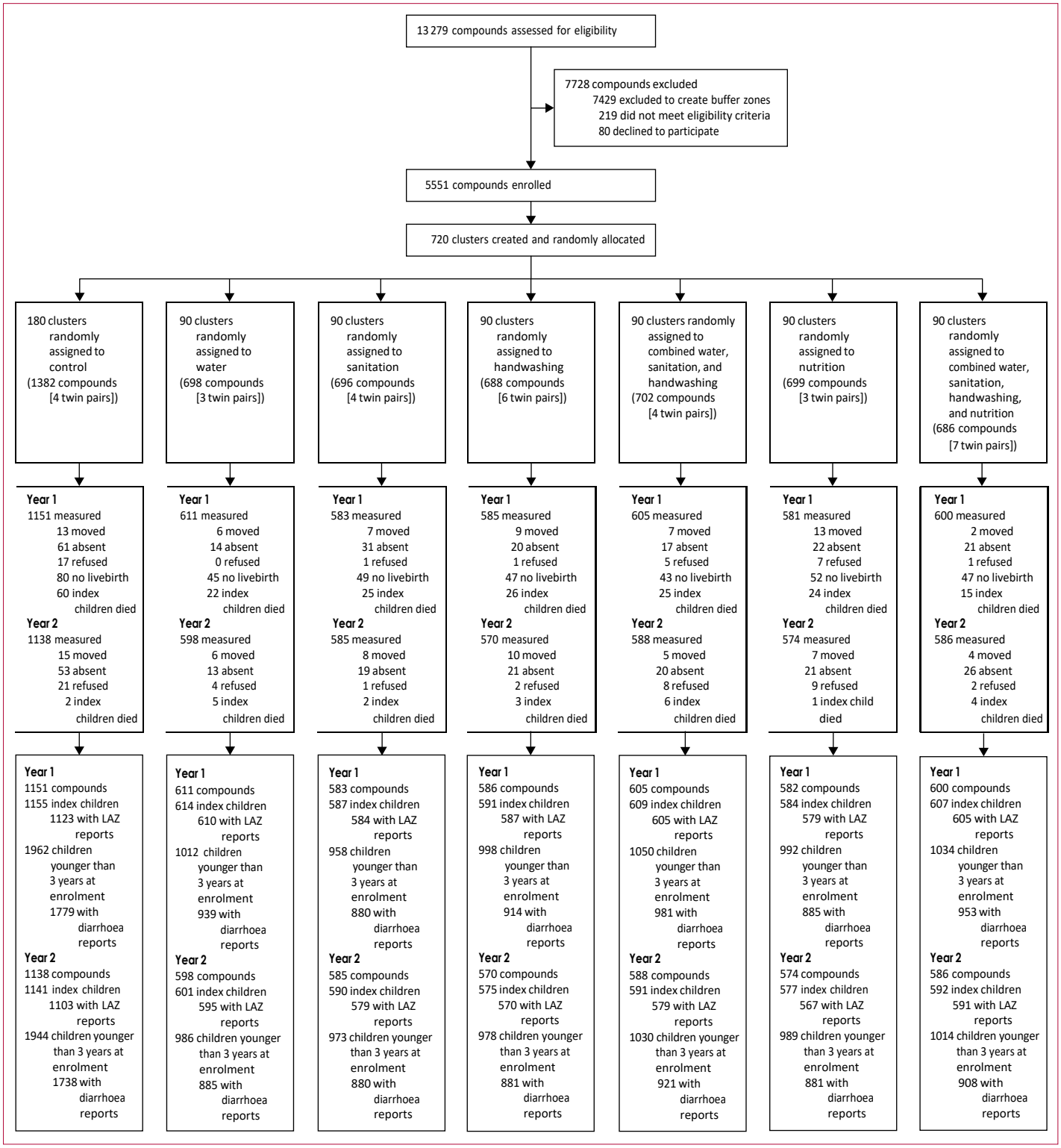


Figure 1: **Trial profile and analysis populations for primary outcomes**
LAZ=length-for-age *Z* scores.

	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 1e3n7rolled 5551 pregnant women in 5551 compounds to an

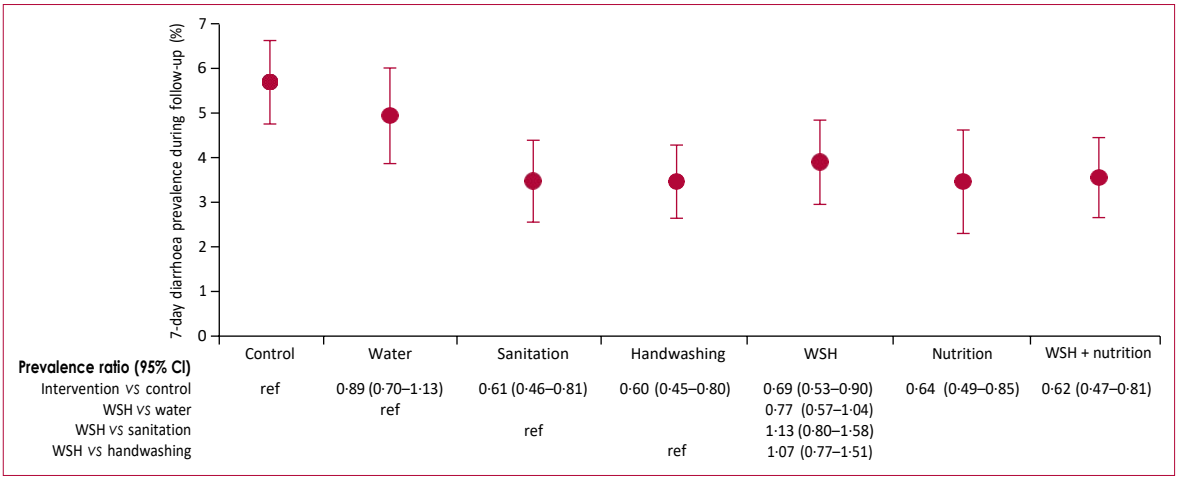


Figure 2: **Intervention effects on diarrhoea prevalence in index children and children younger than 3 years at enrolment 1 and 2 years after intervention**
Data are mean (95% CI). ref=reference. WSH=water, sanitation, and handwashing.

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

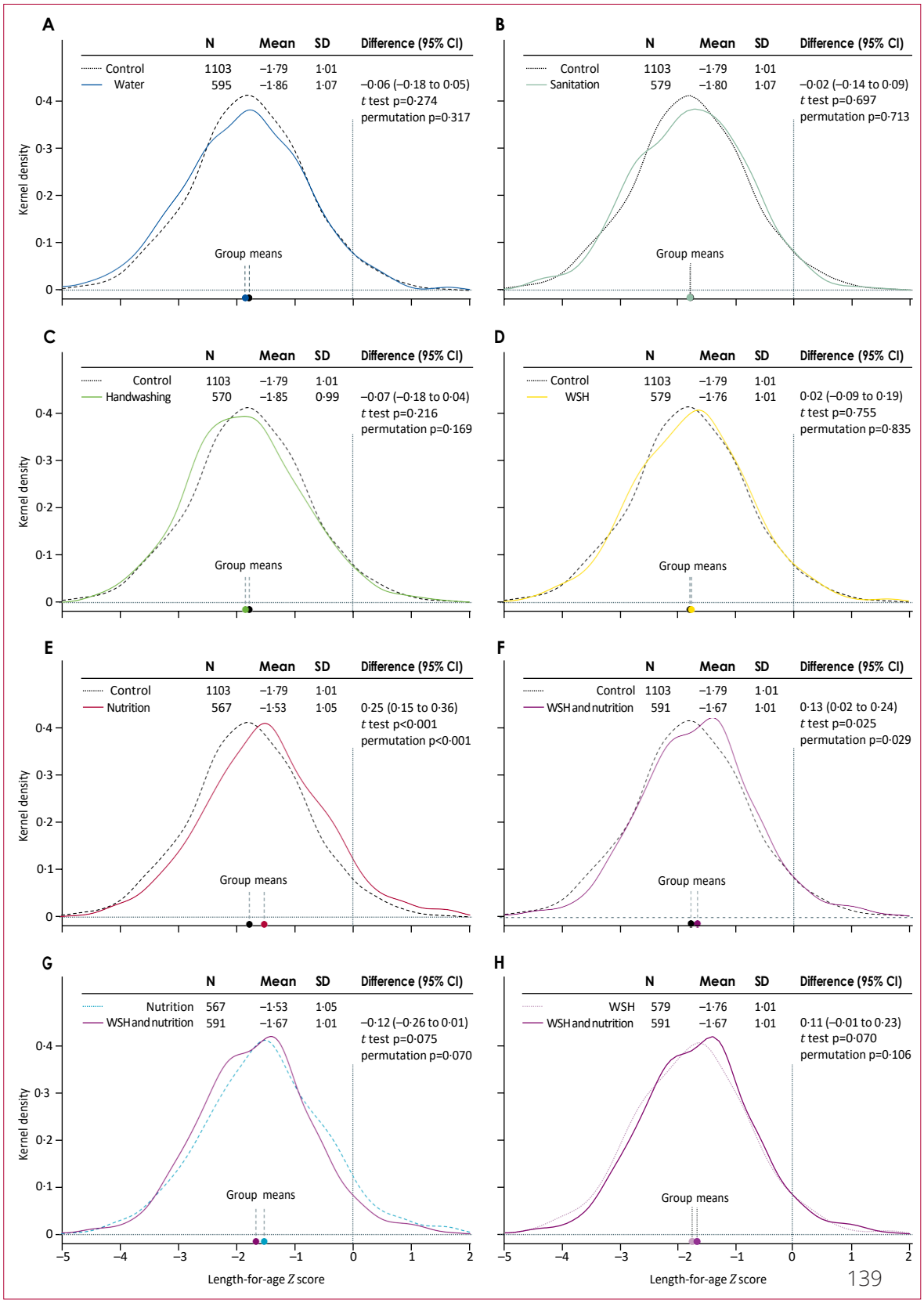


Figure 3: **Intervention effects on length-for-age *Z* scores in 4584 children after 2 years of intervention**

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.

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.

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

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 = $P = .04$).

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.

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.

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-

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-

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.