PHW 250B Week 14 Reader

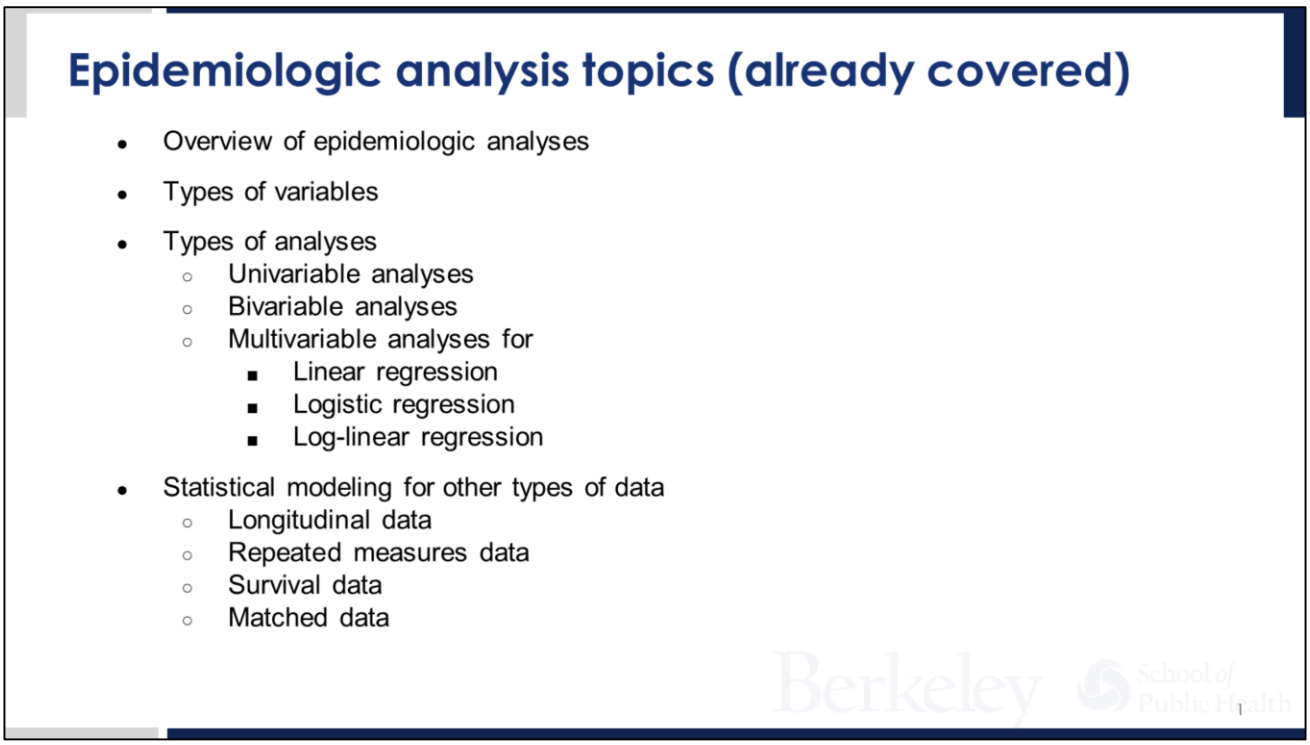
Topic 1: Modern Epidemiologic Analysis Approaches

Topic 2: Statistical Inference

Topic 3: Sample Size and Power

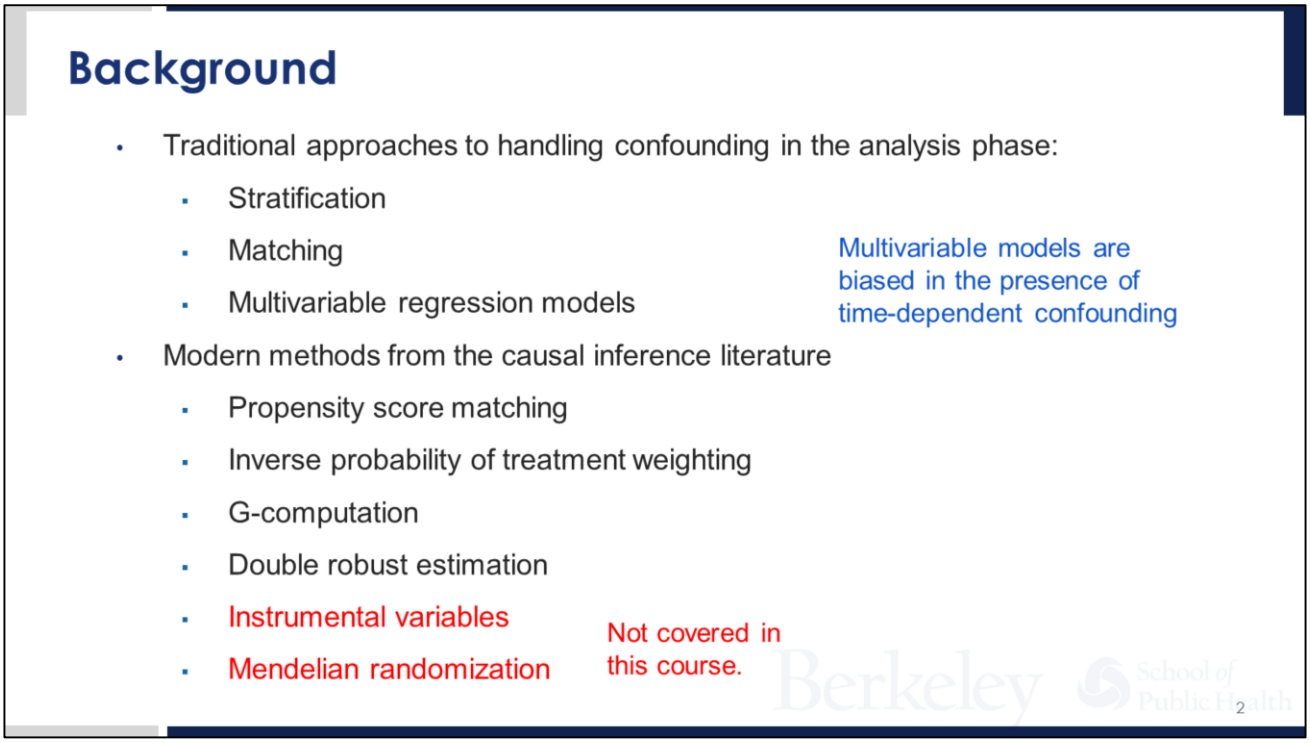
Journal Club

JADE BENJAMIN-CHUNG: In this video, we'll talk about propensity score matching.



Here's a list of the analysis topics we've already covered. So we've talked about different kinds of variables and types of analyses-- univariable, bivariable, and multivariable analyses-- and we went into detail about different types of regression models we can use to conduct adjusted statistical analysis.

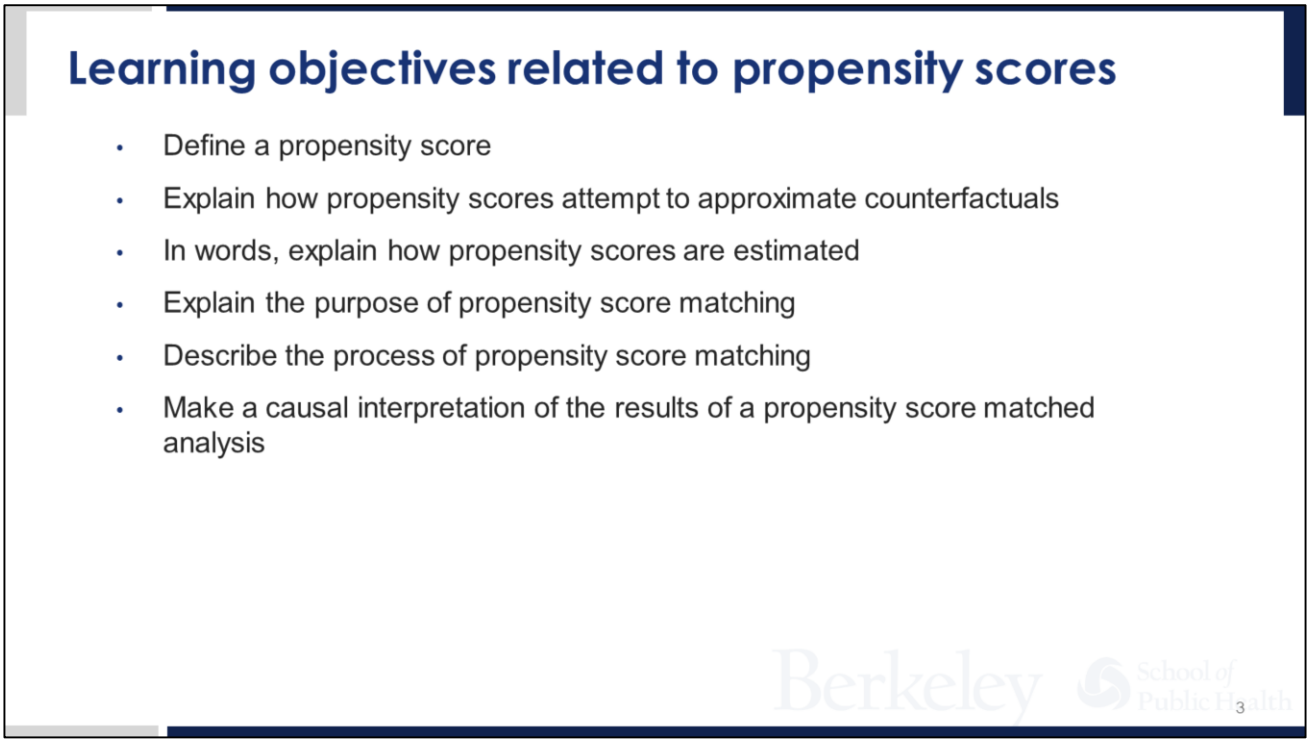
We also briefly touched on statistical modeling for other types of data.



We've learned about three primary different ways that we can handle confounding in the analysis phase of a study-- stratification, matching, and multivariable regression models. Multivariable regression models are most frequently used because they allow us to adjust for multiple potential confounders together. However, when something called time-dependent confounding is present in one of our studies, unfortunately multivariable models are biased. That means our estimated measures of association will not be accurate.

So this particular issue of time-dependent confounding was one thing that helped motivate the development of new, more modern methods. And many of these methods were derived from the causal inference literature, but that doesn't mean that they can only be used to make causal inferences. We can also just use them to estimate new and more potentially useful types of measures of association. So we'll return to that idea in a later video.

But for now, here's a list of some of these different, more modern topics-- propensity score matching, inverse probability of treatment weighting, G-computation, double robust estimation, instrumental variables, and Mendelian randomization. Now, these last two we won't cover in this course, and double robust estimation we're only going to briefly touch on so you understand broadly what it's doing.



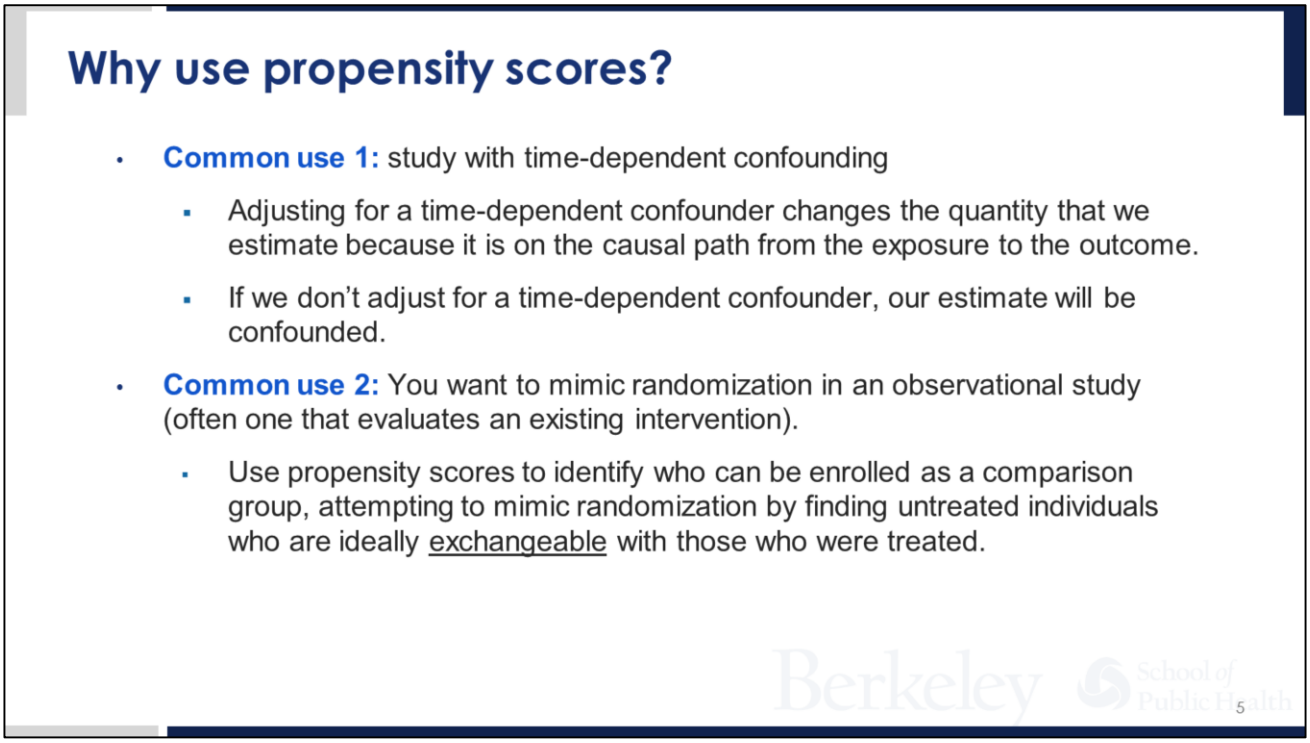
We're focusing in this video on propensity scores. And it may seem quite technical at times in the video, so I just want to remind you of what the learning objectives are for this topic. Really, it's at a high level, so to define a propensity score, be able to explain how propensity scores can help us approximate counterfactuals, explain how they are estimated and the purpose of using them in propensity score matching, the process of propensity score matching, and then finally, be able to make an interpretation, causally, of the results of a propensity score matched analysis.

Definition of propensity scores

- **Propensity score:** the predicted probability ("propensity") of exposure in a particular individual based on a set of characteristics

(Szklo & Nieto, 3rd) Ed. 4

So here is the definition of a propensity score from the Szklo and Nieto textbook. It's the predicted probability or propensity of exposure in a particular individual based on a set of characteristics. Now, this is defined around exposures, but we can also define this around the probability that somebody receives a particular treatment or a particular intervention.



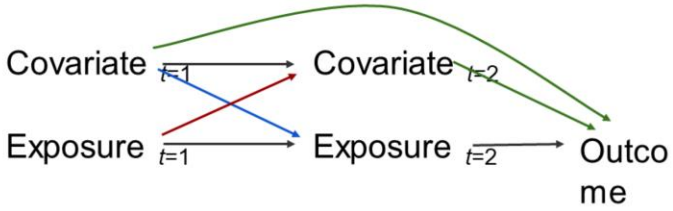
So what are the reasons we would want to use a propensity score matching approach? Well, there's two common uses. The first is that we have a study in which we're concerned about time-dependent confounding, and I'm going to define that in just a moment. To briefly motivate it though, if we adjust for a time-dependent confounder, that changes the quantity that we estimate, and we'll see why when we take a look at the directed acyclic graph that illustrates what a time-dependent confounder looks like. But if we don't adjust for this type of time-dependent confounder, our estimate will be confounded.

OK. So we're going to come back to that. That's just prefacing you for what's coming. The second common use is that we want to mimic randomization in an observational study. And this is really common when we're interested in evaluating a pre-existing intervention. So perhaps there is an NGO or a government that established a new public health intervention, and we want to evaluate whether it improved health, but the intervention was not randomized.

So in this case, we can use propensity scores to identify the comparison groups of people who did not get the intervention so that these individuals in the comparison group are as close to a control group that we would have gotten through randomization as possible. And this mimics randomization by finding untreated individuals who are as close to exchangeable as possible with those who were treated. So we're actually going to spend the majority of this video focusing on this latter example, this latter common use. But for now let me briefly go into time-dependent confounding.

Time dependent confounding

- A time-dependent confounder is a variable that:
  - Is affected by prior exposure
  - Predicts subsequent exposure
  - Associated with / causes the outcome
- In the DAG below we use subscripts to indicate the time (t) of measurement.

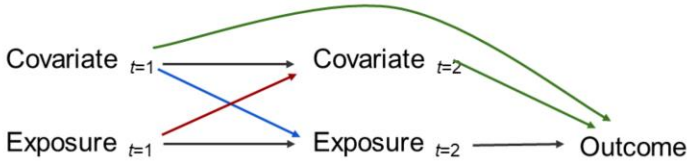


So a time-dependent confounder is a variable that is affected by prior exposure, predicts subsequent exposure, and is associated with or causes the outcome. So take a look at this DAG. We've got a covariate that, at time 1-- so we're using little t here, subscript, to indicate the time of measurement. Covariate at time 1 predicts covariate at time 2. Covariate at time 1 also predicts exposure at time 2. And covariate at time 1 predicts the outcome, which in this case is not subset by time. Perhaps this means that it's measured much later.

And so this is an example of a time-dependent confounder, because what we can see with the red arrow is that, at time 2, the covariate is affected by prior exposure at time 1. And as we can see with the blue arrow, at time 1, the covariate predicts subsequent exposure at time 2. And then the covariate at time 1 and the covariate at time 2 are associated with or cause the outcome.

Time dependent confounding

- A time-dependent confounder is a variable that:
  - Is affected by prior exposure
  - Predicts subsequent exposure
  - Associated with / causes the outcome
- In the DAG below we use subscripts to indicate the time (t) of measurement.



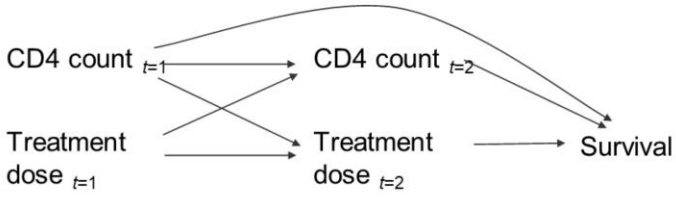
- Adjusting for a time-dependent confounder changes the quantity that we estimate because it is on the causal path from the exposure to the outcome.
- If we don't adjust for a time-dependent confounder, our estimate will be confounded.

So again, why does this matter? If we adjust for a time-dependent confounder, it changes the quantity that we estimate. Now why is this? Well, let's take a look at our DAG here. If we adjust for the covariate, look at how the covariate at time t equals 2 is on the causal path from the exposure at time 1 to the outcome.

So we've touched on this at another point in the class, but if we adjust for a covariate that's along the causal pathway from exposure to outcome, it changes the quantity that we're estimating. And so we're no longer really getting the association or the effect of the exposure on the outcome in isolation. And then if we ignore this covariate, our estimate will be confounded, and we don't want that either.

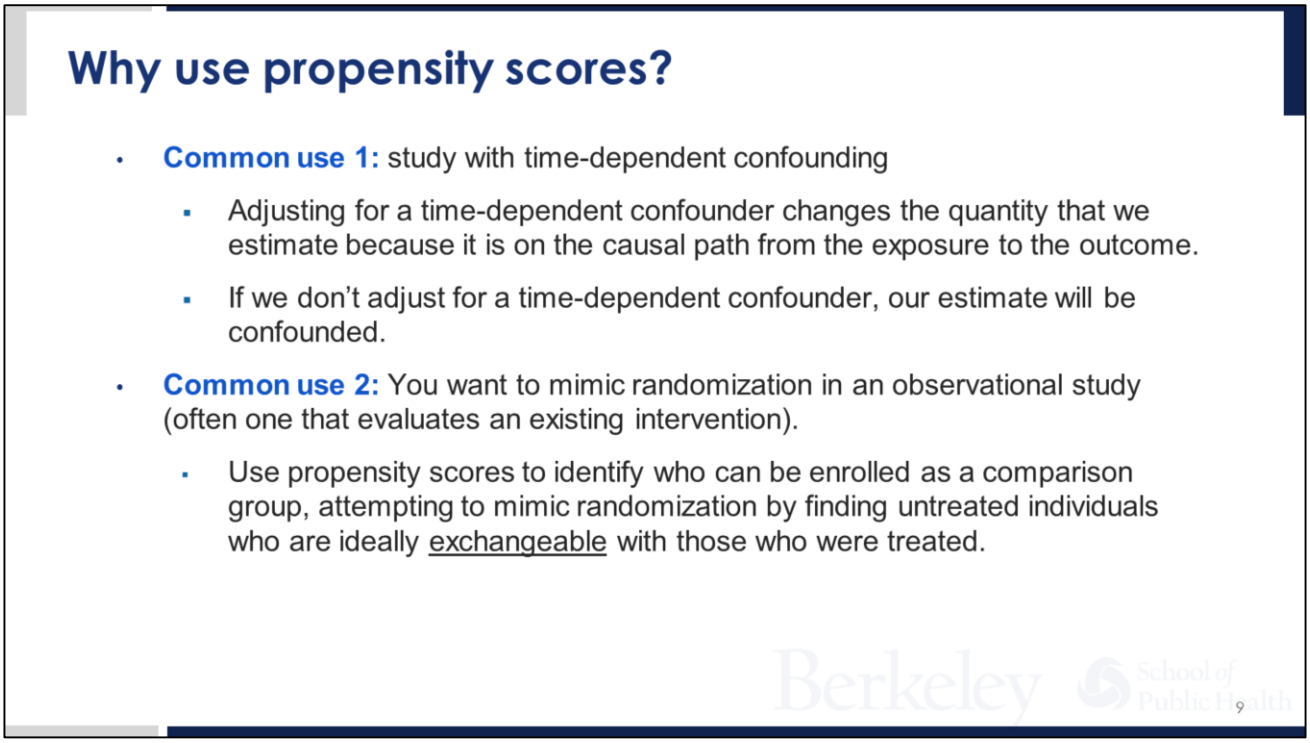
Example of time dependent confounding

- The HIV virus reduces the number of CD4 lymphocyte cells (T cells) in the body when untreated.
- A patient's CD4 count affects a physician's choice of treatment dose for an HIV patient, and the treatment dose can in turn affect later CD4 count.



Here's an example of time-dependent confounding. The HIV virus reduces the number of CD4 lymphocyte cells, which are T cells, in the body when it goes untreated. So this would be, for example, a person who might not know that they're infected with HIV, as the infection progresses, their number of CD4 cells will go down. And then accordingly, a patient's CD4 count can affect their physician's choice of treatment dose. And then the treatment dose can in turn affect later CD4 count.

So here we look at this DAG, and we see that treatment dose at time 1 affects the CD4 count at time 2, which affects survival. Yet, CD4 count at time 1 affects the treatment dose at time 2, which affects survival. So this is a really commonly used example of time-dependent confounding.



OK, so that was common use 1. We're really going to spend the remainder of this video on common use 2. And we're not going to statistically prove to you why using propensity score matching helps us in common use 1, because showing you that requires us to get more technical than we need to in this class. We're more interested in common use 2. So in this case, we want to mimic a randomization when we have observational data.

Propensity score matching to evaluate an existing intervention

- Example: a school-located influenza vaccination program was deployed in 30 schools in the San Francisco Bay Area in 2014.
  - Example is loosely based on an evaluation of the Shoo the Flu Program
- You want to rigorously evaluate whether the program reduced influenza among students and their household members.
- However, you must use an observational design because the program did not want to randomize schools to the program vs. control.

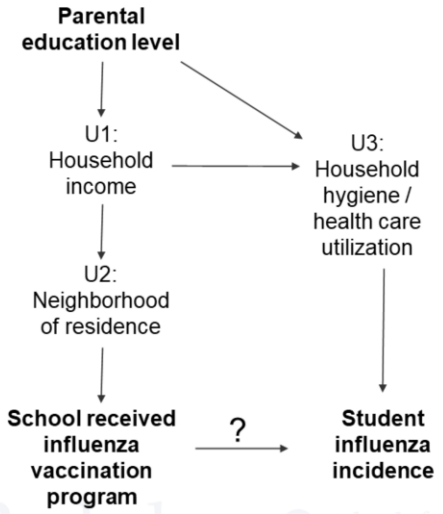
10

And we're going to use an example of a school-located influenza vaccination program deployed in 30 schools in the San Francisco Bay Area in 2014. And so this is loosely based on a real study that we're conducting that's evaluating something called the Shoo the Flu Program. So let's say this program was deployed in 2014, and you came along and you wanted to rigorously evaluate whether the program reduced influenza among students and their household members. And the idea behind this is simply that students of elementary school age are responsible for the majority of influenza transmission, because they congregate in schools and they tend to have poorer hygiene than adults.

Targeting those children for influenza vaccination can potentially reduce transmission enough that we would see benefits not only for those school children but also their household members and community members. Now, in this case, because the program was implemented already, we need to use an observational design. We're not able to randomize schools to receive the program or to control. So this is a situation where propensity score matching can be extremely helpful.

Propensity score matching to evaluate an existing intervention

- **Step 1:** Obtain publicly available pre-intervention data from 2013 from the California Department of Education on all schools in the San Francisco Bay Area.
- Variables include:
  - Average class size
  - Student race
  - Parent education level
  - Standardized test scores
  - % of students receiving free and reduced price lunch
  - % of English language learners
- Goal is to obtain data on variables that are potential confounders.



11

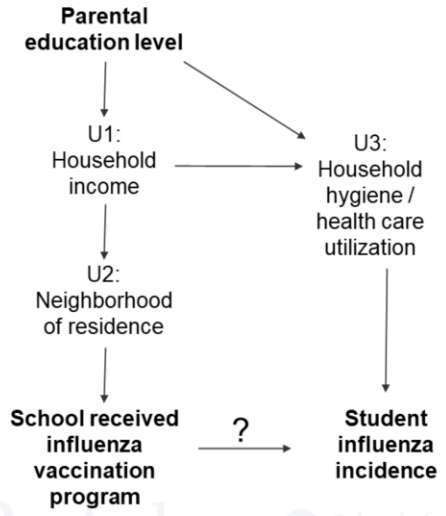
So in the first step, we need to obtain publicly available pre-intervention data. In other studies, we don't necessarily need this to be public data, but this is often the reality, is that we need data that's easy to access, such as census data. And in this example, we used data from the California Department of Education on all schools in the San Francisco Bay Area. So this includes variables at the school level like the average class size, student race, parent education, standardized test scores, et cetera.

And so this data set provides us with information about potential confounders. What we're really interested in is variables that affect whether a school received the influenza vaccination program that also affect student influenza incidence. This DAG shows you an example of how this relationship might be true for parental education level.

So that's something that is measured in this data set from the California Department of Education. Now, we don't have information on household income, neighborhood of residence, or household hygiene and health care utilization. So these are all flagged as U's in the DAG, because they're unmeasured. So parental education level could affect household income, which could in turn affect neighborhood of residence, and that affects the particular school that a child attends. And then parent education level can also affect household hygiene and health care utilization.

Propensity score matching to evaluate an existing intervention

- **Step 1:** Obtain publicly available pre-intervention data from 2013 from the California Department of Education on all schools in the San Francisco Bay Area.
- Variables include:
  - Average class size
  - Student race
  - Parent education level
  - Standardized test scores
  - % of students receiving free and reduced price lunch
  - % of English language learners
- Goal is to obtain data on variables that are potential confounders.



11

We can even draw another arrow in here from U2 to U3, but it doesn't affect our decision at the end of the day. From this DAG, we conclude that there is an open back door pathway between our exposure and our outcome through parental education level. And so this is something that we need to control for in some fashion in our study design.

So to recap, step 1 is all about obtaining data on your population of interest before the intervention. And this will help us eventually select our comparison group. We're going to look for other schools that are very similar to those that received the influenza vaccination program, and enroll them as controls. Why is it important for it to be pre-intervention? Well, if the intervention affects any of these variables, then they wouldn't be able to be used as valid factors to ensure that we have a good comparison group.

Propensity score matching to evaluate an existing intervention

- **Step 2:** Estimate the probability of being treated or exposed conditional on a set of characteristics using a regression model (typically logistic regression)
  - Y = influenza incidence
  - A = child attends school with influenza vaccination program
  - W = confounders

$$\ln\left(\frac{\Pr(A|W=w)}{1-\Pr(A|W=w)}\right) = \beta_0 + \beta_1 W_1 + \beta_2 W_2$$

OK. Step 2-- in step 2, we estimate the probability of being treated or exposed conditional on a set of characteristics using a regression model. Logistic regression is really commonly used in this case, because we usually define treatment as a binary variable. And for statistical reasons, logistic regression is convenient because it ensures that the predicted values coming out of the model will stay within the range of 0 to 1.

Now, let's define Y as our influenza incidence, A as the child attends school with an influenza vaccination program, so that's our treatment intervention variable. And W is a confounder. So to estimate this probability of being treated conditional on a set of characteristics, this is what it would look like in a logistic regression model. So if you're wondering how we got this notation, go back to the video on logistic regression.

This is the logit function right here, the log odds of the probability. But what we've changed is that, in the logistic regression video from before, the probability was of the outcome conditional on the exposure and covariates. This time, the probability is of the exposure or of the treatment conditional on covariates. The outcome doesn't factor into this at all. So we're only modeling the probability of a child attending a school with influenza vaccination.

Propensity score matching to evaluate an existing intervention

- **Step 2:** Estimate the probability of being treated or exposed conditional on a set of characteristics using a regression model (typically logistic regression)
  - Y = influenza incidence
  - A = child attends school with influenza vaccination program
  - W = confounders

$$\ln\left(\frac{\Pr(A|W=w)}{1-\Pr(A|W=w)}\right) = \beta_0 + \beta_1 W_1 + \beta_2 W_2$$

Now in this model, we have two covariates, or two confounders, W1 and W2. In reality though, we would almost always have more than this number of confounders. And that's really the beauty of propensity score matching, is that it allows us to take a large number of potential covariates or confounders and use them to predict treatment in a multivariable fashion, so that we're matching on multiple variables at once. But for the purpose of this example, we're just going to use two to keep things simple.

Propensity score matching to evaluate an existing intervention

- **Step 3:** Use the regression model fit to estimate the probability of being treated or exposed conditional on a set of characteristics for each school

$$\ln\left(\frac{\Pr(A|W=w)}{1-\Pr(A|W=w)}\right) = \beta_0 + \beta_1 W_1 + \beta_2 W_2$$

$$\ln\left(\frac{\Pr(A|W=w)}{1-\Pr(A|W=w)}\right) = -1 + 2\cdot W_1 + 3\cdot W_2$$

$$\Pr(A|W=w) = \frac{1}{1+e^{-(-1+2\cdot W_1+3\cdot W_2)}}$$

\ln\bigg(\frac{\Pr(A|W=w)}{1-\Pr(A|W=w)}\bigg) = \beta_0 + \beta_1 W_1 + \beta_2 W_2

\ln\bigg(\frac{\Pr(A|W=w)}{1-\Pr(A|W=w)}\bigg) = -1 + 2\cdot W_1 + 3\cdot W_2

\Pr(A|W=w) = \frac{1}{1+e^{-(-1 + 2\cdot W_1 + 3\cdot W_2)}}

In step 3, we use the model fit to estimate the probability of being treated or exposed conditional on a set of characteristics for each school. So here's our formula from the last slide. Now, let's say we run our logistic regression, and we find that beta 0 is minus 1, beta 1 is equal to 2, and beta 2 is equal to 3. Now, the left hand side of this second line here is our logit function.

What we really want is the probability of treatment conditional on covariates W. So we can transform the equation in the second line to the equation in the third line using something called the expit function. And that gives us the probability that a school received the program, so probability of A conditional on covariates for that school is equal to 1 over 1 plus e to the negative quantity minus 1 plus 2 times W1 plus 3 times W2. This is really powerful, because this formula allows us to take the data for each school in our data set and multiply the probability that it would have gotten the flu vaccine program conditional on the value for W1 and the value for W2 for that particular school.

Let's take a look at this on the next slide.

Propensity score matching to evaluate an existing intervention

- **Step 3:** Use the regression model fit to estimate the probability of being treated or exposed conditional on a set of characteristics for each school

$$\Pr(A|W = w) = \frac{1}{1 + e^{-(-1+2 \cdot W_1 + 3 \cdot W_2)}}$$

id	A	W_1	W_2	Pr(A|W=w) (Propensity scores)
1	1	1	0	1 / (1 + exp(-(-1 + 2*1 + 3*0))) = 0.29
2	0	0	1	1 / (1 + exp(-(-1 + 2*0 + 3*1))) = 0.88
3	1	1	1	1 / (1 + exp(-(-1 + 2*1 + 3*1))) = 0.89
4	0	0	0	1 / (1 + exp(-(-1 + 2*1 + 3*0))) = 0.27

14

Here is that same formula. And here's some hypothetical data. Each ID here is indicating a different school, so school 1, school 2, 3, and 4. A is whether or not the school actually got the program. W1 could be parent education, W2 could be some other binary covariate, such as whether or not the student is an English language learner.

So these are the actual values in the data set for each school, these three columns here-- A, W1 and W2. And then what we can do in this last column is we can calculate the probability that the school got the vaccine program conditional on the values of W1 and W2. And this has nothing to do with the actual value of A in the data. So for school number 1, let's go through this example.

School number 1 actually got the vaccine program. But when we estimate the probability that they got the program given their values of W1 and W2, this is how we plug in the formula. So it's the same as right up here at the top, except we plug in for W1 this value 1 here, because that's their value. And for W2, we plug in the value 0, because that's our value for that variable. And that gives us, when we do the math, 0.29. So that means there's a 29% probability that that school was treated or that school got a vaccine program conditional on their values of W1 and W2.

Propensity score matching to evaluate an existing intervention

- **Step 3:** Use the regression model fit to estimate the probability of being treated or exposed conditional on a set of characteristics for each school

$$\Pr(A|W = w) = \frac{1}{1 + e^{-(-1+2 \cdot W_1 + 3 \cdot W_2)}}$$

id	A	W_1	W_2	Pr(A|W=w) (Propensity scores)
1	1	1	0	1 / (1 + exp(-(-1 + 2*1 + 3*0))) = 0.29
2	0	0	1	1 / (1 + exp(-(-1 + 2*0 + 3*1))) = 0.88
3	1	1	1	1 / (1 + exp(-(-1 + 2*1 + 3*1))) = 0.89
4	0	0	0	1 / (1 + exp(-(-1 + 2*0 + 3*0))) = 0.27

Now, you may be thinking this is kind of confusing. The school actually got the program, but the probability that they got the program was only 29%. And that's because this is a really made up example, and we're only using two variables. So it's really not going to do a good job of predicting whether a school got the program when you only have two variables. You really need multiple variables for this to work well. Just keep that in mind.

As we continue down, what we see is that the values change. So we plug in the values of W1 and W2 for each school in the study data set. And what we get is different probabilities of that school receiving an influenza vaccine program.

Propensity score matching to evaluate an existing intervention

- **Step 3:** Use the regression model fit to estimate the probability of being treated or exposed conditional on a set of characteristics for each school

$$\Pr(A|W = w) = \frac{1}{1 + e^{-(-1+2 \cdot W_1 + 3 \cdot W_2)}}$$

id	A	W_1	W_2	Pr(A|W=w) (Propensity scores)
1	1	1	0	1 / (1 + exp(-(-1 + 2*1 + 3*0))) = 0.29
2	0	0	1	1 / (1 + exp(-(-1 + 2*0 + 3*1))) = 0.88
3	1	1	1	1 / (1 + exp(-(-1 + 2*1 + 3*1))) = 0.89
4	0	0	0	1 / (1 + exp(-(-1 + 2*0 + 3*0))) = 0.27

88% probability that id 2 was treated

89% probability that id 3 was treated

So just to further illustrate this, in row 2, we have an 88% probability that the school with ID 2 was treated, that they got the program. And in row 3, there's an 89% probability that school with ID number 3 was treated.

Propensity score matching to evaluate an existing intervention

- **Step 4:** Match each treated school with an untreated school with a similar probability of being treated conditional on covariates.

$$\Pr(A|W=w) = \frac{1}{1+e^{-(-1+2\cdot W_1+3\cdot W_2)}}$$

id	A	W_1	W_2	Pr(A|W=w) (Propensity scores)
1	1	1	0	1 / (1 + exp(-(-1 + 2*1 + 3*0))) = 0.29
2	0	0	1	1 / (1 + exp(-(-1 + 2*0 + 3*1))) = 0.88
3	1	1	1	1 / (1 + exp(-(-1 + 2*1 + 3*1))) = 0.89
4	0	0	0	1 / (1 + exp(-(-1 + 2*0 + 3*0))) = 0.27

Untreated id

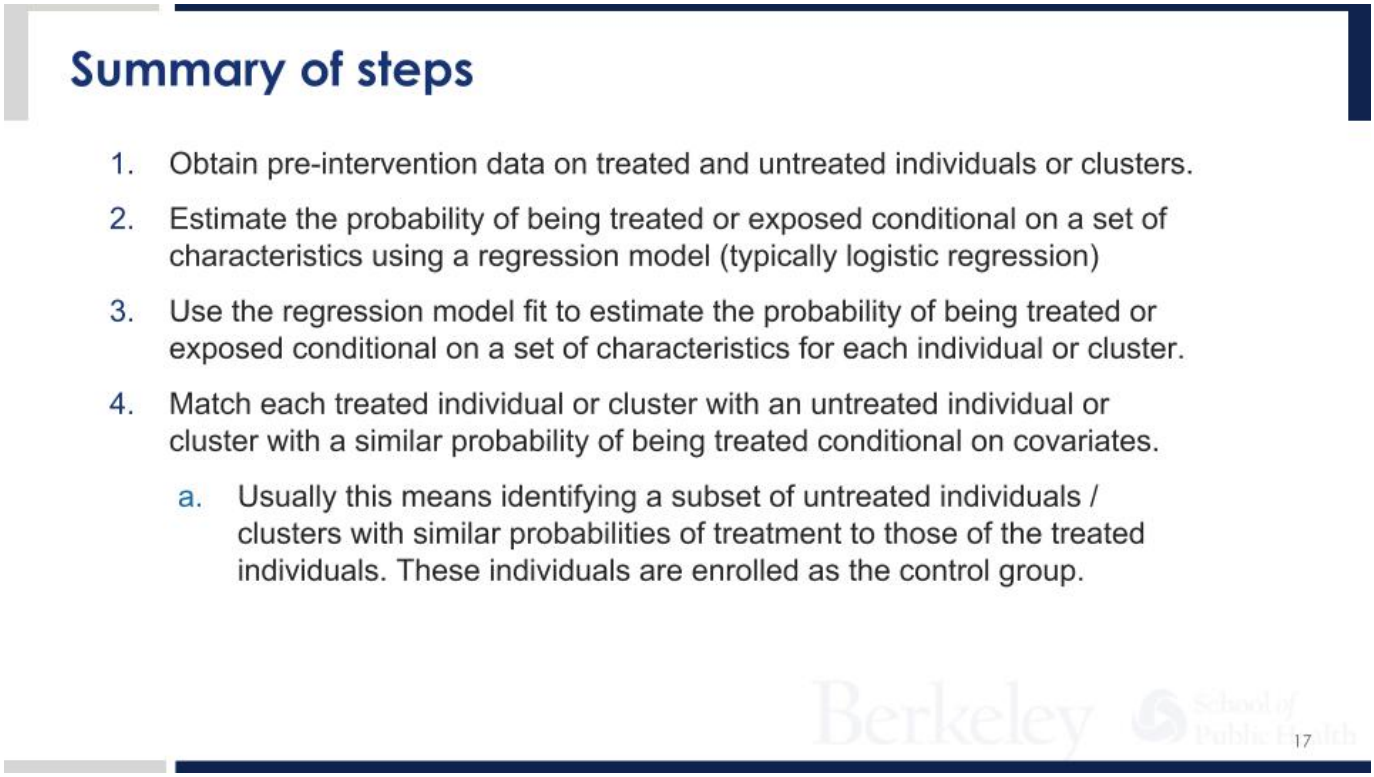
Treated id

Match ids with similar propensity scores

16

So in step 4, we're going to match each treated school with an untreated school that had a similar probability of being treated conditional on covariates. So this is really where the rubber hits the road. Let's look at IDs 2 and 3. So for school with ID number 2, they did not get the program. A is 0. For the school with ID number 3, they did get the program. A is 1.

Even though one got the program and one didn't, their propensity scores are very similar, 0.88 and 0.89. So these two schools would be matched with each other. This school with ID number 2 would serve as the control for the school with ID number 3. And we could do the same thing for IDs 1 and 4, because they have really similar propensity scores, as well. The propensity score for the school with ID 1 is 0.29, and the propensity score for the school with ID 4 is 0.27. So those two could be matched, and the control for school ID number 1 could be school ID number 4.



So to summarize, here are the steps we just went through. We obtained pre-intervention data on treated and untreated individuals or clusters. So our example was based on schools, but you could also do this for individual people or other clustered units. Then we estimated the probability of being treated or being exposed conditional on a set of characteristics. And this is typically done using a logistic regression model, but you could really use almost any type of model.

Then we used the regression model fit to estimate the probability of being treated or exposed conditional on these characteristics. Then we used the regression model fit to estimate the probability of being treated or exposed conditional on a set of characteristics for each individual or cluster. So this step, number 3, is where we get the propensity score for each individual cluster.

And then in step 4, we match each treated school with an untreated school, and then in step 4 we match each treated individual or cluster with an untreated individual or cluster that has a similar probability of being treated, a similar propensity score. So this usually means we're identifying a subset of untreated individuals. So we're not taking all the schools in our example that didn't get the program, the flu vaccine program, we're picking the schools, the subset of schools, that didn't get the program but that had a similar probability of treatment to those that did. And these individuals or clusters are then enrolled as the control group moving forward with your study, treating it as a cohort study.

Analysis and interpretation of propensity score matched data

- The analysis of a propensity score matched study must account for the matching.
- Interpretation: Estimate of the average association between treatment and the outcome for the treated population (if all A=1 matched to an A=0)
- **Statistical notation:**
  - RD: $E[Y|A=1, W] - E[Y|A=0, W]$
  - RD in propensity matched study: $E_w\{E[Y|A=1, W]|A=1\} - E_w\{E[Y|A=0, W]|A=1\}$
- **Counterfactual notation:**
  - RD: $E[Y_1] - E[Y_0]$
  - RD in propensity matched study: $E[Y_1|A=1] - E[Y_0|A=1]$

Let's briefly talk about the analysis and interpretation of propensity score matched data. So when we analyze data from a propensity score matched study, we do need to account for the matching. We're not going to get into details of that here, but that's just something for you to keep in mind. And then when we interpret the results of a propensity score matched study, we're estimating the average association between the treatment or the exposure and the outcome for the treated population. So if all A equals 1 match to an A equals 0.

OK. What does this mean? So let's first look at statistical notation. The regular old risk difference is listed first. So that's the expectation of the outcome Y conditional on A equals 1, controlling for confounders W minus the expectation of Y, conditional on A equals 0, controlling for confounders W. What does the risk difference in a propensity score matched study look like?

Well, we've added this additional conditioning here. So "conditional on A equals 1" is at the end of this term here and at the end of this term here. What this is essentially saying is, we're estimating the risk difference in the treated population. Now let's look at this in counterfactual notation. The regular risk difference is just the expectation of Y sub 1, so that's if everyone got the treatment, minus the expectation of Y sub 0 if no one got the treatment. Now, the counterfactual notation for a risk difference in a propensity score matched study is the mean counterfactual if everyone got the treatment, so Y sub 1, conditional on A equals 1, minus the expectation of the counterfactual of Y sub 0, conditional on A equals 1.

Analysis and interpretation of propensity score matched data

- The analysis of a propensity score matched study must account for the matching.
- Interpretation: Estimate of the average association between treatment and the outcome for the treated population (if all A=1 matched to an A=0)
- **Statistical notation:**
  - RD: $E[Y|A=1, W] - E[Y|A=0, W]$
  - RD in propensity matched study: $E_w\{E[Y|A=1, W]|A=1\} - E_w\{E[Y|A=0, W]|A=1\}$
- **Counterfactual notation:**
  - RD: $E[Y_1] - E[Y_0]$
  - RD in propensity matched study: $E[Y_1|A=1] - E[Y_0|A=1]$

So what is with this "conditional on A equals 1." Well, this is basically indicating that the subset that we've selected through our propensity score matched process is matched on the probability that treatment is 1, A equals 1. So we need to be really careful when we interpret results from propensity score matched studies, because essentially what we're doing is, we're identifying a group of individuals who didn't get the intervention but who would have based on their covariates.

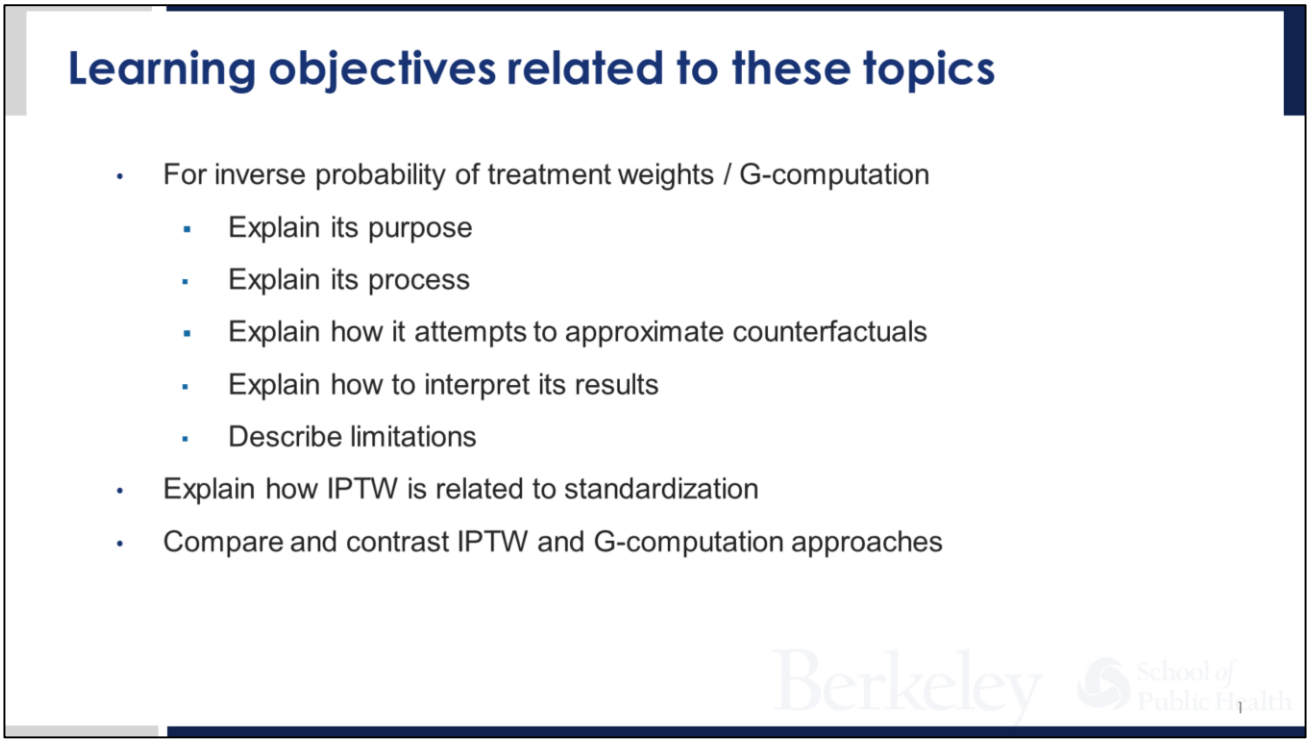
And so this isn't the same as if we took a random sample of individuals or groups who didn't get the intervention in the population. So the interpretation is a little bit more narrow than for a typical risk difference. And we could make the same claims for a risk ratio. I have just chosen the risk difference here for illustration purposes.

Summary of key points

- Propensity score: the predicted probability ("propensity") of exposure in a particular individual based on a set of characteristics
- Common uses:
  - Study with time-dependent confounding
  - You want to mimic randomization in an observational study (often one that evaluates an existing intervention).
- Analysis of propensity score matched data must account for that matching.
- The interpretation of measures of association is different in a propensity score matched study.

To summarize, the propensity score is the predicted probability or propensity of exposure in an individual or cluster based on a set of characteristics. And the common uses for propensity score matching are a study with time-dependent confounding, or you want to mimic a randomized trial using observational data. And this is commonly used to evaluate an existing intervention. Your analysis must account for the matching. We didn't go into detail about that, but it's important to keep in mind. And the interpretation of your measure of association using a propensity score matched analysis will be different than an analysis that does not match on propensity scores.

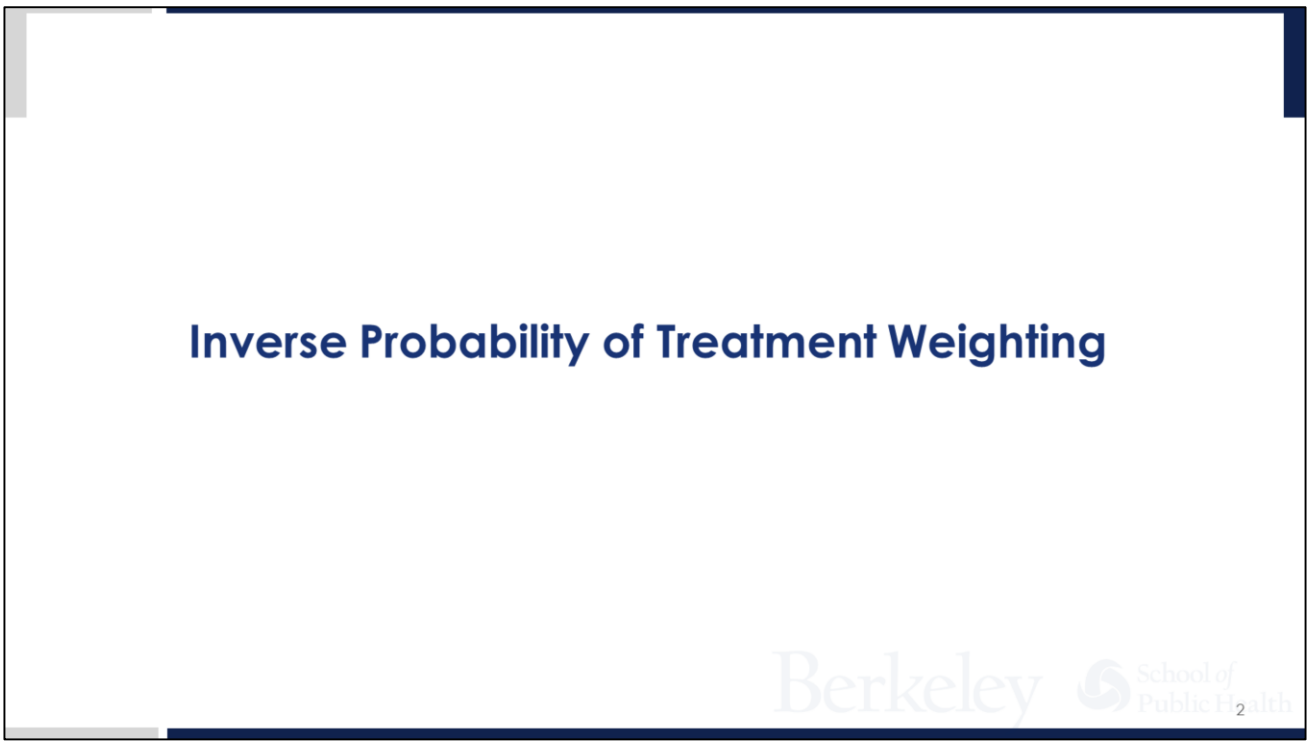
JADE BENJAMIN-CHUNG: In this video, we'll talk about inverse probability of treatment weighting and G-Computation.



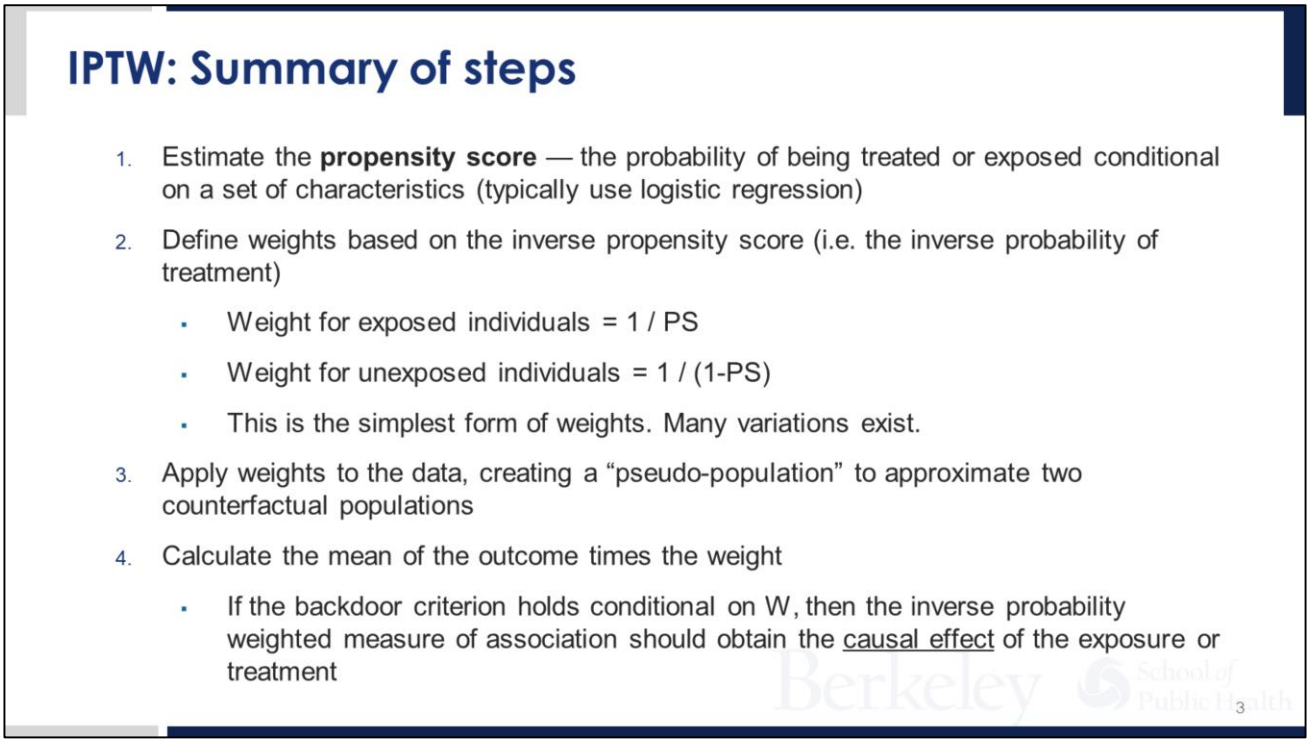
Before I get started, I'd like to quickly go over our learning objectives for these topics, because these topics are a little more advanced and get somewhat technical, and so it's important for you to just keep in mind, as you're watching this video, what we're really hoping for you to take away from this.

So for each of these methods, and I'll go through them one at a time, we'd like you to be able to explain its purpose, so why it's used, the process of implementing each of these methods, including the steps that are involved, how they're related to this idea of approximating counterfactuals in an observational study, how to interpret the results from each method, limitations of each method, how IPTW is related to standardization, and then we'd also like you to be able to compare and contrast these two different approaches.

As we discussed in the propensity score matching video, a motivation for these two methods is also time dependent confounding. So if you'd like to recap that topic, please re-watch the propensity score matching video. It's also a motivation for IPTW and G-Computation.



We'll start with inverse probability of treatment weighting.



This method builds on propensity scores, which you learned about in a previous video, and I'm going to talk you through the steps on this slide, and then we'll go through an example and apply each of these steps in that example to show you how this works.

In step one, we estimate the propensity score. That's the probability that an individual was treated or exposed conditional on a set of characteristics. And typically, we use logistic regression. For statistical reasons, this model often works best. We're not going to get into that level of detail in this course, but it's just good for you to know that logistic regression is commonly used to estimate the propensity score, although there are alternative methods.

In the second step, we're going to use the propensity score to create weights, and these weights are going to be based on the inverse of the propensity score, and that's why this method is called IPTW, inverse probability of treatment weighting. Because remember, that propensity score is the probability of treatment.

Among the exposed individuals, we'll assign them weights equal to 1 over the propensity score. So on this slide, PS stands for propensity score, and among unexposed individuals, we'll assign them a weight 1 over 1 minus the propensity score. This is the most simple way that we can do weighting in IPTW, and there's a lot of different variations, but we're just focusing on the simple example here.

IPTW: Summary of steps

1. Estimate the **propensity score** — the probability of being treated or exposed conditional on a set of characteristics (typically use logistic regression)
2. Define weights based on the inverse propensity score (i.e. the inverse probability of treatment)
   - Weight for exposed individuals = 1 / PS
   - Weight for unexposed individuals = 1 / (1-PS)
   - This is the simplest form of weights. Many variations exist.
3. Apply weights to the data, creating a "pseudo-population" to approximate two counterfactual populations
4. Calculate the mean of the outcome times the weight
   - If the backdoor criterion holds conditional on W, then the inverse probability weighted measure of association should obtain the causal effect of the exposure or treatment

In the third step, we'll apply these weights to the data. So for each person who's exposed, we're going to apply the weight 1 over PS, and for each person who's unexposed, will apply the weight 1 over 1 over PS. And that will create a pseudo population that approximates two counterfactuals. What do we mean by a pseudo population? Well, I think it'll become clearer when we look at this in two by two tables.

But basically, by weighting, we're going to try to approximate a randomized trial, because a randomized trial would help us get as close as possible to using a counterfactual to assess whether or not our treatment or exposure causes our outcome of interest. So we'll come back to that in a moment. And then in step four, we'll calculate the mean outcome times the weight.

And if the back door criterion holds conditional on W, which we're going to use to denote our confounders and covariates, then the inverse probability weighted measure of association should obtain a causal effect of the exposure or the treatment.

Example of IPTW

- **Example**: an observational study whether influenza vaccination reduces the risk of visiting a doctor's office for influenza-like symptoms

	Doctor visit	No doctor visit
Influenza vaccine	128	282
No influenza vaccine	64	26

Crude RR = 0.44

For example, let's focus on an observational study that's assessing whether influenza vaccination reduces the risk of visiting a doctor's office for influenza like symptoms, so cough, fever, runny nose, et cetera.

Now, on our two by two table here, the columns indicate the outcome, whether or not a person visited a doctor for these symptoms. The rows indicate the treatment status, whether a person received a vaccine or not. And if we calculate our crude, relative risk is equal to 0.44, which suggests that there's a reduction in the likelihood or the risk of visiting a doctor for influenza like symptoms among people who are vaccinated for influenza.

Example of IPTW

- **Example**: an observational study whether influenza vaccination reduces the risk of visiting a doctor's office for influenza-like symptoms

	Doctor visit	No doctor visit
Influenza vaccine	128	282
No influenza vaccine	64	26

Crude RR = 0.44

Confounding is present!

Age <65 years		
	Doctor visit	**No doctor visit**
Influenza vaccine	108	252
No influenza vaccine	24	16

RR if Age <65 years = 0.5

Age ≥ 65 years		
	Doctor visit	**No doctor visit**
Influenza vaccine	20	30
No influenza vaccine	40	10

RR if Age ≥ 65 years = 0.5

5

Now, let's stratify by a binary indicator of age under 65 versus greater than or equal to 65. And what we see is that the relative risk controlling for age in each of the strata is equal to 0.5. And so because the relative risks are different when we stratify by age, than the crude relative risk, but the two relative risks are equal to each other, we conclude that confounding is present, and also that effect modification is not present by age.

So in this very simple example, we decide that we need to do something to control for this confounding by age, and I'm going to walk you through the rest of this example, showing you how to use IPTW to control for age. Now, you may be wondering, why do we need to do this more advanced method if we could just use the Mantel Haenszel method to get a summary relative risk that controls for age? Well, you certainly can.

So the example here is just focusing on one confounder to keep things as simple as possible. We would always do this with more than one potential confounder in practice. We're never going to go through with all these steps if we just have a single confounder that we're worried about.

IPTW Step 1: estimate propensity score

	Influenza vaccine	No influenza vaccine
Age <65 years	360	40
Age ≥ 65 years	50	50

Step 1: Estimate the propensity score — the probability of being treated or exposed conditional on a set of characteristics

Propensity score if age < 65 years: 360 / (360 + 40) = 0.90
Propensity score if age ≥ 65 years: 50 / (50 + 50) = 0.50

Notice that the distribution of age is different within the exposed and unexposed groups.

Step one of IPTW is to estimate the propensity score. Now, the two by two table at the top of this slide is a little bit different. We've put the exposure in the columns, and then the confounder in the rows. So influenza vaccination status is in the columns. And then the age indicators in the rows. And in our first step, we need to estimate the probability of being treated or being vaccinated, in this case, conditional on age.

So to do this, we can follow the math on the slide. We'll calculate the propensity score if age is under 65. And to do so, we hone in on the first row of the table, and we take 360, which is the number of people under 65, who are vaccinated, and divide that by the total, people who are under 65, 360 plus 40, and that gives us 0.9. So we can interpret that as meaning that there's a 0.9 or 90% probability of vaccination among people under 65.

And then for people who are 65 or older, we can do the same thing, so we divide 50 by 50 plus 50 and that gives us 0.5. So the probability of vaccination among people over 65 is 50% or 0.5. Now, notice that the distribution of age is different within the exposed and unexposed groups.

In the exposed group, those are the people who had the influenza vaccine, there are quite a few more people under 65 than over. But then in the unvaccinated group, the number is quite similar between those who are under 65 and over 65. Think about what this would look like if we had done a trial where we randomized people to receive the influenza vaccine or not. In a trial, we would have balance across a series of different characteristics that could be potential confounders. And so in a trial, we would hope that if randomization was effective, the age distribution would be very similar across levels of vaccination status. So the next step that we're coming to where we weight the data is going to help us achieve that.

IPTW Step 2: Define weights

	Influenza vaccine	No influenza vaccine
Age <65 years	360	40
Age ≥ 65 years	50	50

Step 2: Define weights based on the inverse propensity score (i.e. the inverse probability of treatment)

Propensity score if age < 65 years: 360 / (360 + 40) = 0.90
Propensity score if age ≥ 65 years: 50 / (50 + 50) = 0.50

- **Weight for age <65 years**
  - Exposed = 1 / 0.90 = 1.11
  - Unexposed = 1 / (1-0.90) = 10
- **Weight for age ≥ 65 years**
  - Exposed = 1 / 0.50 = 2
  - Unexposed = 1 / (1-0.50) = 2

Notice that the distribution of age is different within the exposed and unexposed groups.

7

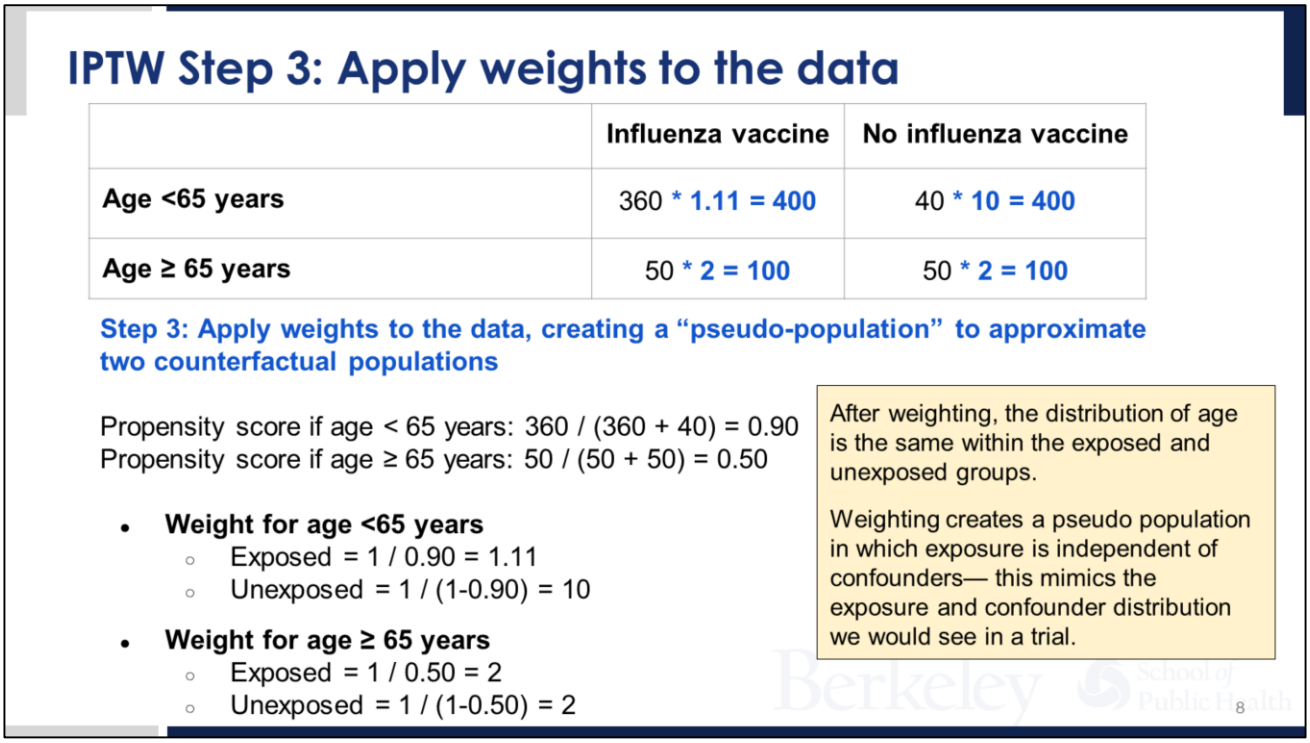
OK, step two. Define the weights based on the inverse propensity score. So that's where this name comes, from IPTW, inverse probability of treatment weighting. I've copied and pasted the propensity scores we estimated on the previous slide at the top here. We're going to use the simple weight definition that I mentioned earlier in the slide with all the steps for IPTW.

So to calculate the weight for people under age 65 among the exposed, we're going to say the weight is 1 over propensity score of 0.9, and that's equal to 1.11. Among the unexposed, the weight is equal to 1 over 1 minus 0.9, and that's equal to 10.

And we can do the same thing for the weights for people who are age 65 or older. So we just plug in the propensity score according to that same formula, and that gives us the weights for the exposed equal to 2 and the weight for the unexposed equal to 2.

Let's take a look at these weights for a moment. So most of the weights are similar in value, except the weight for people under 65, who are unexposed. Now, if we look at our two by two table at the top of the slide, let's look at the people who meet those criteria. So people who are under 65 are in the top row, and those who are unexposed are those without the influenza vaccine.

And so this really large weight of 10 here is essentially going to up weight this cell. When we up weight that cell by a factor of 10, that number 40 is going to become much closer to the 360 number in the adjacent cell. And that's going to help us get a more similar age distribution between our exposed and our unexposed groups, and as I was discussing on the previous slide, that's going to help us better mimic the distribution we would have gotten in a randomized trial.



Step three. Apply weights to the data. This will create a pseudo population to approximate two counterfactual populations, one that's exposed and one that's unexposed. So I've copied and pasted, again, our propensity scores and the weights in the bottom left hand side of the slide. And what we do then is let's start with the weight for people under 65 who are exposed.

So that weight is 1.11, and so we're going to multiply 360 times the weight 1.11, which gives us 400. For the people under 65 who were unexposed, that same row, people without the influenza vaccine will multiply the value 40 times 10, and that gives us a value 400. And then we can do the same for those who are over 65.

After the weighting, our new weighted two by two table has values of 400 and 100 in both the influenza vaccine and no influenza vaccine columns. And so in other words, the distribution of age is the same within the exposed and unexposed groups after we did the weighting. And this creates a pseudo population in which the exposure is independent of the confounder. This is mimicking how we would expect the exposure and confounder distribution to be in a trial.

IPTW Step 3: Apply weights to the data

Age <65 years (Weighted)		
	Doctor visit	No doctor visit
Influenza vaccine	108 * 1.11 = 120	252 * 1.11 = 280
No influenza vaccine	24 * 10 = 240	16 * 10 = 160

Age ≥ 65 years (Weighted)		
	Doctor visit	No doctor visit
Influenza vaccine	20 * 2 = 40	30 * 2 = 60
No influenza vaccine	40 * 2 = 80	10 * 2 = 20

Weight for age <65 years
- Exposed = 1 / 0.90 = 1.11
- Unexposed = 1 / (1-0.90) = 10

Weight for age ≥ 65 years
- Exposed = 1 / 0.50 = 2
- Unexposed = 1 / (1-0.50) = 2

9

Now, in the previous slide, we applied the weights to the two by two table with the exposure on the top and the confounder on the left hand side, and that was really just to illustrate how the weighting process ensures that the exposure and the confounder are independent from each other.

What we actually want to do, in practice, is to apply the weights to a regular style two by two table. So here's what it looks like in our stratified two by two tables. We can take the weight for people under 65, who are exposed, which is 1.11, and multiply that in the first row of the top two by two tables. So the top two by two tables for people under 65, in the first row for those with the influenza vaccine, those people were exposed. We multiply those cells times 1.11.

And then for those who are unexposed in the second row of the first two by two table, we multiply the values times 10 and then the weights are both equal to 2 for both exposure categories for age over 65. So we just multiply each cell times 2. So here's our weighted stratified two by two tables.

IPTW Step 4: Calculate the measure of association

Pooled, weighted data		
	Doctor visit	**No doctor visit**
Influenza vaccine	160	340
No influenza vaccine	320	180

Recall:

- Crude RR = 0.73
- RR if age < 65 = 0.5
- RR if age ≥ 65 = 0.5

Step 4: Calculate the mean of the outcome times the weight

IPTW RR = (160 / (160 + 340)) / (320 / (320 + 180)) = 0.5

The weighting in IPTW has achieved the same RR that we would obtain by calculating a Mantel-Haenszel RR controlling for age.

10

And then at the top of this slide, we've summed up the weighted values from the last slide to get our pooled weighted data.

And in step four, we can calculate our measure of association. So we're just going to do what we would always do to get our relative risk. But instead of using the original two by two table cells, we're using the weighted data. So we can take 160 divided by 160 plus 340. That's our risk among those who are vaccinated, divided by 320 over 320 plus 180. That's our risk among the unvaccinated, and that gives us our IPTW relative risk of 0.5.

Now, recall, the crude relative risk goes 0.44, and the stratified relative risks were 0.5. And since both stratified relative risks were 0.5, we know that if we had done the Mantel Haenszel approach to estimating an adjusted relative risk, adjusting for age, it would have equaled 0.5.

So what we can see in this slide is that the weighting we did through IPTW has achieved the same relative risk that we would obtain if we'd calculated the Mantel Haenszel RR controlling for age.0.5.

IPTW Step 4: Calculate the measure of association

Age <65 years (Weighted)		
	Doctor visit	No doctor visit
Influenza vaccine	120	280
No influenza vaccine	240	160

RR = 0.5

Age ≥ 65 years (Weighted)		
	Doctor visit	No doctor visit
Influenza vaccine	40	60
No influenza vaccine	80	20

RR = 0.5

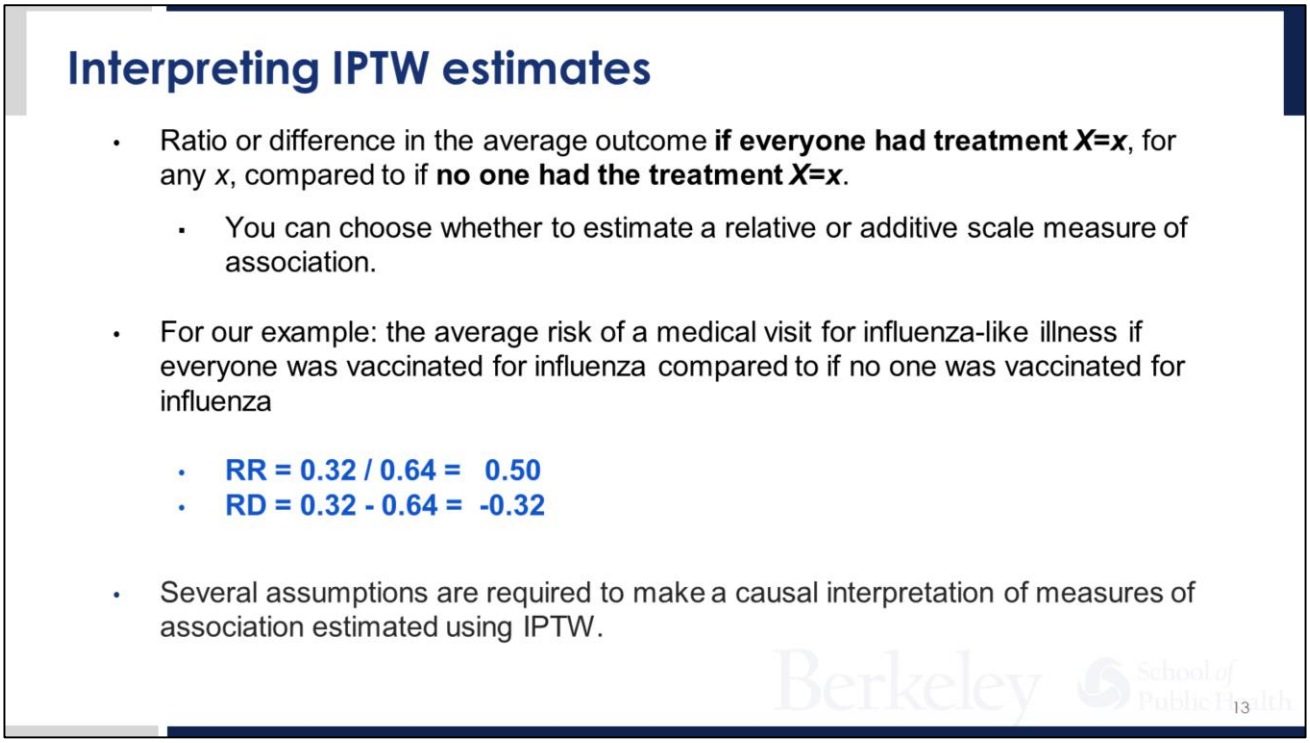
And to further show this, this is the RRs we get on the weighted data. So in the weighted data, stratifying by age, both relative risks are also equal to 0.5.

Creating a pseudo-population

- Weighting creates a pseudo-population.
  - If the weight for subject *i* = 4, then there are 4 copies of subject *i* in the pseudo-population.
- After weighting, the number of participants is balanced between exposed and unexposed within confounder strata.
  - This mimics what we would achieve in a randomized trial: covariate balance between the treatment and control group.

Now that we've gone through the steps of IPTW, let's briefly recap what we did when we created a pseudo population in this method. If the weight for subject i, for example, is equal to 4, the weighting that we did in IPTW creates four copies of subject i in the new pseudo population. And essentially, what this is doing is up weighting certain categories of the exposure and confounders in order to ensure that the number of participants is balanced between the exposed and unexposed within confounder strata. And this mimics what we would achieve in a randomized trial.

Recall when we learned about trials, we always want to check the balance table, which is often table one in a paper, to make sure that the values of a confounder are similar between the treatment and control group after randomization.



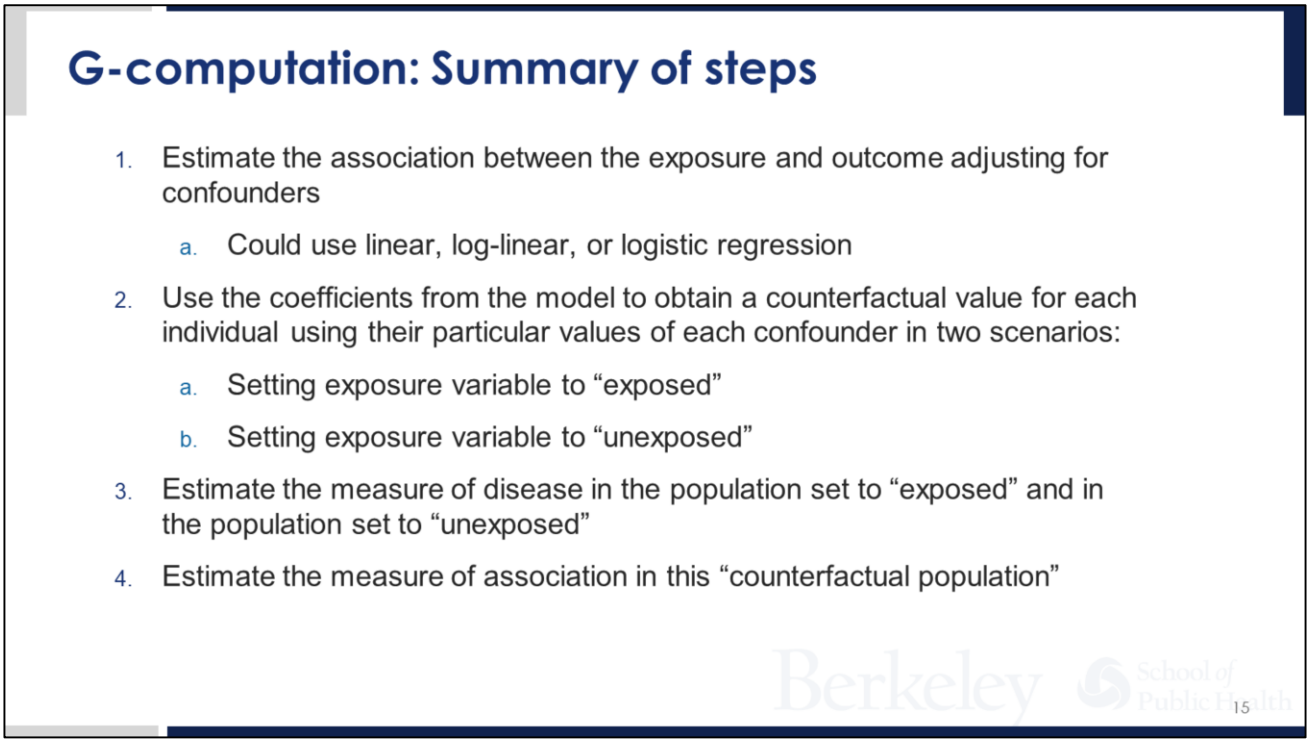
How do we interpret estimates from IPTW? So the IPTW estimate is the ratio or difference in the average outcome if everyone had the treatment for any possible level of treatment compared to if no one had the treatment. And it doesn't matter. You can choose whether you want to use a relative or additive scale measures, so both relative risk and risk difference can be estimated.

And in our example, we are looking at the average risk of a medical visit for influenza like illness if everyone was vaccinated for influenza, compared to if no one was vaccinated for influenza. And the relative risk is 0.5. And we didn't show this, but you could easily calculate the risk difference, which would have been minus 0.32.

So pause for a moment and think about what this interpretation is really saying. We're imagining a world in which we were able to give every person in our population the influenza vaccine and then compare that to a world in which the same people were not vaccinated for influenza. So think about whether that's actually a realistic counterfactual contrast. We're going to come back to this question of defining realistic counterfactuals in a few slides.

And then finally, if we want to make causal interpretations of measures of association, estimated using IPTW, there's a series of different assumptions that we have to make. This isn't really our focus here. There's an entire course at Berkeley on causal inference that details the set of assumptions we need to make to make causal inferences, whether they be about an IPTW results or a regression model results. So we're not going to go into great detail about that here.

All right. Now, let's move on to G-Computation.



And as I did for IPTW, I'll start by summarizing the steps of G-Computation, and then go through an example, applying each of these steps. So in G-Computation, we want to estimate the association between the exposure and the outcome adjusting for confounders. We could use different types of regression models to do this, including linear log linear logistic models.

Then we use the coefficients from the model to obtain a counterfactual value for each individual, using their particular values of each confounder in two different scenarios, one in which we set the exposure variable to exposed for everyone, regardless of their actual exposure value, and another where we set the exposure variable to unexposed, regardless of people's actual exposure value.

As we did in IPTW, in G-Computation, we're imagining a world where everyone was exposed, and then the same world where everyone was unexposed. And in step three, we estimate the measure of disease in the population when we set them to expose in the population, when we set them to unexposed. And then in step four, we estimate the measure of association in these counterfactual populations.

Example of G-computation

- **Example**: an observational study whether influenza vaccination reduces the risk of visiting a doctor's office for influenza-like symptoms
- **Step 1: Estimate the association between the exposure and outcome adjusting for confounders**
  - Y = doctor's office visit for influenza-like symptoms (binary: yes/no)
  - X = influenza vaccination status (binary: yes/no)
  - W = age
  - Could use log-linear regression in R:

    glm.fit = glm(Y~X+W, data = d, family=poisson(link="log"))

$$\ln(E(Y|X,W) = \beta_0 + \beta_1 X + \beta_2 W$$
$$\ln(E(Y|X,W) = -0.223 - 0.693X - 0.288W$$

16

Let's use the same example of an observational study, where we're looking at whether influenza vaccination reduces the risk of visiting a doctor's office for influenza like symptoms. In step one, we're going to estimate the association between the exposure and the outcome adjusting for confounders.

So if we define y as the outcome, doctor's visit, and it's a binary yes no, x is our exposure or treatment, influenza vaccination status as a binary yes no. W we define as a confounder. Age is going to be our focus, again, in this example, and we can also define that as a binary variable. We can use log linear regression in r to estimate this association. So here's the syntax we would use in r. It's just a regular glm model with Poisson family and a log link.

And here's an example of what this model might give us, so we have our beta 0s, our intercept, beta 1 is our coefficient on our x variable for influenza vaccination status, and w is our variable for age, and beta 2 is the coefficient for age, and so at the bottom of the slide here are the particular beta values that we could hypothetically get for this model.

Example of G-computation

- **Step 2: Use the coefficients from the model to obtain a counterfactual value for each individual using their particular values of each confounder in two scenarios:**

$$\ln(E(Y|X,W) = -0.223 - 0.693X - 0.288W$$

Subset of 4 rows:

id	Y	X	W	Counterfactual risk if X=1	Counterfactual risk if X=0
1	1	1	1	exp(-0.223 - 0.693(1) - 0.288(1)) = exp(-1.204) = 0.3	exp(-0.223 - 0.693(0) - 0.288(1)) = exp(-0.511) = 0.6
2	0	1	1	exp(-1.204) = 0.3	exp(-0.511) = 0.6
3	1	0	0	exp(-0.916) = 0.4	exp(-0.223) = 0.8
4	0	0	0	exp(-0.916) = 0.4	exp(-0.223) = 0.8

17

In step two, we can use the coefficients from the model in step one to obtain a counterfactual value for each individual using their particular values of each confounder in two scenarios. That's a lot of words. So let's kind of go through this step by step. Here's example data. So we have four individuals, IDs, 1, 2, 3, and 4. Each person is in a separate row, and then the columns y, x, and w indicate the observed values for these people.

So this is the data we collected in our study. Then what we're going to do is calculate the counterfactual risk if x is equal to 1, and we're going to use our model results to do this. So essentially, what we're going to say is let's estimate the probability of the outcome if this person's x variable was equal to 1. So we take minus 0.223, which is our beta 0, minus 0.693 times 1, because we're assigning x to 1, minus 0.288 times 1.

Now, we multiplied by 1 here because the observed value of w was 1 for this person. And then when we exponentiate this, because this is a log link function in this Poisson model, this gives us a value of 0.3. And then for the same person, we want to estimate the counterfactual risk if x is 0.

Example of G-computation

- **Step 2: Use the coefficients from the model to obtain a counterfactual value for each individual using their particular values of each confounder in two scenarios:**

$$\ln(E(Y|X,W) = -0.223 - 0.693X - 0.288W$$

Subset of 4 rows:

id	Y	X	W	Counterfactual risk if X=1	Counterfactual risk if X=0
1	1	1	1	exp(-0.223 - 0.693(1) - 0.288(1)) = exp(-1.204) = 0.3	exp(-0.223 - 0.693(0) - 0.288(1)) = exp(-0.511) = 0.6
2	0	1	1	exp(-1.204) = 0.3	exp(-0.511) = 0.6
3	1	0	0	exp(-0.916) = 0.4	exp(-0.223) = 0.8
4	0	0	0	exp(-0.916) = 0.4	exp(-0.223) = 0.8

17

So we're going to use the exact same formula, but then for x instead of assigning it to 1, we're going to assign it to 0. And when we exponentiate this sum, it gives us the value of 0.6. And so then we'd fill in the rest of the table accordingly, and each person would then get a different probability of the outcome or counterfactual risk under the value x is 1 and x is 0.

Now note that this column here, counterfactual risk if x is 1, is imagining this counterfactually, because, in fact, people had different observed values of x. So person 3 and 4 have values x of 0 in the study, but in this column, counterfactual risk if x is 1, we're imagining a situation where they actually had x equals 1 counter to fact.

Example of G-computation

- **Step 3: Estimate the measure of disease in the population set to "exposed" and in the population set to "unexposed"**

Subset of 4 rows:

id	Y	X	W	Counterfactual outcome if X=1	Counterfactual outcome if X=0
1	1	1	1	0.3	0.6
2	0	1	1	0.3	0.6
3	1	0	0	0.4	0.8
4	0	0	0	0.4	0.8
				...	...
Mean across all rows:				**0.32**	**0.64**

Counterfactual risk setting X=1 for everyone in the population = 0.32

Counterfactual risk setting X=0 for everyone in the population = 0.64

In step three, we estimate the measure of disease in the population set to expose and in the population set to unexposed. So we've copied the counterfactual values from the previous slide here in these columns, and now, what we want to do is take the mean across all these rows.

So we're just showing you the first four rows in detail, but there's additional rows in this population, because we have more than four people, if you remember from our two by two table example before. So we take the mean of this column here. It's the counterfactual risk setting x as 1 for everyone in the population. And that's equal to 0.32.

And then when we take the mean of the furthest column on the right hand side of the table, this is the counterfactual risk, setting x is 0 for everyone in the population, and that gives us the value of 0.64.

Example of G-computation

- **Step 4: Estimate the measure of association in this "counterfactual population"**

Subset of 4 rows:

id	Y	X	W	Counterfactual outcome if X=1	Counterfactual outcome if X=0
1	1	1	1	0.3	0.6
2	0	1	1	0.3	0.6
3	1	0	0	0.4	0.8
4	0	0	0	0.4	0.8
				...	...
Mean across all rows:				**0.32**	**0.64**

Counterfactual risk setting X=1 for everyone in the population = 0.32

Counterfactual risk setting X=0 for everyone in the population = 0.64

RR = 0.32 / 0.64 = 0.50
RD = 0.32 - 0.64 = -0.32

We obtained the same RR using IPTW and G-computation as we would have obtained by calculating a Mantel-Haenszel RR controlling for age.

19

Now, in step four, we'll estimate the measure of association in these counterfactual populations.

We can take these means and just divide them, because 0.32 is the mean outcome if everyone in the population had been exposed, and 0.64 is the mean outcome if no one had been exposed. And so we can divide 0.32 by 0.64 to get a relative risk of 0.5, or subtract these to get a risk difference of minutes 0.32.

So notice that we obtained the same relative risk using IPTW and G-Computation as we would have obtained by calculating a Mantel Haenszel RR controlling for age.

Interpreting G-computation estimates

- Ratio or difference in the average outcome **if everyone had treatment *X*=*x***, for any *x*, compared to if **no one had the treatment *X*=*x***.
- For our example: the average risk of a medical visit for influenza-like illness if everyone was vaccinated for influenza compared to if no one was vaccinated for influenza
  - **RR = 0.32 / 0.64 = 0.50**
  - **RD = 0.32 - 0.64 = -0.32**
- Several assumptions are required to make a causal interpretation of measures of association estimated using G-computation.

How can we interpret G-Computation estimates? It's the ratio or difference in the average outcome if everyone had treatment for any level of treatment compared to if no one had treatment. So notice that this is the same interpretation that we used for IPTW. And we have the same RR and the same RD estimates.

And as with IPTW, there's additional assumptions that we need to make if we want to make a causal interpretation of a measure of association estimated using G-Computation. We don't expect you to know those for this class, but just keep in mind that doing this procedure alone is not enough to allow us to make causal inferences.

Flexible parameter definition under IPTW and G-computation

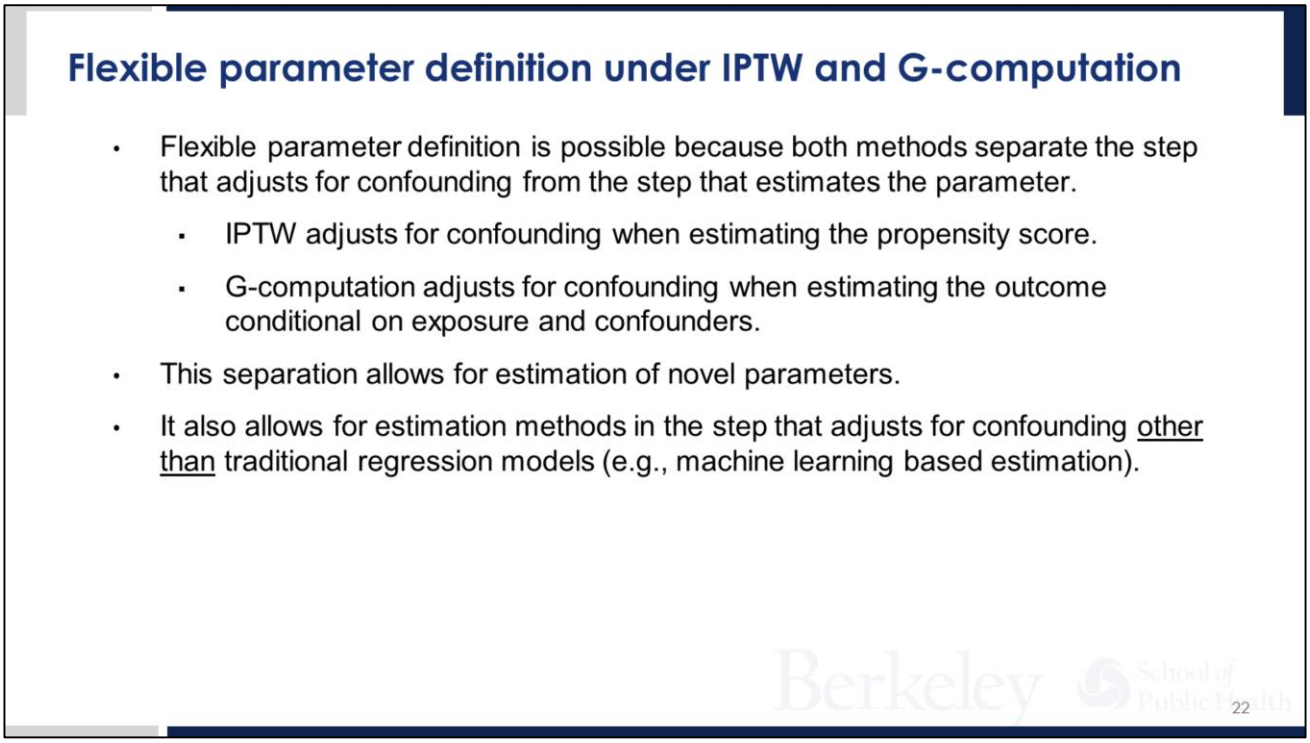
- Typically IPTW and G-computation are used to compare the risk when everyone is exposed compared to when no one is exposed.
- However, sometimes this is not a realistic contrast.
  - E.g., it would be highly unlikely for no one to receive the influenza vaccine in most populations in the U.S.
- IPTW and G-computation can be used to estimate other counterfactual contrasts
- Alternative parameter:
  - What is the difference or ratio of risk comparing:
    - Scenario with current level of vaccination coverage (50%)
    - Scenario with 80% vaccination coverage
- More on this in the unit on population intervention models

A really nice advantage of using these two methods, IPTW and G-Computation is that they allow us to estimate parameters with more flexible definitions. So when I talked through the interpretation of measures of association under these two methods, I explained that these methods typically compare the risk when everyone is exposed to when no one is exposed.

But I sort of hinted that this isn't always a realistic contrast, and in our example of influenza vaccination, it's really unlikely we would ever observe a scenario where no one was receiving the influenza vaccine. A really nice feature of IPTW and G-Computation is that it can be used to estimate other counterfactual contrasts.

So let's think about some alternative parameters for our example. We might be interested in comparing a scenario with the current level of vaccination coverage. So often, in reality, about half of people receive the influenza vaccine each year, versus a scenario in which 80% of people received the influenza vaccine.

And the reason that this is more useful is that this is actually comparing our current level to a level that's realistic, 80%, as opposed to 100%, which we are unlikely to ever attain, unfortunately. And so we'll come back to this in our unit on population intervention models.



Why are we able to estimate more flexible parameters? Well, it's because both these methods separate the step that adjusts for confounding from the step that estimates the parameter. So in IPTW, we adjust for confounding when we estimate the propensity score, and in G-Computation, we adjust for confounding when we estimate the outcome conditional on the exposure and confounders.

So because we're separating these steps, we can estimate these novel parameters, such as population intervention models, which we'll come back to. And it also allows for estimation methods in the step that adjusts for confounding, other than traditional regression models.

So I've been using examples, such as logistic regression and Poisson or log linear regression in this video, but we could also use machine learning based estimation to obtain the propensity score or the mean of the outcome in IPTW and G-Computation, and again, that's beyond the scope of this class, but there's a lot of advantages to using machine learning approaches. And so these methods provide that flexibility as well.

How do we obtain confidence intervals for IPTW and G-computation?

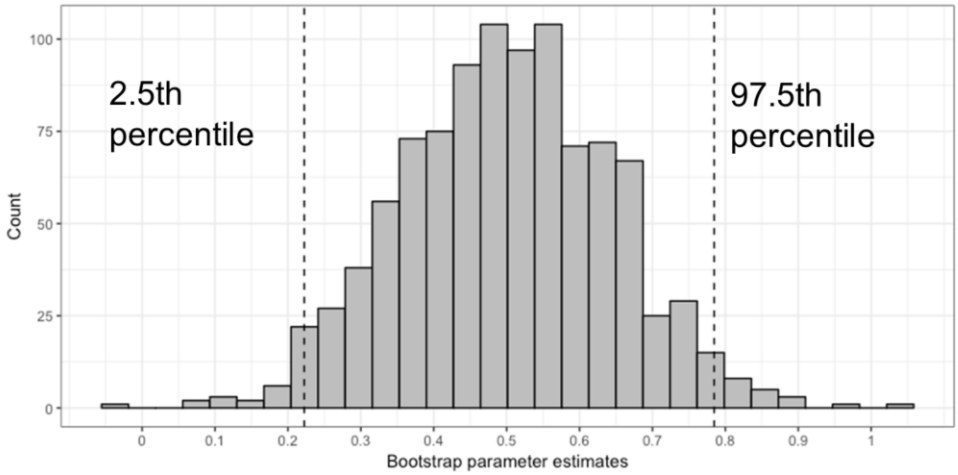
- There is no formula for calculating standard errors of parameter estimates from IPTW or G-computation.
- We must use bootstrapping to obtain standard errors, confidence intervals, and p-values.
- What is bootstrapping?
  1. Take a random sample with replacement from your dataset.
  2. Estimate the parameter in the random sample.
  3. Repeat steps 1-2 many times (1000 - 10,000 times) to obtain a distribution of bootstrapped parameter estimates.
  4. Obtain the standard error by applying the normal distribution to the parameter estimates or using quantiles of the distribution of the parameter estimates.

Something important to mention is that we can't directly calculate standard errors for our parameter estimates from IPTW or G-Computation, because there's no formulas for those. And so we need to use bootstrapping to obtain our standard errors confidence intervals and p values. What is that?

Well, bootstrapping has the following steps. We take a random sample with replacement from our data set. So let's say we have a data set of 1,000 people. In the first step, we can take a random sample of, say 500 from the data set. And then we estimate our parameter of interest in this random sample of 500.

Then we repeat steps one and two many times, 1,000, 10,000 times. So if we repeat it 1,000 times, we get 1,000 parameter estimates in these random samples from our data set. And so this is a distribution of bootstrapped parameter estimates, and then we can obtain our standard error by applying the normal distribution to the parameter estimates or using quantiles of the distribution.

Example of bootstrap distribution

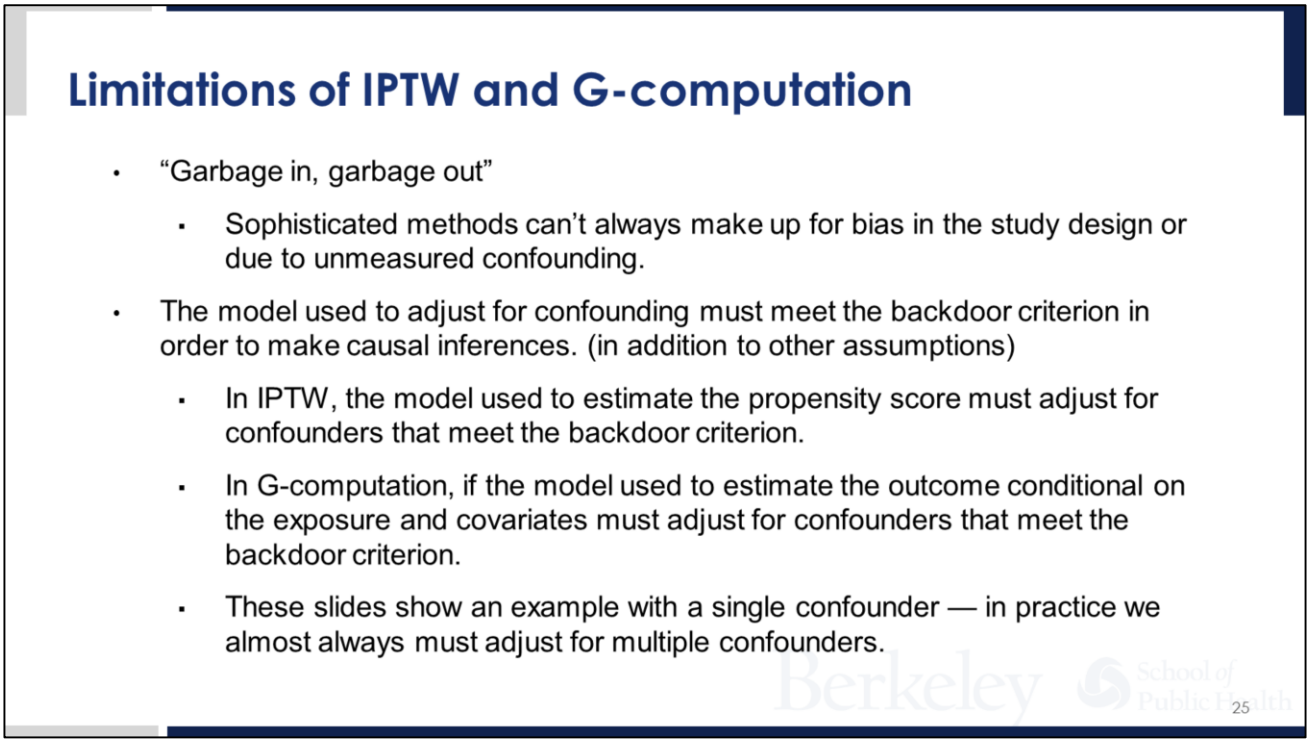


Lower bound = 0.22

Upper bound = 0.78

So on the next slide let's pretend this is our distribution of bootstrapped parameter estimates. So in this example, most of our parameter estimates fall around 0.5. But then there's a few where we have a parameter estimate closer to 1 and closer to 0. And since this distribution is normal, we can take the value of the distribution at the 2.5 percentile and at this 97.5 percentile, and those are equal to 0.22 and 0.78 respectively, and those can become our confidence interval bounds.

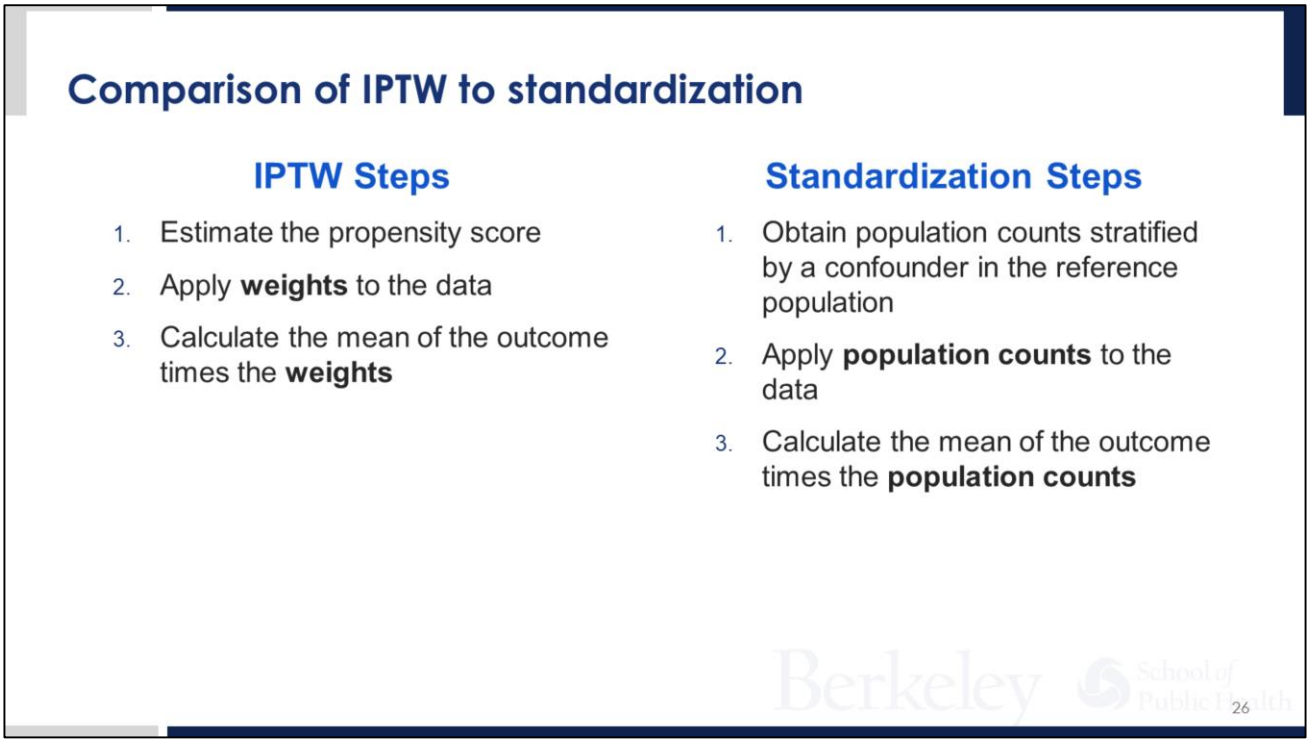
So this is pretty technical, and we really don't expect you to be able to explain bootstrapping in detail in this class. We just wanted to briefly expose you to it, and we want to make sure that you know that this is a method you can use to estimate confidence interval bounds for these methods.



There are limitations of IPTW and G-Computation. A common saying we always hear is garbage in, garbage out. So if we have bias in our study design or unmeasured confounding, we can't really make up for this with just these methods alone. The model we use to adjust for confounding must meet the backdoor criterion, so we learned about that when we learned about data and adjusting for confounding, and that is required for us to make causal inferences in addition to other assumptions that we need to make, that we're not going to go into detail about here.

So in IPTW the model we used to estimate the propensity score must adjust for confounders that meet the backdoor criterion in G-Computation, the model we used to estimate the outcome conditional of the exposure and covariates must adjust for confounders to meet this criterion. And we've only shown you an example with a single confounder in this video, but in practice we almost always mis-address for multiple confounders.

So using these methods on their own is not enough to make valid causal inferences from our data. There's other things that we must do as well.



Now we learned about standardization early in this course, and I want to briefly connect these concepts to standardization, because there are some parallels here when we're thinking about adjusting for a single confounder. So hear are the steps of IPTW on the left and the steps of standardization on the right.

And in IPTW, we estimate the propensity score. We apply weights based on the propensity score to the data, and then we calculate the mean of the outcome times the weights, and in standardization, we obtain population counts stratified by a confounder, often age, in a reference population. We apply those population counts to the data, and then we calculate the mean of the outcome times the population counts.

So if we look at step two in each method, the population counts we use in standardization can be thought of as weights. And so that is a corollary between IPTW and standardization, but the reason IPTW is much more powerful is that it allows us to do this with a multiple confounders at the same time, whereas standardization is typically done with a few variables or one variable.

Double robust estimation

$$P_O(Y, X, W) = \underbrace{P(Y|X, W)P(W)}_{\text{Targeted by G-computation}}\underbrace{P(X|W)}_{\text{Targeted by IPTW}}$$

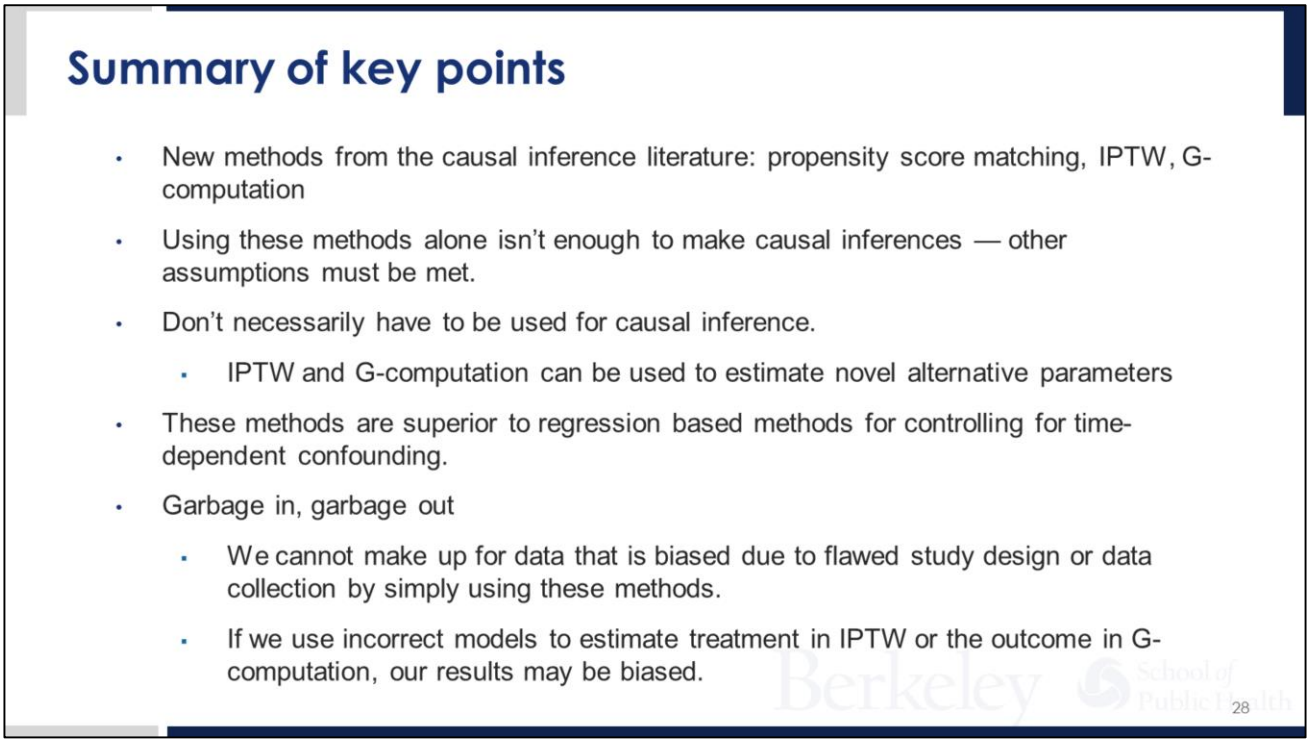
- To obtain an unbiased measure of association, we must use the correct model for the outcome under G-computation and the correct model for the treatment under IPTW.
- What if we get one of them wrong?
- Double robust estimation provides an unbiased estimate if either model is correct.
  - Example: Targeted maximum likelihood estimation

I'd like to briefly touch on double robust estimation, and there could be an entire course on this. So this is just a very high level introduction to this method. So to obtain an unbiased measure of association, we need to use the correct model for the outcome under G-Computation and the correct model for the treatment under IPTW.

But what if we get that model wrong? Well, then we'll have a biased estimate. So there are methods that fall under the category of double robust estimation that can provide unbiased estimates if either model is correct. So take a look at the formula at the top of the slide.

The joint probability of y, x, and w can be written as it is on the right hand side of the equation, where we have probability of y, conditional on x and w times the probability of w times the probability of x conditional on w. And notice that the portions of this equation correspond to what is targeted by G-Computation in the estimation step, and what is targeted by IPTW in the propensity score estimation step.

And so what double robust estimation is doing is essentially combining these two methods, so that as long as you get one of them right, you get the right answer. And so that's a very high level introduction, an example of one of these specific methods is targeted maximum likelihood estimation. So we won't go into further detail than that. But it's good for you to have heard this phrase in case it comes up in your future work.



To summarize, these new methods from the causal inference literature include propensity score matching IPTW and G-Computation. And using these methods alone is not enough to make causal inferences. There's other assumptions that we need to make. We don't necessarily have to use these methods to make causal inferences. We can use them to estimate novel parameters without necessarily having the goal of making a causal inference.

For a situation where we have time dependent confounding, these methods are superior to the traditional multi-variable regression based methods. Even so, we have this phrase we use, garbage in, garbage out to indicate that we can't make up for data that's biased due to a flawed study design or data collection simply by using propensity score matching, IPTW or G-Computation.

And if we use the incorrect model to estimate the treatment in IPTW or the outcome in G-Computation, our results may be biased. So by no means, these are not cure alls for our challenges we face in controlling for confounding, but in the particular case of time dependent confounding, these approaches are a particularly useful and an improvement upon traditional methods.

Sample size & power

This lecture was created by Dr. Ben Arnold @ UC Berkeley

Current use: PH250G

1

PRESENTER: I'd like to talk now about sample size and power. These are very fundamental concepts that are important for epidemiologists to understand as they read studies published by others and as they design studies that they are going to lead themselves to ensure that they're of the right size, not too big, not too small. And I'm drawing from a lecture here prepared by Dr. Ben Arnold in our epidemiology group at UC Berkeley.

Goals for this session

1. Familiarize you with the basic issues that underlie sample size calculations
2. Walk through the steps of real calculations, including one with clustering
3. Point out some additional wrinkles that you may encounter

Slide created by Ben Arnold 2

So the goals that I have for this session is to help you familiarize yourself with the basic issues that underlie sample size calculations, walk you through the steps of real calculations, and then in a subsequent lecture, including one that deals with clustered data, and then talk about some additional wrinkles that you may encounter and then learn more about it in more advanced courses.

Presentation overview

- Big-picture concepts & motivation
- Example 1: Understanding power & marine water exposure / gastrointestinal illness case study
- Example 2: Calculate sample size for an individually randomized trial
- Additional issues and recap

3

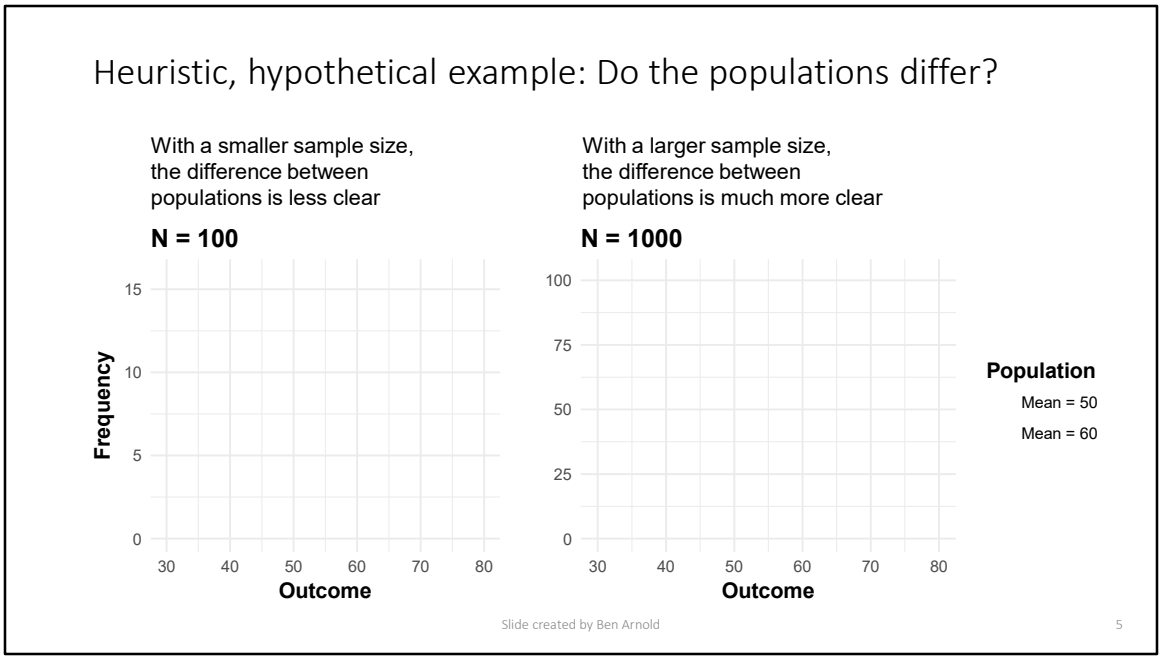
This talk is organized first around some big picture concepts and motivation for sample size and power. Then I'll use a case study from our group, a published article to help explain sample size and power in a setting of marine water exposure and gastrointestinal illness. And then we'll calculate sample size for an individually randomized trial. Circle back to cover some additional issues and recap the main points.

Motivation: need a sufficient sample size

- In the planning stage of a study, a key activity for epidemiologists and biostatisticians is to determine how many participants to enroll.
- A study needs to be large enough to distinguish a true difference from random variation

Slide created by Ben Arnold 4

What's the motivation? Why do we need a sufficient sample size? Well, in the planning stage of a study, a key activity for epidemiologists and statisticians is to determine how many participants need to be enrolled. And a study needs to be large enough to distinguish a true difference from just random variation. But on the other hand, you don't want to make your study too large, because then you can't afford to conduct it.



Let's look at a teaching example. In the data below, the question is being asked, do these two populations differ? So we have a population on the left and a population on the right. The population on the left has as a true mean a value of 50 for whatever the unit of measurement is for this example. And we've measured 100 subjects in each of these two populations and plotted them.

The population on the right has a true mean of 60, also has 100 subjects in the population. And with these relatively small sample sizes of 100 each, it can be difficult to determine whether or not these two populations are truly different from each other, because they overlap each other.

But let's look at what happens if we do a much larger measurement of the two populations. So on the right, we now have 1,000 subjects. And what we see is that the population on the left still has a true mean of 50, of course. And the population on the right still has a true mean of 60.

But it's much easier for us both visually and statistically to separate these two populations, because the larger sample size has given a much clearer vision or view of what is going on with the estimate of the mean of the two populations. So one way to help separate out two groups and distinguish them with respect to their mean value is to greatly increase the sample size.

Motivation: …but samples should not be too big

- Samples that are too large are a problem:
  - Waste time and resources
  - Impose undue burden on the study population
  - At some point study logistics begin to threaten internal validity

Slide created by Ben Arnold 6

But as I mentioned, it can be a problem when samples are too big. If they're too large, they waste time and resources. They impose undue burdens on the study population. And at some point, the logistics of a large study or an over large study begin to threaten the internal validity, because it's hard to conduct and carry out the study.

Sample Size and Minimum Detectable Effects

- When designing a study, we're typically interested in one of these two related questions:

1. Given a desired minimum detectable effect (MDE), how many participants does the study need to enroll? (Example 1)

2. Given a fixed sample size and design, what is the MDE that the study can detect? (Example 2)

- Tied up in both of these questions is a decision of how much power we want the design to have

So when designing a study, we're typically interested in one of two related questions. The first is, given a desired minimum detectable effect, and that's a thing, MDE, we want to use as a frequent concept, so given a minimum detectable effect, how many participants does the study need to Enroll that'll be our first example.

And then our second example will be, in the situation where we already have a fixed sample size, and let's assume a fixed design, what's the NBC MDE that the study can detect? And our fixed sample size might come about because we have a fixed budget.

So we have a sample size that has to be a certain size. And then we're asking the question, what's the minimum detectable effect that we can measure given that the study is that sample size? So lurking in both of these questions is a decision of how much power we want the study design to have.

Presentation overview

- Big-picture concepts & motivation
- **Example 1: Understanding power & marine water exposure / gastrointestinal illness case study**
- Example 2: Calculate sample size for an individually randomized trial
- Additional issues and recap

8

So I'm going to move on now and talk about a specific example to try to understand these concepts. And this is going to be the first situation is trying to figure out the power of a study. And for this, we're going to use a study in which the exposure was going into marine water, ocean water, and the outcome was gastrointestinal illness.

Example 1:
Does swimming in marine water increase the risk of gastrointestinal illness?

WATER RESEARCH 46 (2012) 2176–2186

Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/watres

Using rapid indicators for *Enterococcus* to assess the risk of illness after exposure to urban runoff contaminated marine water

John M. Colford Jr.[a,*], *Kenneth C. Schiff*[b], *John F. Griffith*[b], *Vince Yau*[a,1], *Benjamin F. Arnold*[a], *Catherine C. Wright*[a], *Joshua S. Gruber*[a], *Timothy J. Wade*[c], *Susan Burns*[d], *Jacqueline Hayes*[d], *Charles McGee*[e], *Mark Gold*[f], *Yiping Cao*[b], *Rachel T. Noble*[g], *Richard Haugland*[h], *Stephen B. Weisberg*[b]

Slide created by Ben Arnold 9

So here's the title page from this paper in our group that Ben and I both worked on. And the question we're asking is, does swimming in marine water increase the risk of gastrointestinal illness?

General setup using the recreational water example

- Scientific Question: Does swimming in marine water increase the risk of gastrointestinal illness?
- Swimmers ($A = 1$) and non-swimmers ($A = 0$) are enrolled at the beach, and followed for 2 weeks for incident GI illness (Y).
- One comparison of interest, θ, is the risk difference between groups:

$$\theta = \text{risk among swimmers} - \text{risk among non-swimmers}$$
$$= p_2 - p_1 = Pr(Y=1 \mid A=1) - Pr(Y=1 \mid A=0)$$

- Sample size question: How large does the study need to be to credibly detect an increase of 1 percentage point in the incidence proportion from a base of 3.4% to 4.4% ?

Slide created by Ben Arnold 10

So our scientific question, as I said, is, does swimming in marine water increase the risk of gastrointestinal illness.

Let's identify the exposures and outcomes in this study, so the swimmers are going to be referenced by the random variable a. And A can take on the value one or zero. So A equals one is going to refer to a swimmer. A equals zero will be a non swimmer. And these swimmers and non swimmers are enrolled at the beach.

And then they're followed for two weeks, during which we ask them whether or not they develop new or incident gastrointestinal illness. And that'll be their random variable y. So y one would be someone with illness. Y zero will be someone without illness.

So there are many different questions we could ask. But let's formulate a specific hypothesis that we want to test. And let's refer to this measure here as theta, as the risk difference between two groups.

General setup using the recreational water example

- Scientific Question: Does swimming in marine water increase the risk of gastrointestinal illness?
- Swimmers ($A = 1$) and non-swimmers ($A = 0$) are enrolled at the beach, and followed for 2 weeks for incident GI illness (Y).
- One comparison of interest, θ, is the risk difference between groups:

$$\theta = \text{risk among swimmers} - \text{risk among non-swimmers}$$
$$= p_2 - p_1 = Pr(Y=1 \mid A=1) - Pr(Y=1 \mid A=0)$$

- Sample size question: How large does the study need to be to credibly detect an increase of 1 percentage point in the incidence proportion from a base of 3.4% to 4.4% ?

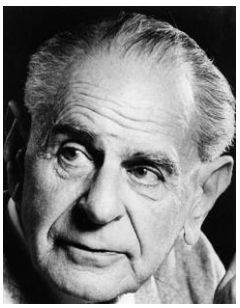
So we want to see if the risk in the swimming group is larger than the risk in the non swimming group. We could have done a risk ratio and set it up that way. But for the purposes of teaching sample size and power, let's first talk about a risk difference, because it's a concept that's a little easier to grasp right at the beginning.

So we're going to define theta as the difference between the probability in one group minus the probability in another group. We're calling them group P2 and P1 here. The first probability is when disease is present, y equals 1, given that exposure is present.

So these are the swimmers. So that first probability term is the probability of illness in the swimmers. And to get the risk difference, we're going to subtract the probability of illness in the non swimmers. That's the second term, the probability of y equals 1 given that A equals zero.

So our sample size question is, how large does our study need to be to credibly detect an increase of one percentage point in the incidence proportion from a base of 3.4% to 4.4%. So the idea here is we assume without any exposure that the background rate of illness is 3.4%. And this comes from outside information. But we want to be able to detect if it increases to 4.4% or greater, because if it's greater than 4.4, it's even easier to detect, so the minimum detectable effect, our MDE here is 1%, or going from 3.4 to 4.4 in this particular example.

Statistical Power



Sir Karl Popper (1902-1994)

https://en.wikipedia.org/wiki/Karl_Popper

- It is common to use statistical power to help determine the size of a study.
- Statistical power is defined as the probability of rejecting the null hypothesis when it is false.
- Implicit in its definition is that an investigator must specify a null and alternative hypothesis.
  - i.e., power is defined in terms of a specific hypothesis test
- Grounded in Popper's epistemology of falsification: scientific knowledge progresses through testing falsifiable statements.

Slide created by Ben Arnold 11

It's common to use statistical power to help determine the size of a study, and that's what we're doing here. If we want to detect this minimum detectable effect of 1% from 3.4 to 4.4, how large do we need to study to be? And statistical power is defined as the probability of rejecting the null hypothesis when it is false.

So we're going to set up a null hypothesis of the two groups being equal and see if we can reject that. It's implicit in this definition that the investigator specifies a null and alternative hypothesis. In other words, the particular power found in a study is defined in terms of a specific hypothesis test. If you test a different hypothesis, you're going to have a different power result. So the power is linked to what the specific hypothesis is that's being tested.

Now this idea of setting up a null hypothesis and rejecting it has a deep scientific background that many of you are probably familiar with, grounded in the work of Karl Popper and the epistemology of falsification. And that's the idea that scientific knowledge progresses through testing falsifiable statements.

So if we set up a statement, and we can reject it, then we only have to reject it once, because it's provably false. If we were to do this in a reverse way and set up statements that were true, showing that it was true in one case and another case and another case, would not prove that it's true all the time. Whereas rejecting it, falsifying it, once is all we need to do to prove that it's false. It only has to be false once.

1 : specify the null and alternative hypotheses

From the recreational water example:

- The null hypothesis, H_0, is that $\theta_0 = p_2 - p_1 = 0$
  (there is no effect)
- An alternative hypothesis, H_a , is that $\theta_a = p_2 - p_1 = 0.044 - 0.034 = 0.01$
  (water exposure increases risk by 1 percentage point)

Think of these scenarios as true alternatives – both cannot be true at the same time.

OK, so our first step is to specify the null and alternative hypotheses. And in our recreational water example, our null hypothesis abbreviated as H sub 0 is that theta zero, the difference is equal to P2 minus P1. And the null hypothesis here would be that the two groups don't differ. So the null hypothesis would be that this difference is zero. In other words, there's no effect.

An alternative hypothesis abbreviated as H sub A would be that the theta sub a is equal to P2 minus P1, where P2 is given as 0.044 minus P! as 0.34, with a difference of point 0.01. And that comes from the 4.4% and the 3.4% that we developed on the earlier slide.

3.4% was our known background rate of illness in the non swimmers. We wanted to see how big a study do we need to be able to detect a difference up to 4.4. So this is our alternative hypothesis that the exposure increases illness by 1% or more.

So think of these scenarios as true alternatives. By scenarios, I mean the null hypothesis, H0, and the alternative hypothesis, H1. It's not possible for both of them to be true at the same time.

2: pick a test statistic

- Given a hypothesis, the next step is to choose a test statistic
- One example of a (commonly used) test statistic is a Wald statistic, which is the difference between groups divided by its standard error.
- A Wald statistic, X, will have a standard normal distribution under the null hypothesis with mean 0 and standard deviation 1.

$$X = \frac{\hat{\theta}_0}{SE(\hat{\theta}_0)} \sim N(0,1)$$

Slide created by Ben Arnold 13

So next, we want to pick a test statistic against which to measure the actual difference that we observe. And then to see whether that is larger or smaller than the test statistic that we've chosen. So given a hypothesis, the next step is to choose a test statistic.

And one example of a commonly used test statistic is called the Wald statistic, W-A-L-D. And this is the difference between groups divided by its standard error. So remember the difference between groups was that Greek letter theta that we chose to use to represent the difference between the two groups, the P2 minus P1.

So a Wald statistic, which is a chi-square type of statistic, will have a standard normal distribution under the null hypothesis with a mean of zero and a standard deviation of one. To use this statistic, we take our parameter theta zero that we estimate, and we divide by the standard error of that estimate. And we believe that this is distributed as a normal distribution with a mean of zero and a standard deviation of one.

3 : specify critical values based on Type I & Type II errors

Decision made (based on test statistic)	H_0 is true	H_a is true
Accept H_0	correct decision	Type II error (β)
Reject H_0	Type I error (α)	correct decision

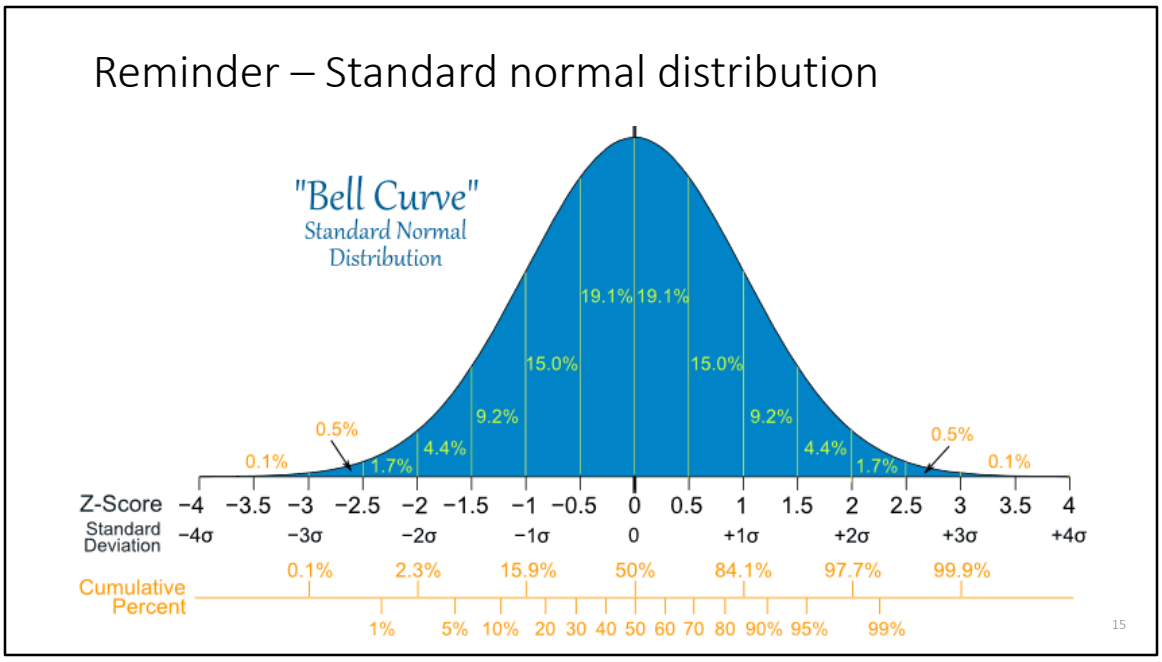
- Type I error: incorrectly reject the null when there is truly no difference (false positive)
- Type II error: incorrectly accept the null when there is, in fact, a difference (false negative)

Slide created by Ben Arnold 14

So next we want to specify the critical values based on type one and type two errors that might arise when we conduct our testing. So let's just quickly review the difference between type one error and type two errors. Now these are laid out, and I know these are review for most of you from prior classes.

But a type I error is the situation in which we incorrectly reject the null hypothesis, when there truly is no difference. So in other words, if the null hypothesis is actually true and we reject it because of our testing, that's a type I error, or sometimes called a false positive.

And a type two error would be to incorrectly accept the null when there is, in fact, a difference. And we would call this a false negative. So the way to think about these type one and type two errors are as false positive and false negative results.

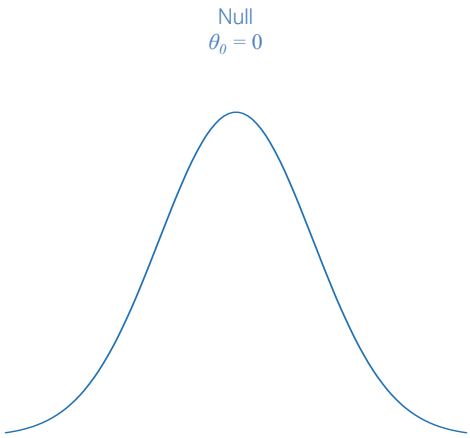


Just a quick reminder of the standard normal distribution with a mean of zero and standard deviation of one. So on the x-axis, here we actually use something called the z-score. Again, I'm assuming this is review for all of you. But it shows underneath that the cumulative percentage of the curve that is covered as you move further and further to the right, in this case.

So that if you go out to a far tail of the distribution, you can see how much of the curve is covered in blue to the left. So we're going to need to use these z-scores as we do some of this statistical testing. Again, this is just the standard normal distribution that you've seen many times.

What is power?

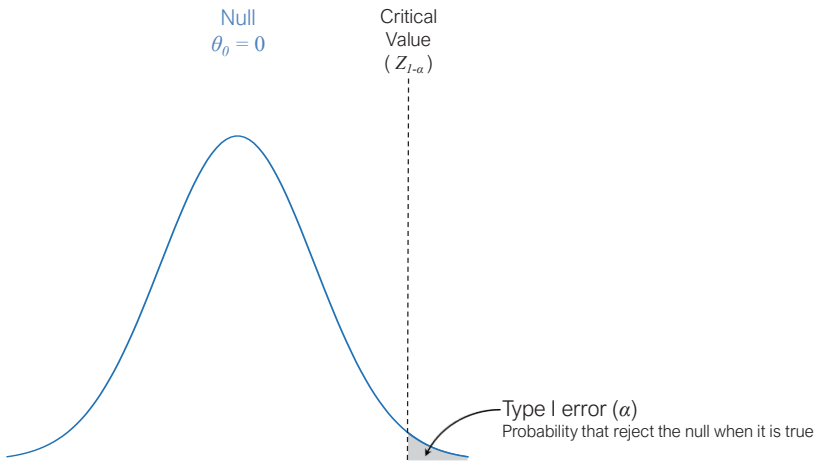
Start with the null distribution of X (2 slides ago)
It is normally distributed with mean 0 and standard deviation 1



Slide created by Ben Arnold 16

So what is power? Well, let's start with the null distribution of our test statistic. And that was two slides ago. This is normally distributed with a mean of zero and a standard deviation of one. And our null value here, our normal idea is that the difference theta zero is equal to zero. In other words, no difference.

Type I error (α) probability is the area under the null distribution of X to the right of the **critical value**
If α = 0.05, then $Z_{1-\alpha}$ is the 95th percentile of the standard normal distribution.



Slide created by Ben Arnold

17

So let's decide what we want to pick as a critical value. That is, some point on this normal distribution where our type one error is set to the right of this critical value. So that's the grayed out area on the slide here, and that's the type one error. We also call that alpha.

Because this distribution represents all possible values for theta given a known true value of zero. The values above our critical value are the ones that would be incorrectly identified as a true difference, the values corresponding to a false positive. That means that the proportion of values above this critical value is the probability of a false positive or the probability of a type one error.

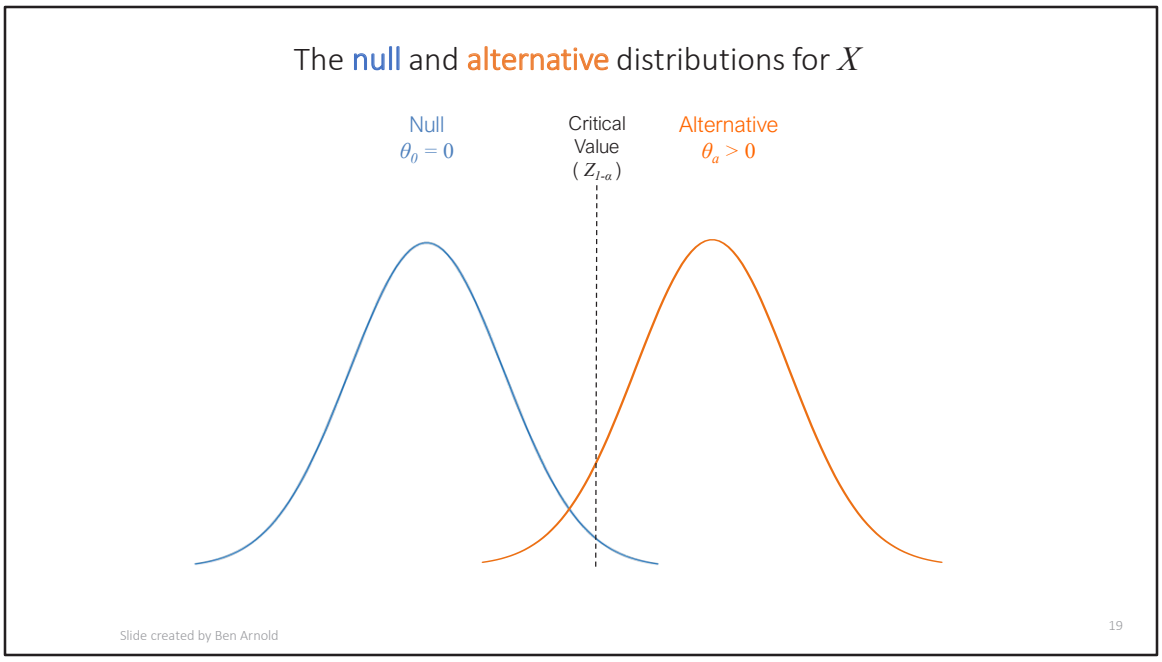
QUESTIONS

- Can we just set alpha = zero to make sure our result/finding is not due to random error?
  - No – Some random error (i.e, some sampling error) is unavoidable
- What if we increase our sample size until it includes all sampling units?
  - This is a census, not a sample
    - In which case don't need to estimate parameters we can just calculate them

18

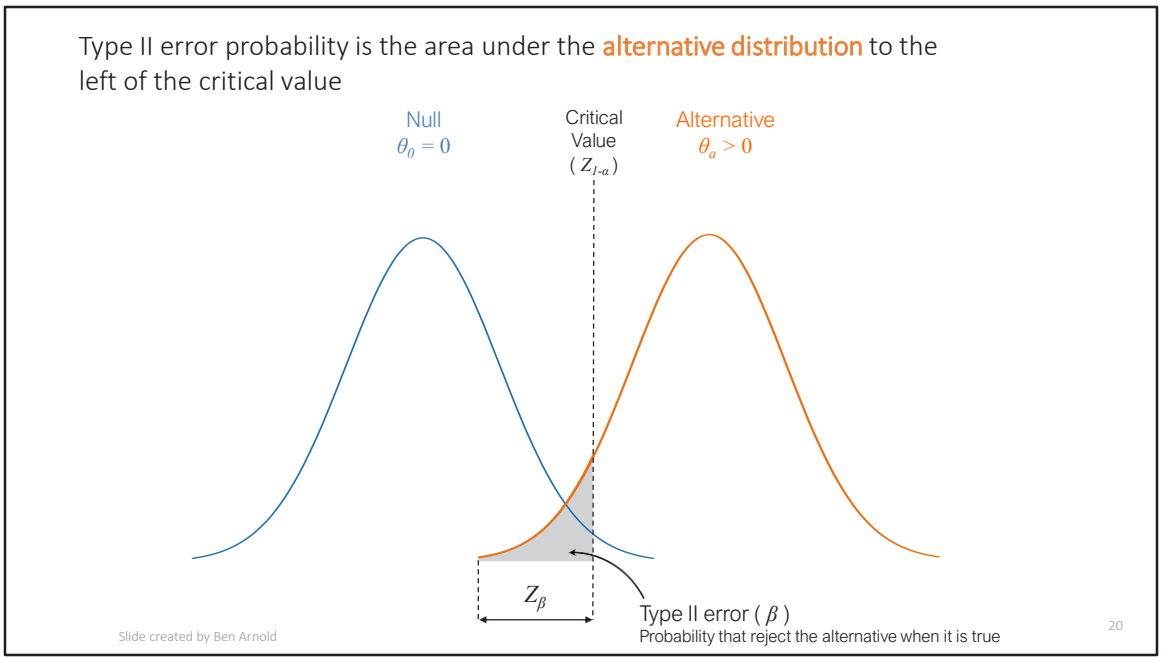
So could we just set alpha equal to zero to make sure our result or finding is not due to random error? Well, that's not practically possible because there's always going to be some random error or some sampling error that's an unavoidable in any study. Well, what if we increased our sample size until it includes all sampling units?

Well, if that's the case, then that's a census not a sample. So if we are conducting a census where we enroll the entire population, we wouldn't need any of these kinds of measures, because we're not taking a sample against which we're trying to draw inference to the whole population. If we measure the whole population, that's a census. So we don't need any of these techniques to estimate sample size, because we essentially have the entire population.



So let's look at the null and alternative distributions for our test statistic here, this chi value that we're looking at. So we have a null distribution on the left and an alternative distribution on the right. And recall that for the null distribution, we said that theta zero is equal to zero. That is no difference. And our alternative idea was that theta zero was greater than zero, that the difference between the two populations was greater than that theta of no difference.

And we've put our critical value that we set up on the earlier slide for the null distribution. You see that critical value. And notice that it also cuts at some point the alternative distribution.

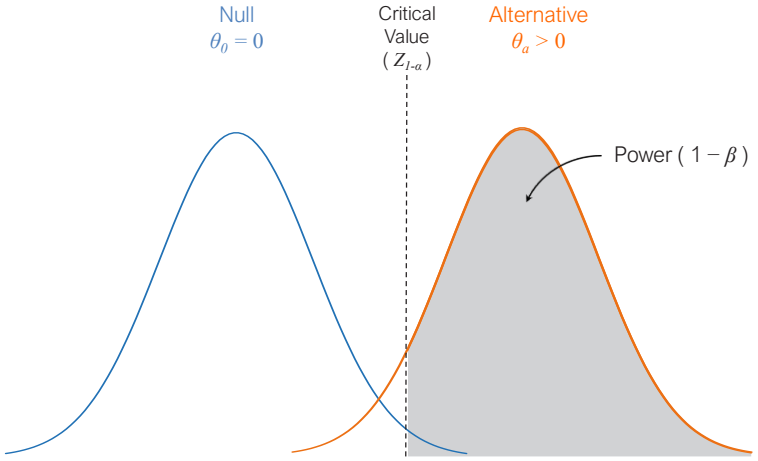


And what we're going to do is first look at the area to the left of the critical value as it cuts off the alternative distribution. So remember the alternative distribution was this idea that the two values, one is greater than the other. One population risk of illness is greater than the other population risk of illness.

So that area to the left is called the type two error, and it's also referred to as beta. That's the probability that we reject the alternative hypothesis even when it's true. So notice how this critical value is setting up areas both on the null distribution and on the alternative distribution.

In this slide, we're just looking at the area to the left the critical value underneath the curve for the alternative distribution. That's the type two error area. And we call that area beta.

Power ($1-\beta$) is the area under the alternative distribution to the right of the critical value.



Slide created by Ben Arnold 21

Now what do we call the area to the right of the critical value under the alternative distribution? Well, the grayed out area on this curve is referred to as the power of the study. And that, of course, is equal to one minus beta.

Well, why is that? Well, the area to the left of the curve was beta. And the area to the right of the curve then must be one minus beta, because the proportion of the area under the entire curve is equal to one.

So we have beta to the left of the critical value on the alternative distribution. Power to the right of the critical value on the alternative distribution. Another term for power is one minus beta.

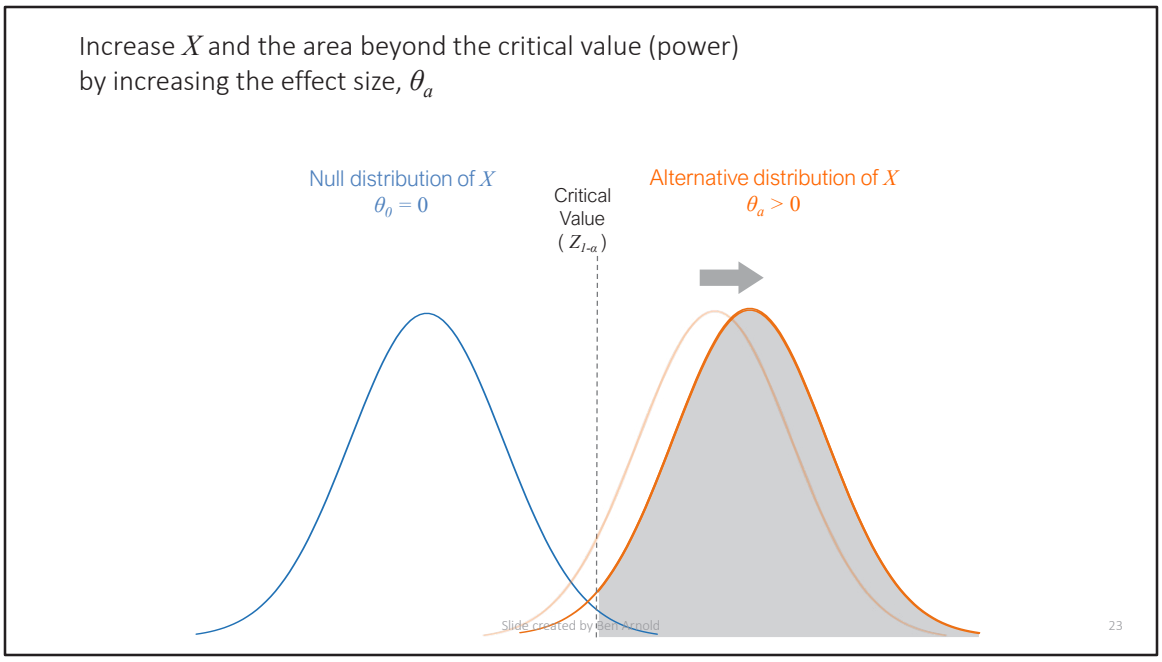
How to increase power?

- To increase power, need to shift the alternative distribution further away from the null distribution.
- How to increase this test statistic? $X = \frac{\hat{\theta}_a}{SE(\hat{\theta}_a)}$

Slide created by Ben Arnold 22

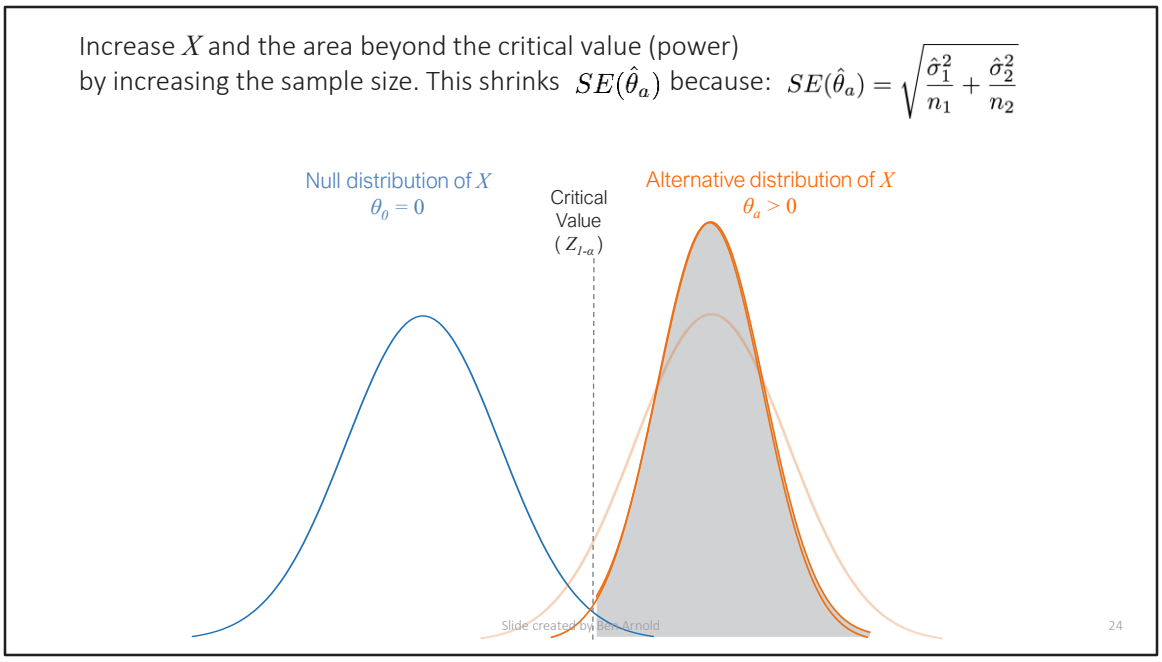
So how can we increase the power of the study? How can we increase that grayed out area to the right? Well, we need to shift the alternative distribution further away from the null. Right? That'll make more of that area gray. More of that curve on the right will become gray.

So how do we increase this test statistic? How do we make the value of this chi value here greater? In this formula where we have theta sub a divided by the standard error of theta sub a, how do we make that a larger number?



So how do we increase that grayed out area that represents power on the right side? Well, one way to do it is to increase the effect size. In other words, to move the means of the two curves farther apart from each other. So if we further separate the null and alternative distributions, it's going to increase the area that is gray under the curve on the right.

So if we slide the curve to the right, we keep the critical value where it was with respect to the null distribution, but the critical value has changed location, in a sense, on the curve on the right. So now you see that there's more gray represented under the curve than there was before on the right side of the alternative distribution.



So another way to increase the power or the gray area under the curve is to increase the sample size. That's because increasing the sample size will shrink the standard error. Now here's a standard error formula for something where we have two samples. And we're using the two samples here with the two distributions referred to here as n1 and n2.

As both n1 and n2 get larger, as the sample size gets larger, the denominators are bigger, so the standard error is going to get smaller. So the standard error shrinks, which makes the curve tighter. The distribution of the curve is tighter. And that's going to further increase the area to the right of the critical value. The two ends of the tails on the curve on the right are going to come closer to each other.

Example calculation from the recreational water study

1. Specify hypotheses
   - $H_0 : \theta_0 = 0$; $H_a : \theta_a > 0$
2. Pick critical values
   - Power (aka, Type 2 error rate): 80% = $Z_{0.80} = 0.84$
   - Precision (aka, Type 1 error rate): 5% for a one-sided test = $Z_{1-\alpha} = Z_{0.95} = 1.64$
3. Specify parameters required in the calculation
   - $p_1 = 0.034$ cumulative incidence among non-swimmers from a nearby study (Colford 2007)
   - $p_2 = 0.044$ cumulative incidence among swimmers (for a 1% increase)
4. Calculate sample size (by hand or with software – next slides)

Slide created by Ben Arnold 25

Let's now do an example calculation using the recreational water study situation. And I think the concept will then become even clearer as we work through with numerical demonstration. So first, let's specify our hypotheses.

Our null hypothesis is that the difference between the two groups, theta zero, is zero. The alternative hypothesis, we're going to use what we call a one sided hypothesis here. We're testing the idea that the alternative hypothesis is that the swimmers have more illness than the non swimmers, so that the difference between the two groups would be greater than zero.

Second step is we want to pick the critical values that we want to use in doing these calculations. So for power, or type two error rate, it's usually chosen to use 80%. And in order to get a point on the normal distribution where 80% of the curve is covered, we choose a z that gives us 0.80 under the curve. So that's an actual z-score of 0.84.

Now we address the precision that we want in the study, or the type one error rate. And usually, we choose a 5% value here. Because this is a one sided test, we need z to give us a one minus alpha of 0.95.

Example calculation from the recreational water study

1. Specify hypotheses
   - $H_0 : \theta_0 = 0$; $H_a : \theta_a > 0$
2. Pick critical values
   - Power (aka, Type 2 error rate): 80% = $Z_{0.80} = 0.84$
   - Precision (aka, Type 1 error rate): 5% for a one-sided test = $Z_{1-\alpha} = Z_{0.95} = 1.64$
3. Specify parameters required in the calculation
   - $p_1 = 0.034$ cumulative incidence among non-swimmers from a nearby study (Colford 2007)
   - $p_2 = 0.044$ cumulative incidence among swimmers (for a 1% increase)
4. Calculate sample size (by hand or with software – next slides)

Slide created by Ben Arnold 25

And that z-score on the normal distribution is 1.64. If you go back and look at the normal distribution curve, you'll see that 1.64 z-score would place 5% of the area above that point in the right tail. So the value we want to use is 1.64

Next, we want to specify the parameters required in our calculation. Well, this was given to us earlier. The probability of illness or cumulative incidence in the non swimmers we said was 0.034. And as I mentioned, that came from a prior study. So we had a background estimate of what the incidence of illness would be in the non swimmers. And we're trying to be able to detect at least a 1% increase in the swimmers, so that would be 4.4%.

So next, we calculate the sample size by hand or with software. And that'll be on the next slides.

Sample size equation for a risk ratio of proportions

(algebra challenge: derive the equation in panel 2 of Schulz and Grimes 2005)

$$m = \frac{(Z_{1-\beta} + Z_{1-\alpha/2})^2 \, [p_1(1-p_1) + p_2(1-p_2)]}{(p_1 - p_2)^2}$$

- m = sample size in each group (swimmers, non-swimmers; assumed equal size)
- p_1 = cumulative incidence among swimmers
- p_2 = cumulative incidence among non-swimmers
- Z_x = the x'th percentage point of the standard normal distribution

Slide created by Ben Arnold 26

So in order to do this, we have to use some algebra. And we have to use a sample size equation. And here, the equation is given for us.

Let me just explain each of the terms in this equation. So m in the equation is the sample size for each group, the swimmers and the non swimmers. And in this example, we're assuming them to be equal sizes. There are variations of the formula that allow you to adjust for differently sized groups.

P1 is the cumulative incidence among the swimmers. P2 is the cumulative incidence among the non swimmers. Z sub x is the x percentage point of the standard normal distribution.

Recreational water example, sample size arithmetic

$$
\begin{aligned}
m &= \frac{(Z_{1-\beta} + Z_{1-\alpha/2})^2 \left[p_1(1-p_1) + p_2(1-p_2)\right]}{(p_1 - p_2)^2} \\
&= \frac{(0.84 + 1.64\ )^2 \left[0.044(1-0.044) + 0.034(1-0.034)\right]}{(0.044 - 0.034)^2} \\
&= 4{,}607
\end{aligned}
$$

Slide created by Ben Arnold 27

So in our problem, we're going to put in the following values. For Z of one minus beta, we chose 0.84, because that's 80% power. For Z of one minus alpha, we chose 1.64, because on the z-score, that's the critical value that places 5% of the curve above that point.

P1 was 0.044. So then we have one minus P1. Plus P2 was 0.034. Then we have the quantity one minus P2 or one minus 0.034. The denominator is P1 minus P2 squared. We plug those numbers in.

"Most hand [sample size] calculations diabolically strain human limits, even for the easiest formula"

(Schulz & Grimes 2005)

Recommend that you practice writing functions in R that use these formulas.
See example in the video on sample size for cluster randomized trials.

So an important concept to keep in mind was made clear in the Schulz and Grimes article, that most hand or sample size calculations diabolically strain human limits even for the easiest formula. So this sort of thing is usually always done by software. And you may want to practice writing functions in R that use these formulas. So there is an example in the video on sample size for cluster randomized trials where you can do this.

Presentation overview

- Big-picture concepts & motivation
- Example 1: Understanding power & marine water exposure / gastrointestinal illness case study
- **Example 2: Calculate sample size for an individually randomized trial**
- Additional issues and recap

29

Next, let's calculate the sample size for an individually randomized trial.

Example #2: Mobile app RCT (individual)

JMIR MHEALTH AND UHEALTH Goyal et al

Original Paper

A Mobile App for the Self-Management of Type 1 Diabetes Among Adolescents: A Randomized Controlled Trial

Shivani Goyal[1,2*], BEng, MSc, PhD; Caitlin A Nunn[3*], MSc; Michael Rotondi[4], PhD; Amy B Couperthwaite[4], MSc; Sally Reiser[5], RD; Angelo Simone[5], MD; Debra K Katzman[6,7], MD, FRCP(C); Joseph A Cafazzo[1,2,8], PhD, PEng; Mark R Palmert[3,6,9], MD, PhD

[1]Centre for Global eHealth Innovation, Techna Institute, University Health Network, Toronto, ON, Canada
[2]Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON, Canada
[3]Division of Endocrinology, The Hospital for Sick Children, Toronto, ON, Canada
[4]School of Kinesiology & Health Science, York University, Toronto, ON, Canada

Slide created by Yoshika Crider

And for the example here, we're going to look at a mobile app randomized controlled trial in this article by Goyal, et al.

Given parameters

"Sample size was determined based on a nominal 2-sided type 1 error rate of 5% and 80% power. Estimates of standard deviation in HbA1c ranging from 0.50 to 0.75 were used to determine the minimum number of participants required to detect a clinically relevant (>=0.5) change in HbA1c levels." (Goyal et al, 2017)

Parameter	Value
Type 1 error rate (alpha)	0.05
Power (1-beta)	0.80
MDE ("clinically relevant" change in HbA1c levels)	>=0.5
Standard deviation	0.50-0.75
Baseline HbA1c (from eligibility criteria)	8.0-10.5

Slide created by Yoshika Crider

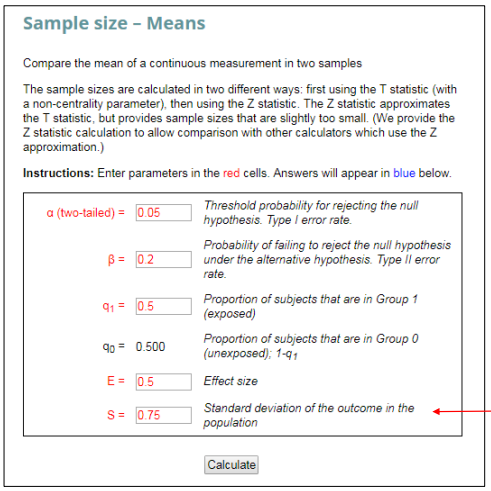
And so let's look at the parameters that are given to us.

So it says in this study that sample size was determined based on a nominal two sided type one error rate of 5% and 80% power. Estimates of standard deviation in hemoglobin A1c ranging from 0.50 to 0.75 were used to determine the minimum number of participants required to detect a clinically relevant, that is a greater than or equal to 0.5 change in hemoglobin A1c levels.

So what are the values that we choose to put in here? Well, for our type one error rate, most often, 0.05 is used, just as we used in our last example. Similarly for power, a value of 0.80 is usually used.

In this study, the authors wanted to detect a minimally detectable effect of greater than or equal to 0.5. The standard deviation was given to us as 0.50 to 0.75. And the baseline hemoglobin A1c given to us in the eligibility criteria was eight to 10.5.

http://www.sample-size.net/sample-size-means/



Sample size – Means

Compare the mean of a continuous measurement in two samples

The sample sizes are calculated in two different ways: first using the T statistic (with a non-centrality parameter), then using the Z statistic. The Z statistic approximates the T statistic, but provides sample sizes that are slightly too small. (We provide the Z statistic calculation to allow comparison with other calculators which use the Z approximation.)

Instructions: Enter parameters in the red cells. Answers will appear in blue below.

α (two-tailed) =	0.05	*Threshold probability for rejecting the null hypothesis. Type I error rate.*
β =	0.2	*Probability of failing to reject the null hypothesis under the alternative hypothesis. Type II error rate.*
q_1 =	0.5	*Proportion of subjects that are in Group 1 (exposed)*
q_0 =	0.500	*Proportion of subjects that are in Group 0 (unexposed); 1-q_1*
E =	0.5	*Effect size*
S =	0.75	*Standard deviation of the outcome in the population*

Calculate

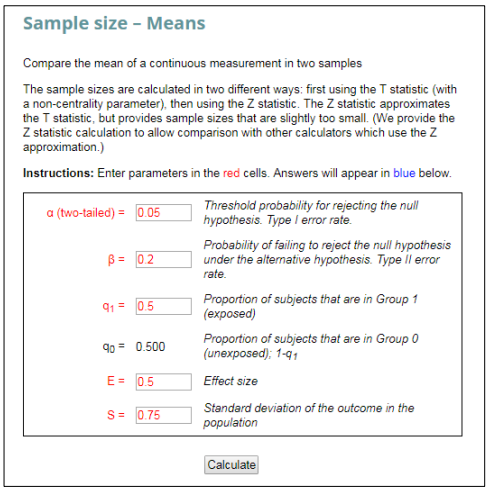
Let's make the most conservative assumption here

Slide created by Yoshika Crider

So there are many online calculators that will allow you to do this, but here's one. You see the website at the top. This is a plug and chug sort of example.

But if you put in the values in red here, you will then press the calculate button. And we've chosen to use the most conservative standard deviation of the outcome in this population. So let's press the calculate button.

http://www.sample-size.net/sample-size-means/



Sample size – Means

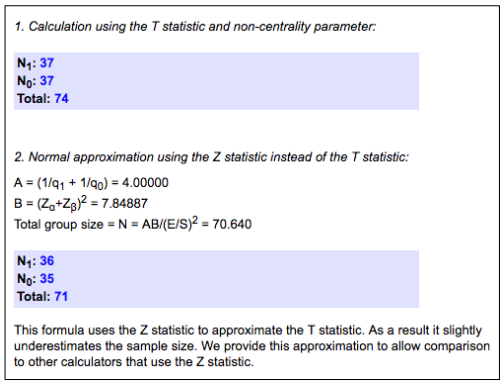
Compare the mean of a continuous measurement in two samples

The sample sizes are calculated in two different ways: first using the T statistic (with a non-centrality parameter), then using the Z statistic. The Z statistic approximates the T statistic, but provides sample sizes that are slightly too small. (We provide the Z statistic calculation to allow comparison with other calculators which use the Z approximation.)

Instructions: Enter parameters in the red cells. Answers will appear in blue below.

α (two-tailed) =	0.05	*Threshold probability for rejecting the null hypothesis. Type I error rate.*
β =	0.2	*Probability of failing to reject the null hypothesis under the alternative hypothesis. Type II error rate.*
q_1 =	0.5	*Proportion of subjects that are in Group 1 (exposed)*
q_0 =	0.500	*Proportion of subjects that are in Group 0 (unexposed); 1-q_1*
E =	0.5	*Effect size*
S =	0.75	*Standard deviation of the outcome in the population*

Calculate



1. Calculation using the T statistic and non-centrality parameter:

N_1: 37
N_0: 37
Total: 74

2. Normal approximation using the Z statistic instead of the T statistic:

$A = (1/q_1 + 1/q_0) = 4.00000$
$B = (Z_\alpha + Z_\beta)^2 = 7.84887$
Total group size = $N = AB/(E/S)^2 = 70.640$

N_1: 36
N_0: 35
Total: 71

This formula uses the Z statistic to approximate the T statistic. As a result it slightly underestimates the sample size. We provide this approximation to allow comparison to other calculators that use the Z statistic.

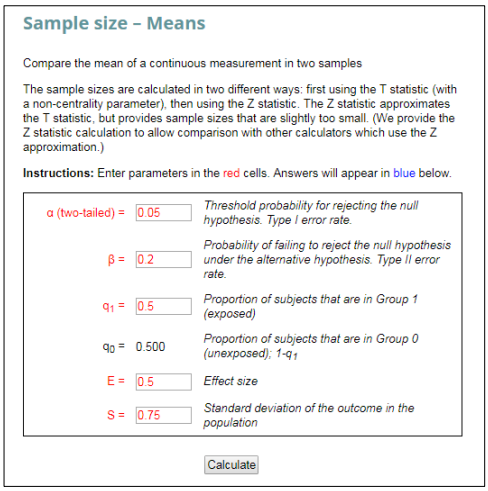
But their final sample size was 46 per arm, 92 total participants.... Why?

Slide created by Yoshika Crider

Then we get the following results. The two groups should each have 37 people in them for a total of 74 people. And you could do a slightly different calculation using the z statistic instead of the t statistic, which is what the default in this situation is. But you get a very similar result, 71.

These sample size calculations don't need to be exactly to the decimal point the same. You're trying to get a sense of how large the sample size was. But the authors actually chose a final sample size of 92 or 46 per arm. Why do you think that is?

http://www.sample-size.net/sample-size-means/



Sample size – Means

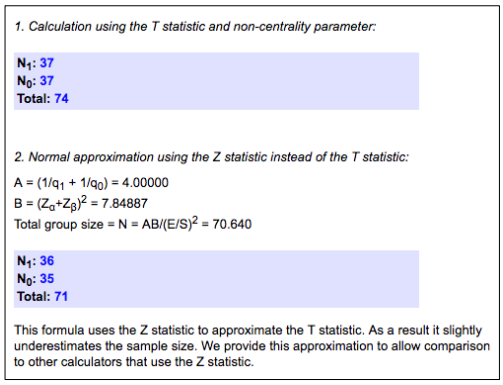
Compare the mean of a continuous measurement in two samples

The sample sizes are calculated in two different ways: first using the T statistic (with a non-centrality parameter), then using the Z statistic. The Z statistic approximates the T statistic, but provides sample sizes that are slightly too small. (We provide the Z statistic calculation to allow comparison with other calculators which use the Z approximation.)

Instructions: Enter parameters in the red cells. Answers will appear in blue below.

α (two-tailed) =	0.05	*Threshold probability for rejecting the null hypothesis. Type I error rate.*
β =	0.2	*Probability of failing to reject the null hypothesis under the alternative hypothesis. Type II error rate.*
q_1 =	0.5	*Proportion of subjects that are in Group 1 (exposed)*
q_0 =	0.500	*Proportion of subjects that are in Group 0 (unexposed); 1-q_1*
E =	0.5	*Effect size*
S =	0.75	*Standard deviation of the outcome in the population*

Calculate



1. Calculation using the T statistic and non-centrality parameter:

N_1: 37
N_0: 37
Total: 74

2. Normal approximation using the Z statistic instead of the T statistic:

$A = (1/q_1 + 1/q_0) = 4.00000$
$B = (Z_\alpha + Z_\beta)^2 = 7.84887$
Total group size = $N = AB/(E/S)^2 = 70.640$

N_1: 36
N_0: 35
Total: 71

This formula uses the Z statistic to approximate the T statistic. As a result it slightly underestimates the sample size. We provide this approximation to allow comparison to other calculators that use the Z statistic.

Why?

Buffered for up to 25% loss to follow up (37*1.25=46)

Slide created by Yoshika Crider

Well, they were buffering for up to a 25% loss to follow up. So they took the 37 and multiplied it by 1.25. So that's a very conservative approach. They were afraid they might lose that many people. So they made the study larger than it needed to be.

Additional issues you may need to consider

- Loss to follow-up
  - Increase your sample size by your anticipated attrition rate (e.g., 10%)
- Study design for multiple primary outcomes
  - Repeat the exercise and find the limiting outcome
- Study design for planned subgroup analyses
  - Size the study around subgroups (may require sampling based on subgroup)
- Binary outcomes with repeated measures
  - Similar, but slightly modified equations (Leon 2004)
- Sample size / power calculations for complicated designs
  - Consider the use of Monte Carlo simulation (Feiveson 2002, Arnold 2011)

Slide created by Ben Arnold 35

So there are additional issues you may need to consider in sample size calculations. The first is loss to follow up. And we just saw an example of what some authors did to increase their sample size in that prior example by 25%. Usually, we tend to want to make them just 10% larger, but sort of depends on the context of your study.

You might have a study design where there are multiple primary outcomes. In which case, you need to repeat the exercise and find the limiting outcome. If you're going to do subgroup analysis, you need to plan for them in advance and size the study around subgroups. So you may need to make the overall study much larger than you intended so that the results can be analyzed in the subgroups for the subgroup to have a large enough sample size, which would then inflate the entire study, of course.

Binary outcomes with repeated measures are a special situation. But there are modified equations for this, if you're repeatedly measuring some outcome yes, no, yes, no, yes, no over time.

And then there are approaches to sample size and power calculations for complicated designs. And one of them uses the Monte Carlo simulation. And there's two references here that can help you with that.

Summary of Key Points

- Sample size calculations require that you specify:
  1. basic design (fraction treated/exposed), parameter of interest, and hypothesis
  2. outcome variability (get this automatically for binomial outcomes)
  3. minimum detectable effect
  4. desired level of power and Type I error (alpha)

- Sample size, MDE, and power equations are all just re-arranged versions of the same equation.

Slide created by Ben Arnold 36

So to summarize, sample size calculations require that you specify the basic design, which includes the fraction treated or exposed, the parameter of interest, and hypotheses, the variability in the outcome, and you can get this automatically for binomial outcomes if you know P1 and P2, the minimum detectable effect, and then the desired level of power and type one error or alpha.

Sample size, minimum detectable effect, and power equations are all just rearranged versions of the same equation. So you could rearrange that algebraic equation we saw earlier if you needed to find a different entity within it.

Additional comments

- These calculations are important, but ultimately they rely on a lot of guesswork. In practice, there is a lot of interplay between the science, the budget, and the sample size calculations. This is the "sample size samba" (Schulz & Grimes 2005)
- These skills are extremely useful to have in your study design tool box
- If you are interested in the derivation that underlies the basic sample size and power equations, Steve Selvin's book *Statistical Analysis of Epidemiologic Data* has a nice introduction (Ch 3)

Slide created by Ben Arnold 37

So some additional comments. These calculations are important. But ultimately, they rely on a lot of guesswork. So in practice, there's a lot of interplay between the science, the budget, and the sample size calculation.

Schulz and Grimes calls this the sample size samba. You scale the sample size up or down based on both your needs and the study, but also the budget you have. These skills are extremely useful to have in your study design tool box. And if you're interested in the derivation that underlies the basic sample size and power equations, Steve Selvin's book Statistical Analysis of Epi Data has a nice introduction in Chapter three.

Additional sample size resources

- www.power-calculator.org
  - RCT, cluster RCT calculations for continuous and binary outcomes (can be super buggy and slow though!)
- https://jadebc.shinyapps.io/samplesize/
  - Individually randomized trial calculations for continuous and binary outcomes
  - Shows curves to visualize trade-offs in parameters
- http://www.sample-size.net/
  - Individual and clustered design options
  - No visualization, just table output
- https://ssc.researchmethodsresources.nih.gov/ssc/
  - Group trial calculations
  - Lots of parameters that can get a little complicated

Slide created by Yoshika Crider

Here's some additional sample size resources, websites where you can go and put in values and carry out these calculations, and a list of references you can use.

References

- Arnold, B.; Hogan, D.; Colford, J. & Hubbard, A. Simulation methods to estimate design power: an overview for applied research. BMC Med Res Methodol, 2011, 11, 94
- Arnold, B. F.; Null, C.; Luby, S. P.; 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
- Colford JM, Schiff KC, Griffith JF, Yau V, Arnold BF, Wright CC, et al. Using rapid indicators for Enterococcus to assess the risk of illness after exposure to urban runoff contaminated marine water. Water Res. 2012;46: 2176–2186.
- Colford, J. M.; Wade, T. J.; Schiff, K. C.; et al. Water quality indicators and the risk of illness at beaches with nonpoint sources of fecal contamination. Epidemiology, 2007, 18, 27-35
- Campbell, M. J.; Donner, A. & Klar, N. Developments in cluster randomized trials and Statistics in Medicine. Stat Med, 2007, 26, 2-19
- Duflo, E.; Glennerster, R. & Kremer, M. Using Randomization in Development Economics Research: A Toolkit. 61 Handbook of Development Economics, 2007, Volume 4, 3895-3962
- Feiveson, A. H. Power by simulation. Stata Journal, 2002, 2, 107-124
- Leon, A. C. Sample-size requirements for comparisons of two groups on repeated observations of a binary outcome Eval Health Prof, 2004, 27, 34-44
- Murray, D. M.; Varnell, S. P. & Blitstein, J. L. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health, 2004, 94, 423-432
- Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. Int J Epidemiol. 2015;44: 1051–1067.
- Schulz, K. F. & Grimes, D. A. Sample size calculations in randomised trials: mandatory and mystical. Lancet, 2005, 365, 1348-1353
- Selvin, S. Statistical Analysis of Epidemiologic Data (2nd ed). Oxford University Press, 1996
- Victora, C. G.; Adair, L.; Fall, C.; Hallal, P. C.; Martorell, R.; Richter, L.; Sachdev, H. S. Maternal and child undernutrition: consequences for adult health and human capital. Lancet, 2008, 371, 340-357
- Victora, C. G.; de Onis, M.; Hallal, P. C.; Blossner, M. & Shrimpton, R. Worldwide timing of growth faltering: revisiting implications for interventions. Pediatrics, 2010, 125, e473-e480

Sample size and power for cluster randomized trials

Jack Colford
Adapted from slides by Ben Arnold

PH250G

1

JACK COLFORD: Let's continue our discussion of sample size and power. But now let's examine a common situation in epidemiology when we're working with clustered data or clustered designs.

Presentation overview

- Big-picture concepts related to clustering
- Example: Clustering-related considerations & nutritional intervention / child growth case study
- Additional issues and recap

2

First we'll talk about some big picture concepts related to clustering in general. These apply to both trials and observational studies. Then with a case-based example we'll talk about the considerations related to clustering when we study nutritional interventions and their impact on child growth.

And we use a case study from the WASH Benefit study. And then we'll circle back and close with some additional issues and a recap of the main issues.

Clustered designs : common in epidemiologic studies

- Often it makes the most sense (scientifically and/or logistically) to deliver an intervention or program to a group of individuals (Murray 2004)
- Changes the physical or social environment (handwashing behavior change)
  - Cannot be delivered to individuals (centralized water treatment)
  - Investigators wish to capture group-level dynamics (deworming campaigns)
- Outcomes are often measured at the individual level, where individuals are grouped into clusters.
- Clusters can be defined by space (e.g., village membership) or by any shared characteristic that connects group members within a cluster (e.g., health practitioner)

Slide created by Ben Arnold 3

Cluster designs are very common in epidemiology because often it makes the most sense, either scientifically, and/or logistically, or both to deliver an intervention or program to a group of individuals.

So there might be situations where we want to change the physical or social environment, such as a handwashing behavior change, or a situation that clearly can't be delivered to individuals, like centralized water treatment. And in these situations, investigators often wish to capture group level dynamics.

The outcomes in these studies are often measured at the individual level, but then the individual results are grouped into clusters. Clusters could be defined by space, such as membership in a village, or by any shared characteristic that connects group members within a cluster. For example, they might be all the patients who attend a clinic run by a specific health practitioner.

Clustered designs : Independence assumptions

- A defining characteristic of clustered designs is that in the analysis, repeated observations within the cluster are not assumed to be independent. They are assumed to be correlated.
- However, investigators typically design studies so that clusters are independent (i.e., no interference between individuals in different clusters)
- The result of correlated outcomes within each cluster is that each measurement tends to provide less information than it would if all outcome measurements were independent

Slide created by Ben Arnold 4

Cluster designs rely on some independence assumptions. Defining characteristics of cluster designs is that in the analysis repeated observations within the cluster are not assumed to be independent. They are assumed to be correlated.

However, investigators typically design studies so that clusters are independent. That is there's no interference between individuals in different clusters.

The result of correlated outcomes within each cluster-- not across the cluster, but within each cluster-- is that each measurement tends to provide less information than it would if all the outcome measurements were independent.

Clustered designs : Within-cluster correlation

- When observations within a cluster are correlated, a common way to summarize the correlation is with the intraclass correlation coefficient (ICC):

$$ICC = \rho = \frac{\tau^2}{\tau^2+\sigma^2}$$

where τ^2 is the between-cluster variance and σ^2 is the within-cluster variance.

- The ICC is the fraction of the total variance ($\tau^2 + \sigma^2$) that is explained by the between-cluster variance (τ^2).

- If cluster membership explains a lot of the variability in the outcome, then:
  - the ICC will be larger
  - outcomes within each cluster will be more correlated than if cluster membership had no effect on the outcome

Slide created by Ben Arnold 5

So when we talk about cluster designs we have to understand the concept of within cluster correlation and how to measure that. So when observations within a cluster are correlated, a common way to summarize that correlation is with something called the intraclass correlation coefficient, or ICC.

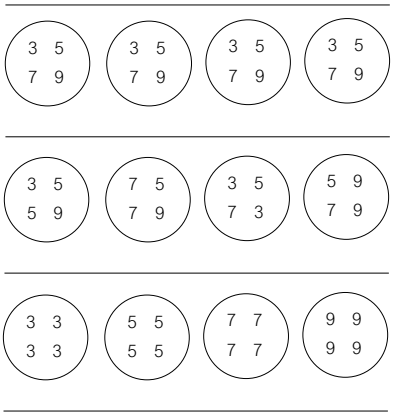
This is usually abbreviated with the Greek letter rho. And the formula for it is given by these Greek letters, Tao squared over Tao squared plus sigma squared. And tao squared, you see in the numerator, represents the between cluster variance. And sigma squared is the within cluster variance.

The ICC is the fraction of the total variance that's explained by the between cluster variance. So if you just think about this in broad terms-- numerator over the nominator-- the numerator is the between cluster variance. And the denominator is the total variance. So the fraction is giving the proportion of the variance that's explained by the between cluster variance.

If the cluster membership explains a lot of the variability of the outcome, then the ICC will be larger and the outcomes within each cluster will be more correlated than if cluster membership had no effect on the outcome.

A picture of difference ICCs (ρ)

Clusters (circles) with individual level outcomes

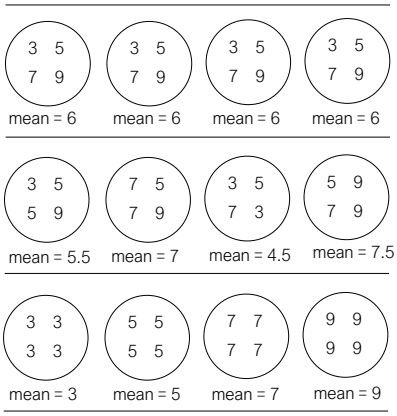


Slide created by Ben Arnold 6

Let's consider an example. Think of each of these circles as a village. So within each village there are four subjects, each with individual level outcomes measured. You see here three, five, seven, and nine. But let's compare the first study, to the second study, to the third study-- that is the first row, to the second row, to the third row.

A picture of difference ICCs (ρ)

Clusters (circles) with individual level outcomes



$\rho = 0$

No between cluster variation
No within cluster correlation
Cluster membership is not informative

$\rho = 0.18$

Some between cluster variation
Some within cluster correlation
Cluster membership is a bit informative

$\rho = 1$

High between cluster variation
Perfect within-cluster correlation
Cluster membership tells you everything

Slide created by Ben Arnold 7

In the first row the ICC is zero because between the clusters there is no variability at all. All four of them are exactly the same. They all have the same mean and the same four values, in fact. There's just no variability between them.

So the cluster membership in this first study let's say doesn't tell us anything. Having four people within each study is no different than having one person within each study. We don't learn anything by membership, having additional subjects in each cluster.

Let's talk now about the second row. Here we have some between cluster variation and some within cluster correlation. But cluster membership is only a bit informative. It tells us a little bit but not a lot based on what cluster the individual is in.

And finally, the farther extreme, here we see very high between cluster variation. The clusters are very different between each other. There's perfect within cluster correlation. So within each village, within each cluster, all the measurements are the same.

In the first village all measurements are three. In the second village all measurements are five. And so forth. So here, cluster membership tells you everything.

Clustered designs : The Design Effect

- The ICC influences sample size calculations through the design effect (*Deff*):

$$D_{eff} = 1 + (m - 1)\rho$$

where m is the average number of observations per cluster

- The design effect is the ratio between the variances of a design with group-level correlation versus a design where all units are independent.
  - $\rho = 0 \Rightarrow$ have as much power as an individually randomized trial
  - $\rho = 1 \Rightarrow$ effective sample size is the number of clusters

- Even if ρ is small, the design effect can be large if cluster size (m) is large.

Slide created by Ben Arnold 8

So let's start to learn how to work with rho. So the first place we want to pay attention to where rho has an impact is on something called the design effect. That's given by this formula here, one plus the quantity m minus one times rho, where m is the average number of observations per cluster.

So on that previous slide there were four observations in each cluster. So m would be four in that situation. The ICC then is going to influence sample size calculation through this design effect.

And the design effect is the ratio between the variances of a design with group level correlation versus a design where all the units were independent.

So in a situation where rho equals zero-- go back and look at the first row of the earlier slide-- there we have as much power as if all the individuals in all the clusters were individually randomized because there's no correlation between the different clusters.

Clustered designs : The Design Effect

- The ICC influences sample size calculations through the design effect (*Deff*):

- The design effect is the ratio between the variances of a design with group-level correlation versus a design where all units are independent.
  - $\rho = 0 \Rightarrow$ have as much power as an individually randomized trial
  - $\rho = 1 \Rightarrow$ effective sample size is the number of clusters

- Even if ρ is small, the design effect can be large if cluster size (m) is large.

Slide created by Ben Arnold 8

In the last row, when rho is equal to one, the effective sample size was only four because there were only four clusters. And we didn't know anything more about each individual in each cluster because they were so correlated within each cluster. That's what a rho equal to one means.

So even if rho is still small-- like more than zero but still small-- the design effect can be large if the cluster size is large. And you can see that from the formula above. The design effect is one plus the quantity m minus one times rho. So even if rho is small, if m is very large you could still have a large design effect.

Sample Size and MDE equations for clustered designs with a continuous outcome (example from Duflo 2007)

$$k = \frac{\sigma^2(Z_{1-\beta}+Z_{1-\alpha})^2}{d^2P(1-P)m} \times [1+(m-1)\rho]$$

Design effect

$$MDE = d = (Z_{1-\beta}+Z_{1-\alpha})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{k}} \times \sqrt{\rho+\frac{1-\rho}{m}}$$

- σ^2 : variability of the outcome, Y (here: assumed equal in both groups)
- P : proportion of units allocated to treatment (optimal allocation is 50%)
- d : minimum detectable effect between groups
- k : total number of clusters enrolled in the study
- m : average number of individuals per cluster
- ρ : cluster level ICC

Another good reference is Rutterford et al. 2015 IJE (see references at end of slides)

9

To calculate the sample size that we need for clustered data, we use a sample size formula that adds in this design effect and also puts in m, number of individuals in each cluster, into the denominator of the first part of the equation.

So the algebra here isn't really so important. But we see that first we have to calculate the total number of clusters in our study. And that's given to us by this top formula for k. And then back to our prior concept of detecting the minimum detectable effect or, here, the distance between two groups, the d, we see the sample size formula once again with an extension that includes rho and m at the end of the formula.

Components here include the variability of the outcome sigma squared, what proportion of the units are allocated to treatment. If it's a one to one allocation to treatment controlled, then that p is 50%, or 0.5. The d we talked about was the minimum detectable effect between groups.

The k was the total number of clusters enrolled in the study. m was the average number of individuals in each cluster. And rho was the cluster level ICC.

Clustered designs : Practical considerations

- In clustered designs, investigators must choose the number of clusters per treatment arm (*k*) and the number of individuals (*m*) to measure within each cluster.
- As a general rule of thumb, gains in power are small for m > 1/ρ (Campbell 2007)
- Designs with more clusters and fewer observations per cluster are usually optimal from a statistical perspective
  - … but are often sub-optimal from a cost / logistical perspective
- Statistical power for clustered designs is usually driven by the number of clusters per arm and by the intraclass correlation coefficient (*ρ*) (Murray 2004)

Slide created by Ben Arnold 10

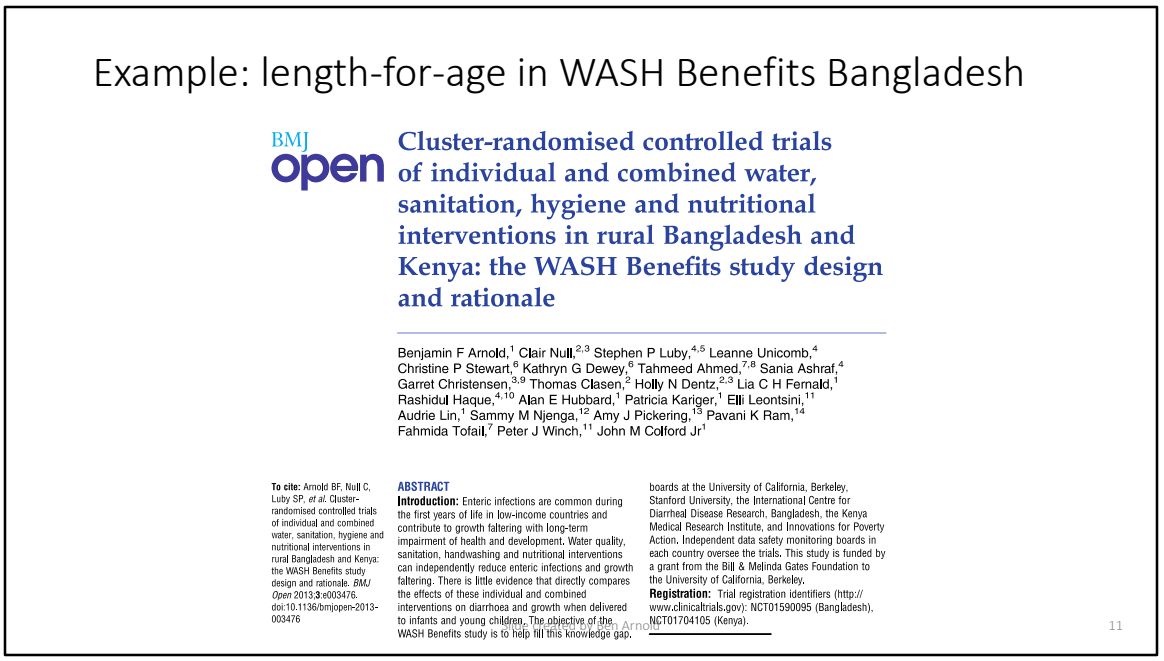
So some practical considerations for cluster designs-- the investigator has to choose the number of clusters per treatment arm, which is k, and the number of individuals, which is m, to measure within each cluster.

As a general rule of thumb we don't gain a lot in power when m, the average number of individuals in each cluster, is greater than one over rho, the cluster level ICC. So if rho were 0.1, there's not much to be gained in power for any value of m greater than one over 0.1, which would be 10.

Designs with more clusters and fewer observations per cluster are usually optimal from a statistical perspective. But they can often be suboptimal from a cost and logistical perspective because it's very expensive to go get additional clusters as opposed to just getting more people within one cluster.

Statistical power for clustered designs is usually driven by the number of clusters per arm and by rho, the intraclass correlation coefficient that we've learned about already.

Example: length-for-age in WASH Benefits Bangladesh



BMJ open

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

Benjamin F Arnold,[1] Clair Null,[2,3] Stephen P Luby,[4,5] Leanne Unicomb,[4] Christine P Stewart,[6] Kathryn G Dewey,[6] Tahmeed Ahmed,[7,8] Sania Ashraf,[4] Garret Christensen,[3,9] Thomas Clasen,[2] Holly N Dentz,[2,3] Lia C H Fernald,[1] Rashidul Haque,[4,10] Alan E Hubbard,[1] Patricia Kariger,[1] Elli Leontsini,[11] Audrie Lin,[1] Sammy M Njenga,[12] Amy J Pickering,[13] Pavani K Ram,[14] Fahmida Tofail,[7] Peter J Winch,[11] John M Colford Jr[1]

To cite: 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. doi:10.1136/bmjopen-2013-003476

ABSTRACT

Introduction: Enteric infections are common during the first years of life in low-income countries and contribute to growth faltering with long-term impairment of health and development. Water quality, sanitation, handwashing and nutritional interventions can independently reduce enteric infections and growth faltering. There is little evidence that directly compares the effects of these individual and combined interventions on diarrhoea and growth when delivered to infants and young children. The objective of the WASH Benefits study is to help fill this knowledge gap.

boards at the University of California, Berkeley, Stanford University, the International Centre for Diarrheal Disease Research, Bangladesh, the Kenya Medical Research Institute, and Innovations for Poverty Action. Independent data safety monitoring boards in each country oversee the trials. This study is funded by a grant from the Bill & Melinda Gates Foundation to the University of California, Berkeley.

Registration: Trial registration identifiers (http://www.clinicaltrials.gov): NCT01590095 (Bangladesh), NCT01704105 (Kenya).

So let's work with an example using length-for-age as the outcome in the WASH Benefits Bangladesh study that we've talked about before.

Presentation overview

- Big-picture concepts related to clustering
- **Example: Clustering-related considerations & nutritional intervention / child growth case study**
- Additional issues and recap

12

So what we're going to do now is look at clustering related considerations when our outcome is child growth and our intervention is a nutritional intervention in WASH benefits.

Nutritional intervention in WASH Benefits

Behavior Change

Exclusive breastfeeding through 6 months
Continued breastfeeding until 24 months
Encourage micronutrient dense food

www.aliveandthrive.org

\+

Nutritional Supplement

Nut-based daily supplement 6 – 24 months
118 kcal/day + fatty acids + micronutrients

Slide created by Ben Arnold

13

So the nutritional intervention WASH benefits had several components. But one component was behavior change, which meant to encourage the mothers to use exclusive breastfeeding through the first six months of life and then to continue breastfeeding until 24 months, and also to encourage micronutrient-dense food.

Additionally, there was a nutritional supplement that was a nut-based daily supplement given for months six to 24 that had a low number of calories-- 100 calories per day-- along with fatty acids and micronutrients. So it's not a complete nutritional replacement. It's a supplement.

Nutritional interventions and length-for-age

- A primary outcome in the Bangladesh trial was a child's length-for-age Z-score (LAZ) measured 2 years after intervention
  - A child's length (height) is mapped to international standards based on age and sex. The score is continuous; LAZ < −2 indicates the child is stunted.
- **Stunting** is low height-for-age (ie, low LAZ) –> often due to poor nutrition
  - **Wasting** is low weight-for-height –> often due to periods of insufficient food intake
- Stunting by age 24 months is associated with life-long deficits in health and human capital (Victora 2008)
- A testable hypothesis: The nutritional intervention would improve LAZ compared to standard practices (control group) after 24 months of intervention.

Slide created by Ben Arnold 14

So one of the primary outcomes in the Bangladesh trial was a child's length-for-age z-score measured two years after intervention. A child's length-- or height-- is mapped to international standards based on age and sex. The score is continuous. And an LAZ less than negative two indicates that the child is stunted.

Stunting is low height for age, that is a low LAZ, which is often due to poor nutrition. Wasting is a different outcome. That's low weight for height. And that's often due to periods of insufficient food intake.

Stunting by age 24 months is associated with lifelong deficits in health and human capital. So our testable hypothesis here is that our WASH benefits nutritional intervention would improve the LAZ compared to standard practices in the control group after 24 months of intervention.

Sample size steps in this example

1. Specify the parameter of interest and H_0, H_a
2. Obtain measures of outcome variability and the ICC
3. Calculate preliminary sample size and MDE estimates
4. Check the budget and logistics

   (iterate steps 3 + 4 multiple times, potentially with multiple outcomes)
5. Settle on a final sample size / design
6. Calculate the minimum detectable effect size

Slide created by Ben Arnold 15

Here are the steps to use in thinking through the clustered sample size in WASH Benefits. First we're going to specify the parameter of interest and what our null hypothesis and alternative hypotheses are. We want to have some measures of outcome variability in the intraclass correlation coefficient. And these could come from a pilot study or from other published data.

Then we're going to calculate preliminary sample size and minimum detectable effect estimates. Then we have to iterate a bit, because we need to check our budget and logistics to see if we can afford to do what our equations seem to be telling us we might want to do. Eventually, we'll settle on a final sample size and design. And then we'll calculate the minimum detectable effect size for that sample size and design that we've settled on.

Step 1: Identify the parameter of interest & hypotheses

- We were interested in estimating the difference in LAZ between the nutrition arm and the comparison arm.
- Our parameter of interest, θ, was the mean difference between groups:

$$\theta = E(Y \mid A = 1) - E(Y \mid A = 0)$$

where Y is LAZ and A is a dichotomous indicator equal to 1 if a child received the nutrition intervention and 0 if a child was in the control arm.

- Our null hypothesis, H_0, was that $\theta = 0$ (no effect).
- Our alternative hypothesis, H_a, was that $\theta > 0$ (intervention beneficial).

Slide created by Ben Arnold 16

In step one we want to identify the parameter of interest and our hypotheses. In our study we were interested in estimating the difference in LAZ between the nutrition arm and the comparison arm. Our parameter of interest, theta, was the mean difference between these two groups.

So we could write our parameter of interest as theta is the expectation of y given that a equals one minus the expectation of y given that a equals zero. This is looking at the group a equals one that has the intervention minus the group that doesn't have the intervention. And the y we're looking at is the length-for-age z-score.

Our null hypothesis was that theta was equal to zero, that this difference between the two groups would be zero. That is that there's no effect. So we want to see, as we learned before, whether we can falsify this null hypothesis. Our alternative hypothesis was that theta was greater than zero, that is that the intervention was beneficial.

Step 2: Obtain measures of outcome variability and the ICC

- We had LAZ measurements from 982 children < 3 years old from rural Bangladesh that were part of an existing cohort. (Huda 2012)
- To estimate the variability of LAZ you can calculate the standard deviation in R using summarise function in the dplyr package

```
anthro %>% summarise(sd_laz = sd(laz, na.rm=TRUE))

sd_laz
1.24
```

Slide created by Ben Arnold 17

The second step in our process is to obtain measures of outcome variability and the intraclass correlation coefficient, or rho. So in other work we had LAZ measurements from 982 children under three years old from rural Bangladesh that were part of an existing cohort.

So to estimate the variability of length-for-age z-score, we can calculate the standard deviation. And if you do this in r using the summarize function in the DPLYR package that you've seen-- so here's the r code for this. What it returns is the variability for the length-for-age z-score of the standard deviation for the length-for age z-score.

Step 2: Obtain measures of outcome variability and the ICC

- and estimate the ICC using the ICC package in R.

```
install.packages("ICC", dependencies = TRUE)
library(ICC)
ICCest(x = clusterid, y = laz, data = anthro)
$ICC[1] 0.008
```


Next we want to estimate rho using the ICC package in r. You see that here.

Steps 3, 4 & 5: Calculate preliminary sample size and MDE estimates, check budget/logistics, iterate

- In the case of this study, we repeated MDE calculations for a range of designs and settled on the following:
  - A single post-treatment measurement of LAZ after 2 years of intervention
  - 7 newborns per cluster (fixed by demographics)
  - 90 clusters per arm
  - A double-sized control arm (due to multiple treatment comparisons)

Slide created by Ben Arnold 19

In our next steps we want to calculate the preliminary sample size and the minimum detectable effect estimates. Then we want to check our budget and logistics and iterate these steps a few times.

So in this study, we repeated the MDE calculations for a range of designs, and settled on the following. We decided to do a single post-treatment measurement of LAZ after two years of intervention. We wanted to enroll seven newborns per cluster. This was determined by the demographics.

We determined we needed 90 clusters per arm. And we wanted a double-size control arm due to multiple treatment comparisons that we were going to make, as we've talked about before. All the different in other arms we're going to be compared against the control group.

Step 6 : Given a design, calculate the minimum detectable effect size

$$
\begin{aligned}
MDE &= (Z_{1-\beta} + Z_{1-\alpha})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{k}} \times \sqrt{\rho + \frac{1-\rho}{m}} \\
&= (0.84 + 1.64)\sqrt{\frac{1}{0.33 \times (1-0.33)}}\sqrt{\frac{1.24^2}{270}} \times \sqrt{0.008 + \frac{1-0.008}{7}} \\
&= 0.154
\end{aligned}
$$

- Note: $k = 270$ and $P = 0.33$ because we are comparing 1 treatment arm (90 clusters) to a double-sized control arm (180 clusters)
- The design will be able to detect a difference of $+0.15$ LAZ with 80% power and a one-sided $\alpha = 0.05$ for the comparison of any treatment arm to the control arm. At age 24 months, this is equivalent to ≈ 0.45 cm.

Slide created by Ben Arnold 20

Given the design features we discussed on the prior slide, we want to calculate the minimum detectable effect size. So you see the algebra for it here. Let's talk about the various pieces that are plugged in.

So for power we chose 80% power. And on the z-scale, on the normal distribution, that's represented by a z-score of 0.84. And for our one-sided test we chose an alpha represented on the z-scale by 1.64.

And then let's look a little bit at the denominator p times one minus p. We said we wanted a double-sized control arm. So in this comparison, in the one arm it's 33%. And the other arm is 67%. So that's a one to two ratio. So that's a double-sized control arm compared to an intervention arm.

Sigma squared is 1.24 squared. And we saw earlier where we calculated that variability. The number of clusters here is 270, or k. And some of this, of course, is driven by budget and logistics. You choose and then you play around with this to see what it does as you change this around.

Step 6 : Given a design, calculate the minimum detectable effect size

$$
\begin{aligned}
MDE &= (Z_{1-\beta} + Z_{1-\alpha})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{k}} \times \sqrt{\rho + \frac{1-\rho}{m}} \\
&= (0.84 + 1.64)\sqrt{\frac{1}{0.33 \times (1-0.33)}}\sqrt{\frac{1.24^2}{270}} \times \sqrt{0.008 + \frac{1-0.008}{7}} \\
&= 0.154
\end{aligned}
$$

- Note: $k = 270$ and $P = 0.33$ because we are comparing 1 treatment arm (90 clusters) to a double-sized control arm (180 clusters)
- The design will be able to detect a difference of $+0.15$ LAZ with 80% power and a one-sided $\alpha = 0.05$ for the comparison of any treatment arm to the control arm. At age 24 months, this is equivalent to ≈ 0.45 cm.

Slide created by Ben Arnold 20

And then we know rho is 0.008. And so the rest of that calculation under the square root is shown to you there. Plus one minus rho over m, the number of children in each cluster which was fixed by the area in which we were working. That was what we felt was reasonable to get in the area in which we were working, seven children.

So the MDE is .154. So this means the design would be able to detect a difference of 0.154 length-for-age z-score with 80% power and a one-sided alpha of 0.05 for the comparison of any treatment arm to the control arm. So if you convert this LAZ score, which is a score on a normal distribution to actual centimeters, that works out to about a difference in groups of 0.45 centimeters, which is a fairly substantial growth difference on average between groups.

Very easy to implement in R!

```
#--------------------------------------
# mde function
# k:     total number of independent units (clusters)
# m:     average number of repeated measures per unit
# sd:    standard deviation of the outcome in the population
# rho:   intra-class correlation of the outcome within units
# P:     proportion of clusters allocated to treatment (optimal = 0.5)
# alpha: Type-II error rate (for a one-sided test, double the alpha)
# power: 1 - Type II error rate
#--------------------------------------
mde <- function(k,m,sd,rho,P=0.5,alpha=0.05,power=0.8) {
  Za <- qnorm(1-(alpha/2))
  Zb <- qnorm(0.8)
  return( (Za+Zb) * sqrt(1/(P*(1-P)))*sqrt(sd^2/k) * sqrt(rho+(1-rho)/m) )
}
# one-sided hypothesis test, consistent with the example (double alpha from 0.05 to 0.1)
mde(k=270,m=7,sd=1.24,rho=0.008,P=0.33,alpha=0.1,power=0.8)
[1] 0.1544049
# two-sided hypothesis test
mde(k=270,m=7,sd=1.24,rho=0.008,P=0.33,alpha=0.05,power=0.8)
[1] 0.1739726
```

Slide created by Ben Arnold 21

So this is all very easy to implement in r. And we show you the code for that here allowing you to change the different parameters and decide what your different minimum detectable effect sizes are with different parameters put into your calculations.

Example of an intermediate step (alternate scenarios):

- Considering alternate designs with a fixed total N (1,890), but different allocation:

Children per cluster (m)	Total clusters (k)	MDE (d)
7	270	0.154
10	189	0.156
14	135	0.158
21	90	0.162
30	63	0.167

- Not much increase in MDE from enrolling fewer, larger clusters so this would be a preferred strategy if logistics permit (in this real example, we were limited by demographics and timeline)
- Note: since ρ = 0.008, our rule of thumb suggests we could enroll up to 1/0.008 = 125 children per cluster without a large loss in power

Slide created by Ben Arnold 22

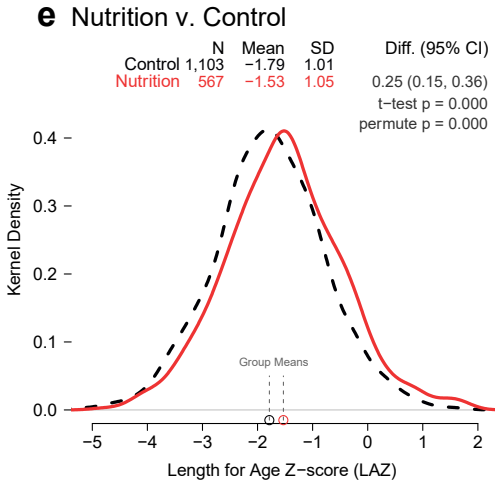
So here's the result of some of the runs with r. If we change the children per cluster from seven to 10 to 14 to 21 to 30, you see the impact on our minimum detectable effect size as the number of clusters changes as well.

In our study we were able to support a study that would find the smallest minimal detectable effect size, which is the 0.154, the seven children per cluster with lots of clusters, 270. Notice that there's not a lot of increase in MDE from enrolling fewer larger clusters. So this would be a preferred strategy. The logistics required it. But we were limited by demographics and timeline because, as I said, we didn't think we could easily get more than seven children per cluster.

And note since the rho is equal to 0.008, our rule of thumb suggests we could enroll up to one divided by 0.008, or 125 children per cluster without a large loss in power.

… and in the end: WASH Benefits Nutrition results

Luby et al. 2018



- Assumptions were slightly conservative
- SD was 1.01 rather than 1.24
- Actual MDE was about 0.11, slightly smaller than the designed MDE of 0.15

- Nutrition intervention improved LAZ by +0.25
  - Still very far from removing all growth faltering (mean LAZ = -1.53)

Slide created by Ben Arnold 23

So after all was said and done and we conducted the study, here are the results that we published. So this is comparing just the nutrition arm against the control arm. Of course the full paper has many other comparison of other arms against the control arm. But this is just the nutrition arm in red against the control arm in black.

And what you see here is this is a very statistically significant difference between these two curves. The t-tests suggest a highly significant result, less than 0.000-- so less than a chance in 1,000 that this happened by just randomness.

So our assumptions were slightly conservative. The standard deviation turned out to be 1.01 rather than 1.24, which works in our favor. Think about why that is. The actual minimum detectable effect was about 0.11-- slightly smaller than the design that we set up for an MDE of 0.15.

So the nutrition intervention improved LAZ 0.25 z-scores. But this is still very far from removing all the growth deficit in these children where they start with a mean LAZ of minus 1.53 when they're born. So improving it by 0.25 the z-scale is good. But it certainly doesn't overcome all the altering that's present at birth.

Presentation overview

- Big-picture concepts related to clustering
- Clustering-related considerations & nutritional intervention / child growth case study
- Additional issues and recap

24

Let's just summarize some additional issues and then recap what we've done.

Summary of Key Points (1)

- Sample size calculations require that you specify:
  1. basic design (fraction treated/exposed), parameter of interest, and hypothesis
  2. outcome variability (get this automatically for binomial outcomes)
  3. minimum detectable effect
  4. desired level of power and Type I error (alpha)

- Additionally, for clustered designs you also need to specify:
  5. number of units per cluster
  6. the cluster level ICC

Slide created by Ben Arnold 25

So sample size calculations require you to specify the basic design, which includes the fraction treated and exposed, the parameter of interest, and the hypothesis. You have to specify the outcome variability. Now, if you're working with a binary outcome you can just use a formula for that.

You need to specify a minimum detectable effect-- that's the MDE. And then you have to say what your desired level of power and your type one error, or alpha, are going to be.

Additionally, for clustered designs you also need to specify how many units per cluster-- remember that was seven in our study because that's the number of children we thought we could comfortably get in each cluster. And then you'd have to specify the cluster level intraclass correlation coefficient, which as you recall, was represented by rho.

Summary of Key Points (2)

- The power of clustered designs is most sensitive to the number of clusters per arm and the ICC.
- Even if the ICC is small, there can be large design effects if cluster size is large
- Use estimates of variability and the ICC from similar populations to your study.
  - If estimates are not available to you, calculate sample size and the MDE over a range of plausible values for these parameters.

Slide created by Ben Arnold 26

The power of using a clustered design is most sensitive to the number of clusters per arm and the ICC. And even when the ICC is small there can be large design effects in situations where the cluster size is large.

We use estimates of variability in the ICC from similar populations to our study. So remember, we had that other study where we had a lot of children-- almost 1,000 children-- where we could get the ICC and the variability.

And if these estimates aren't available to you, you have to calculate the sample size in the MDE over a range of plausible values for these parameters. So you could set up a table and keep track of different assumptions you might make about what these are, and see their impact on your sample size and power.

And details of everything we've covered here are in this reference bank.

References

- Arnold, B.; Hogan, D.; Colford, J. & Hubbard, A. Simulation methods to estimate design power: an overview for applied research. BMC Med Res Methodol, 2011, 11, 94
- Arnold, B. F.; Null, C.; Luby, S. P.; 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
- Colford JM, Schiff KC, Griffith JF, Yau V, Arnold BF, Wright CC, et al. Using rapid indicators for Enterococcus to assess the risk of illness after exposure to urban runoff contaminated marine water. Water Res. 2012;46: 2176–2186.
- Colford, J. M.; Wade, T. J.; Schiff, K. C.; et al. Water quality indicators and the risk of illness at beaches with nonpoint sources of fecal contamination. Epidemiology, 2007, 18, 27-35
- Campbell, M. J.; Donner, A. & Klar, N. Developments in cluster randomized trials and Statistics in Medicine. Stat Med, 2007, 26, 2-19
- Duflo, E.; Glennerster, R. & Kremer, M. Using Randomization in Development Economics Research: A Toolkit. 61 Handbook of Development Economics, 2007, Volume 4, 3895-3962
- Feiveson, A. H. Power by simulation. Stata Journal, 2002, 2, 107-124
- Leon, A. C. Sample-size requirements for comparisons of two groups on repeated observations of a binary outcome Eval Health Prof, 2004, 27, 34-44
- Murray, D. M.; Varnell, S. P. & Blitstein, J. L. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health, 2004, 94, 423-432
- Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. Int J Epidemiol. 2015;44: 1051–1067.
- Schulz, K. F. & Grimes, D. A. Sample size calculations in randomised trials: mandatory and mystical. Lancet, 2005, 365, 1348-1353
- Selvin, S. Statistical Analysis of Epidemiologic Data (2nd ed). Oxford University Press, 1996
- Victora, C. G.; Adair, L.; Fall, C.; Hallal, P. C.; Martorell, R.; Richter, L.; Sachdev, H. S. Maternal and child undernutrition: consequences for adult health and human capital. Lancet, 2008, 371, 340-357
- Victora, C. G.; de Onis, M.; Hallal, P. C.; Blossner, M. & Shrimpton, R. Worldwide timing of growth faltering: revisiting implications for interventions. Pediatrics, 2010, 125, e473-e480

Sample size calculations for the design of health studies: a review of key concepts for non-statisticians

***Alistair Merrifield*[A,C] *and Wayne Smith*[B]**

[A]*Centre for Epidemiology and Evidence, NSW Ministry of Health*
[B]*Environmental Health Branch, NSW Ministry of Health*
[C]*Corresponding author. Email: amerr@doh.health.nsw.gov.au*

Abstract: Sample size calculations before conducting a health study or clinical trial are important to provide evidence that the proposed study is capable of detecting real associations between study factors. This review aims to clarify statistical issues related to the calculation of sample sizes and is illustrated with an example of a recent study design to improve health outcomes related to water and sewage in NSW Aboriginal communities. The effect of power, significance level and effect size on sample size are discussed. Calculations of sample sizes for individual-based studies are modified for more complex trial designs by multiplying individual-based estimates by an inflationary factor.

Sample size calculations are an important consideration when designing a health study.[1,2] Investigators need to provide suitable calculations to ensure that a study is capable of detecting a real effect due to an intervention. While there are articles available to assist researchers who have some statistical background with sample size calculations,[2] there are few available for those with limited statistical knowledge. This review is based on a literature review of relevant articles that the authors have found useful. It provides a background understanding for the researcher to be able to more easily communicate with the statistician during the sample size calculation process. We introduce important concepts in a clear and non-technical account to a reader who is uneasy with basic statistics. Suitable references will be given to enable the interested reader to go beyond the scope of this review.

While studies may be conducted to examine differences between treatment groups or to estimate some population statistic,[1] here we focus on the former. We introduce the reader to the steps involved in calculating a sample size for an individual-based randomised control trial with treatment and control groups and a binary outcome (two categories). These principles apply to other types of outcomes. The review also discusses the calculation of sample sizes for more complex study designs.

Calculation of sample sizes for studies in which individuals are randomised

Three fundamental factors are involved in calculating sample sizes: significance level, power and effect size (defined in Table 1). We recommend Kirby et al. for a more detailed discussion.[2] When consulting a statistician for a sample size calculation, a researcher can help assist the process with a knowledge of these three parameters. Various sample size calculators are available online, which further explain the relationship of these three components to sample size (these tools should be used with appropriate statistical advice).[3]

The process for calculating a sample size is:[4]

1. Specify the null and alternative hypotheses, power, effect size and significance level.
2. Define the study population.
3. Estimate the required parameters (e.g. means, standard deviations) from the available data. These estimates are often derived from pilot studies and literature searches.
4. Calculate a range of sample sizes for a range of parameters (to provide different scenarios).
5. Choose the most appropriate sample size from these scenarios, given the study constraints.

Example

A proposed study to examine the intervention of improved water and sewage on health outcomes in discrete NSW Aboriginal communities (Aboriginal Communities Water and Sewage Program Health Outcomes Evaluation) provides an illustration of sample size calculations. The health outcome under consideration is the presence of intestinal infections. The measure for the study is expressed as a relative risk (RR), which is the ratio of the probability of intestinal infections in the Aboriginal communities before and after the intervention. Sample size formulae for binary outcomes (presence or absence of

10.1071/NB11017

Table 1. The fundamental components of sample size estimation

Component	Definition	Example
Null hypothesis	A statement that the intervention has no effect (treatment groups are equivalent), defined in terms of an appropriate measure calculated for the treatment and control groups.	Examples include differences in means or probabilities, relative risks and hazard ratios.
Significance level	The significance level (α) is defined as the chance that the study will incorrectly report that the two treatment groups differ when they are equivalent (Type I error, false positive).	Typical values of α include 5% and 1%. If the study (at the 5% level) was rerun 20 times, we expect to incorrectly reject the null hypothesis once.
Power	Power is defined as the chance that the study will correctly report that the two treatment groups differ. The power is the chance that the study will **not** make a Type II error (a false negative).	Common values of power include 80% and 90%. In practice, power and significance level involve trade-offs with one another. Increasing power will come at the cost of a higher significance level.
Effect size	The **alternative hypothesis** is the hypothesis that the two treatment groups differ by at least some pre-specified amount. This amount is the effect size (δ), the detectable difference between the two treatment groups.	

Table 2. Total (treatment and control) sample sizes for various effect sizes for studies in which individuals are randomised, assuming the probability of intestinal infection before intervention to be 0.051 and equal numbers in the two groups, using the Housing for Health study*

	Effect size			
	Worst case	Housing for Health intervention*	Intermediate case	Best case
Effect size (reduction)	20%	43%	50%	60%
Relative risk	0.80	0.57	0.50	0.40
Power = 80%, α = 10%	10 798	2154	1550	1034
Power = 80%, α = 5%	13 604	**2686**	1928	1280
Power = 80%, α = 1%	20 052	3912	2796	1844
Power = 90%, α = 10%	14 806	2914	2090	1384
Power = 90%, α = 5%	18 078	3536	2530	1670
Power = 90%, α = 1%	25 440	4932	3518	2314

α: significance level.
Note: sample sizes are rounded up to be conservative.
*Closing the gap: 10 years of Housing for Health in NSW. NSW Health 2010.

intestinal infections) are given in Wittes and Campbell et al. (with formulae for other situations).[4,5]

Sample size calculations are based on a set of assumptions. For this example we assume that information from the previous Housing for Health in NSW study[6] holds true for our proposed study. From this study, we estimate the probability of intestinal infection before the intervention as 0.051. We assume that there are equal numbers of people in both the treatment and the control groups, significance level 5% and power 80%. We alter effect size (the difference between probabilities before and after intervention) to give us a range of sample sizes corresponding to different scenarios. In addition to the reduction in the prevalence of intestinal infections of 43% seen in the Housing for Health in NSW study, we also present a worst case of a 20% reduction, an intermediate reduction of 50% and a best reduction of 60%. The resulting sample sizes for the Aboriginal Communities Water and Sewage Program Health Outcomes Evaluation are calculated from formulae 7B and 7C in Wittes[4] (Table 2). From Table 2, we see that the smaller the detectable difference, the larger the sample size required (if all other parameters are held constant). The Housing for Health in NSW study reported a relative

reduction of 43%; the probability of intestinal infection after the intervention is $(1-0.43)\times 0.051 = 0.029$. The absolute effect size is $0.051 - 0.029 = 0.022$. The corresponding sample size is 2686 (Table 2).

Figure 1 shows the effect that power, significance level and effect size have on sample size. Figure 1(a) shows the relationship of different effect sizes on sample size. A decreasing relative reduction means a smaller difference to be detected between treatment and control outcomes which requires a larger sample size. The effect of significance level is shown in Figure 1(b). Ideally, a study should mistakenly reject a true null hypothesis of no treatment effect as few times as possible. For this to occur, a smaller significance level and consequently a larger sample size are required. Figure 1(c) shows the effect of power on sample size. Increased power means a study is more likely to correctly reject a null hypothesis of no treatment effect and a larger sample size is required. A study with more precise estimates of treatment effects will have higher power and lower significance level; this situation comes at the cost of a larger sample size. We recommend Kirby et al. to describe the relationship of significance level, power and effect size on sample size.[2]

Calculation of sample sizes for studies in which clusters of individuals are randomised

The Aboriginal Communities Water and Sewage Program Health Outcomes Evaluation study is a more complicated design as the community (not the individual) receives the intervention. The intervention is an improved water and sewage program. Such an intervention cannot feasibly be delivered to individuals. The clusters are communities and the intervention is randomised to clusters. The sample size calculation for a cluster study involves calculating the corresponding sample size for an individual study and multiplying this by an inflationary factor to account for

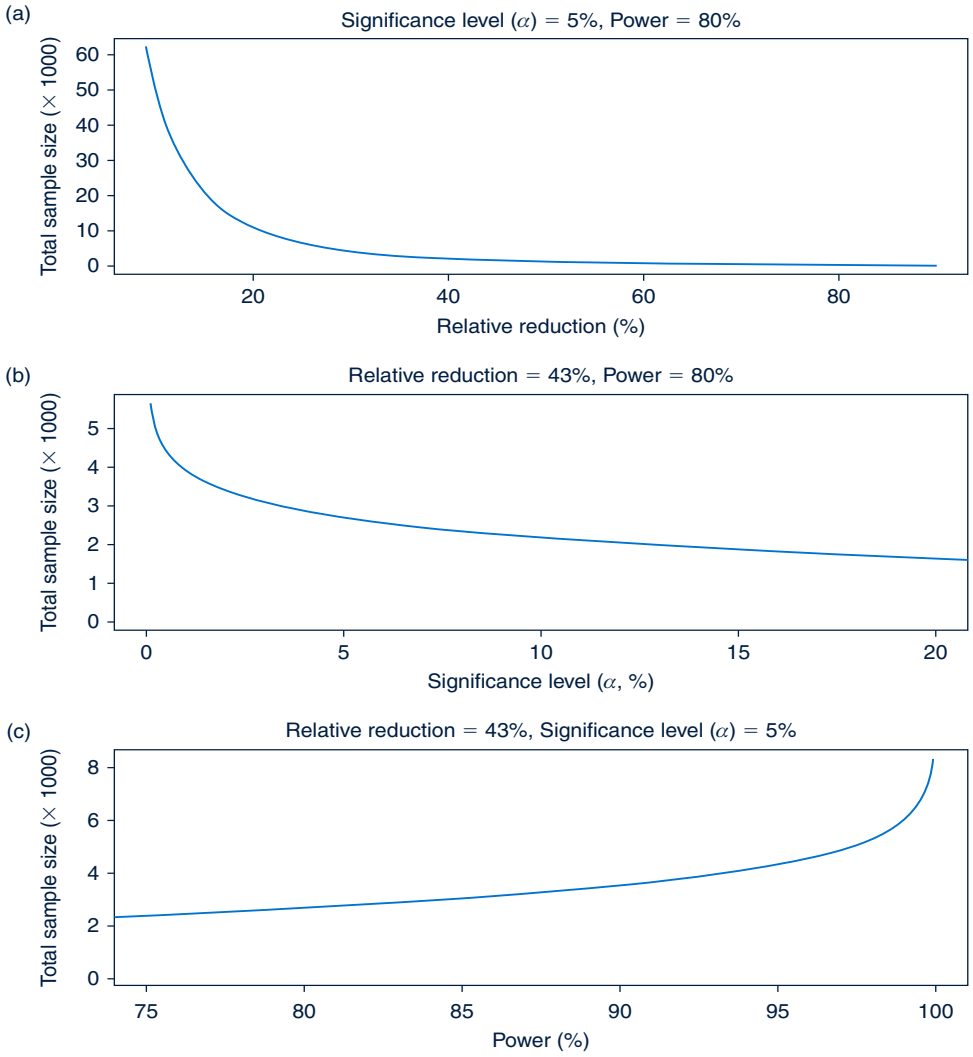


Figure 1. The effect of effect size (a), significance level (b) and power (c) on sample size. Calculations assume the probability of intestinal infection before the intervention to be 0.051 and equal numbers in the two groups.

the more complex trial design.[7–10] This inflationary factor is called the **design effect** (DE). Eldridge et al. provide formulae for design effects for various continuous (e.g. blood pressure, weight) and binary (e.g. whether the patient has the disease or not) outcomes.[7] The estimation of a design effect for cluster randomised control trials involves three factors: mean size of clusters, variation of cluster size and **intra-cluster correlation** (ICC).

The intra-cluster correlation can be regarded as a measure of the degree of similarity in outcomes between clusters.[11] There have been previous papers presenting intra-cluster correlations for different cluster units and populations.[12,13] Appropriate intra-cluster correlations for binary outcomes are discussed in Ridout et al.[14] These outcomes have an associated variance, which can be modelled as two components: variation in outcomes **between** clusters and variation in outcomes **within** each cluster. The intra-cluster correlation is the ratio of the between-cluster variation to total variation (the sum of the between and the within). The intra-cluster correlation is between 0 and 1. Small values of intra-cluster correlation imply that variation within clusters is much greater than variation between clusters and the clustering effect of individuals is less important. If the intra-cluster correlation is zero, outcomes can be regarded as being the same between clusters. The intra-cluster correlation is estimated from available data on cluster sizes and the number of outcomes (intestinal infections) within each cluster.

Example

Information about the size of clusters must be included in our study. We base this on the Housing for Health in NSW study.[6] The clusters are of different sizes and therefore we estimate a mean cluster size and cluster size variation (using standard deviation). From the Housing for Health in NSW study,[6] the mean cluster size = 150.7 and standard deviation = 103.5.

We estimated the intra-cluster correlation using the cluster information from the Housing for Health in NSW study (formula 7 in Ridout et al.[14]). The intra-cluster correlation is estimated as 0.007. We present estimated sample sizes in Table 3. In addition to the reduction in the prevalence of intestinal infections of 43% seen in the Housing for Health in NSW study, we also present a worst case of a 20% reduction, an intermediate of 50% and a best of 60% (assuming 80% power, 5% significance level and equal numbers in groups). We multiply the individual sample sizes presented in Table 2 by the design effect to obtain the estimates in Table 3. From Table 3, the corresponding sample size is 7074.

Figure 2 shows the relationships of intra-cluster correlation and cluster size on sample size. From Figure 2(a), it is apparent that the estimate of the intra-cluster correlation will have a large impact on sample size. As outcomes between clusters become more heterogeneous, the intra-cluster correlation increases. This decreases precision in the resulting outcome estimates from the clusters, and larger samples are thus needed. If the intra-cluster correlation is zero and there is no variation between clusters, the design effect (DE) = 1 and the resulting sample size is equivalent to an individual-level trial size.

Individual-level studies are more efficient than cluster-level studies[7] which is reflected by the larger sample size in response to increased (mean) cluster size shown in Figure 2(b). All other things being equal, an increasing cluster size standard deviation results in increased sample size (Figure 2(c)). Intuitively, increased standard deviation reflects increasing disparity between the size of the clusters. Due to less precise estimates, a larger sample size is required. Trials are more statistically efficient for similar sized clusters and need smaller sample sizes. Larger samples are required for increasing mean and standard deviation.

Table 3. Total (treatment and control) sample sizes for various scenarios for studies in which clusters of individuals are randomised. Corresponding design effects are shown in brackets. Sample sizes are derived from Table 2 (design effect multiplied by sample size with 80% power and 5% significance level, subject to rounding), using the Housing for Health study*

	Worst case	Housing for Health intervention*	Intermediate case	Best case
Effect size (reduction)	20%	43%	50%	60%
Relative risk	0.80	0.57	0.50	0.40
ICC = 0.001 (DE = 1.22)	16 620	3282	2356	1564
ICC = 0.005 (DE = 2.11)	28 680	5662	4064	2698
ICC = 0.007 (DE = 2.63)	35 830	**7074**	5078	3372
ICC = 0.01 (DE = 3.21)	43 754	8638	6202	4116
ICC = 0.05 (DE = 12.04)	164 356	32 450	23 294	15 464
ICC = 0.1 (DE = 23.16)	315 108	62 216	44 658	29 648

ICC: intra-cluster correlation.
DE: design effect.
*Closing the gap: 10 years of Housing for Health in NSW. NSW Health 2010.

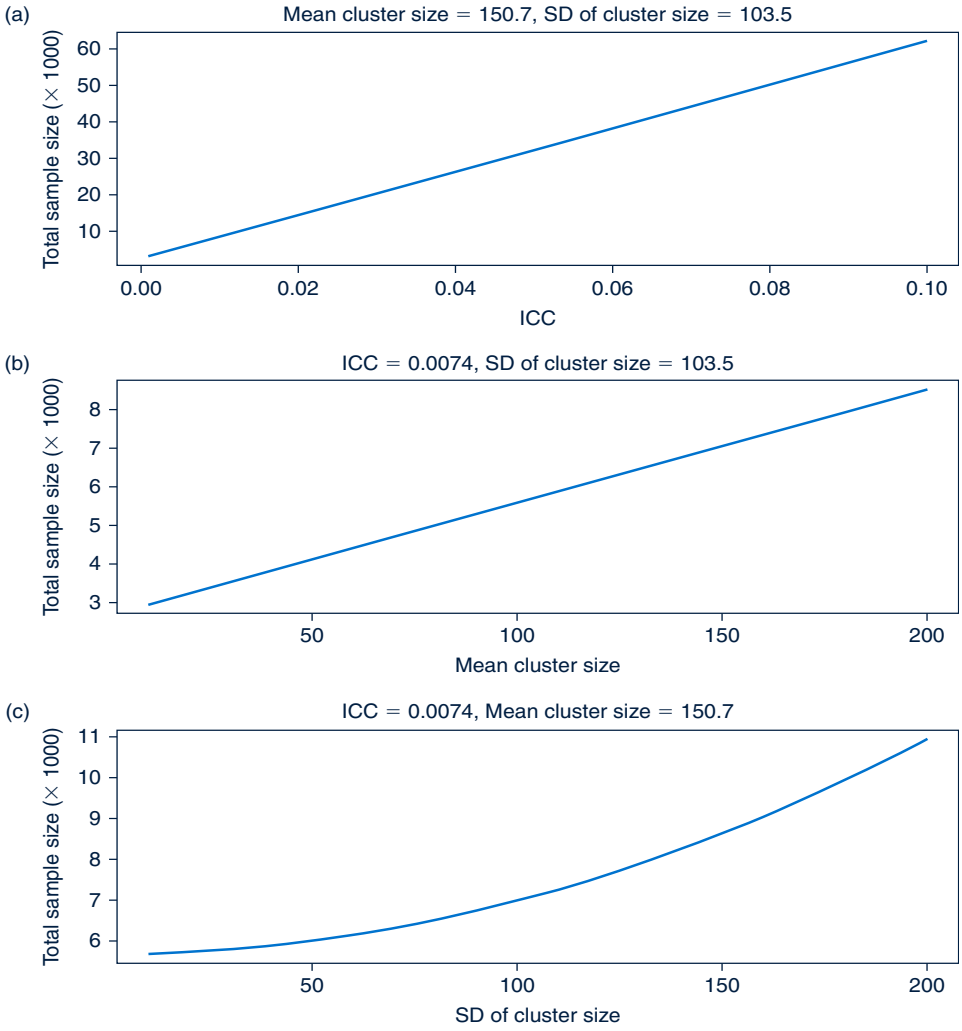


Figure 2. Effect of intra-cluster correlation (ICC) (a), mean cluster size (b) and standard deviation (SD) of cluster size (c) on sample size. Calculations assume the parameter estimates from the Housing for Health in NSW study are correct, reduction of 43%, power = 80%, significance level = 5% and equal numbers in the two groups.

Other factors affecting sample size calculations

There are other important factors that need to be accounted for in sample size calculations, including losses to follow-up, unequal treatment group sizes and the noncompliance of subjects to the intervention.[1,2,4] If the study investigator is able to provide an estimate of these factors to the statistician, the calculation of the required sample size will be improved.

Sample size calculations for more complex study designs can be regarded as multiplying the estimated sample size from an equivalent individual-level study by a design effect. Additional considerations involved in the calculation of this design effect include estimating the intra-cluster correlation and the sizes of the clusters, losses to follow-up and noncompliance.

Discussion

The calculation of sample sizes is based on several parameters; the researcher should at least be aware of power, significance level and effect size. Increased power, smaller significance level and smaller effect sizes translate into larger sample sizes. The researcher and statistician are faced with selecting the most appropriate sample size from an appropriate set of parameters (subject to financial and logistical constraints).

Sample size calculations are an important and complex part of study design and should be discussed by study investigators and statisticians as early as possible during the design of a study design.

Acknowledgment

AM was employed as part of the NSW Biostatistical Officer Training Program funded by the NSW Ministry of Health while undertaking this work based at Environmental Health Branch, NSW Health. We thank the reviewers for their comments.

References

1. Whitley E, Ball J. Statistics review 4: Sample size calculations. *Crit Care* 2002; 6: 335–41. doi:10.1186/cc1521
2. Kirby A, Gebski V, Keech AC. Determining the sample size in a clinical trial. *Med J Aust* 2002; 177: 256–7.
3. Campbell MK, Thomson S, Ramsay CR, MacLennan GS, Grimshaw JM. Sample size calculator for cluster randomised trials. *Comput Biol Med* 2004; 34: 113–25. doi:10.1016/S0010-4825(03)00039-8
4. Wittes J. Sample size calculations for randomised controlled trials. *Epidemiol Rev* 2002; 24(1): 39–53. doi:10.1093/epirev/24.1.39
5. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *BMJ* 1995; 311: 1145–8. doi:10.1136/bmj.311.7013.1145
6. Closing the gap: 10 years of Housing for Health in NSW. An evaluation of a healthy housing intervention. Aboriginal Environmental Health Unit, NSW Health. 2010. Available from: http://www.health.nsw.gov.au/pubs/2010/pdf/housing_health_010210.pdf (Cited 12 August 2011).
7. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomised trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol* 2006; 35(5): 1292–300. doi:10.1093/ije/dyl129
8. Bland JM, Kerry SM. Statistics notes: trials randomised in clusters. *BMJ* 1997; 315: 600. doi:10.1136/bmj.315.7108.600
9. Kerry SM, Bland JM. Statistics notes: sample size in cluster randomisation. *BMJ* 1998; 316: 549. doi:10.1136/bmj.316.7130.549
10. Kerry SM, Bland JM. Statistics notes: analysis of a trial randomised in clusters. *BMJ* 1998; 316: 54. doi:10.1136/bmj.316.7124.54
11. Kerry SM, Bland JM. Statistics notes: the intracluster correlation coefficient in cluster randomisation. *BMJ* 1998; 316: 1455. doi:10.1136/bmj.316.7142.1455
12. Knox SA, Chrondros P. Observed intra-cluster correlation coefficients in a cluster survey sample of patient encounters in general practice in Australia. *BMC Med Res Methodol* 2004; 4: 30. doi:10.1186/1471-2288-4-30
13. Mickey RM, Goodwin GD. The magnitude and variability of design effects for community intervention studies. *Am J Epidemiol* 1993; 137(1): 9–18.
14. Ridout MS, Demétrio CG, Firth D. Estimating intraclass correlation for binary data. *Biometrics* 1999; 55(1): 137–48. doi:10.1111/j.0006-341X.1999.00137.x

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

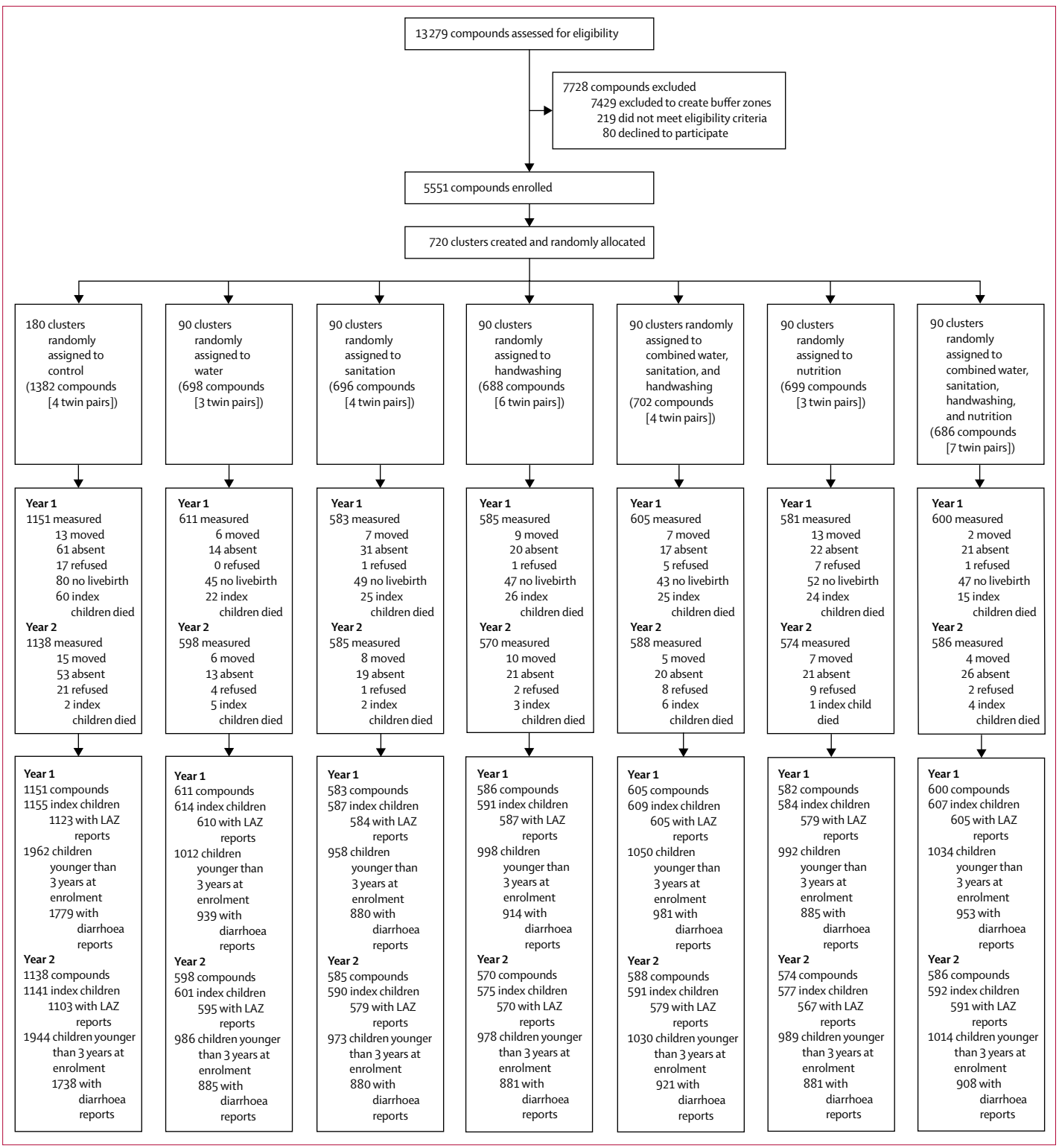


Figure 1: **Trial profile and analysis populations for primary outcomes**
LAZ=length-for-age *Z* scores.

	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 enrolled 5551 pregnant women in 5551 compounds to an

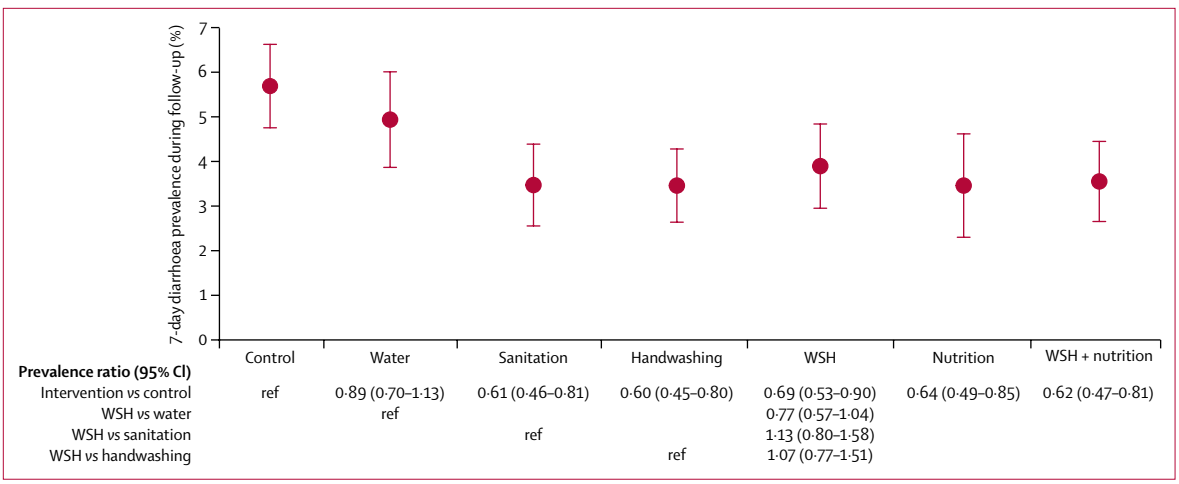


Figure 2: **Intervention effects on diarrhoea prevalence in index children and children younger than 3 years at enrolment 1 and 2 years after intervention**
Data are mean (95% CI). ref=reference. WSH=water, sanitation, and handwashing.

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

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

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

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

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

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

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

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

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

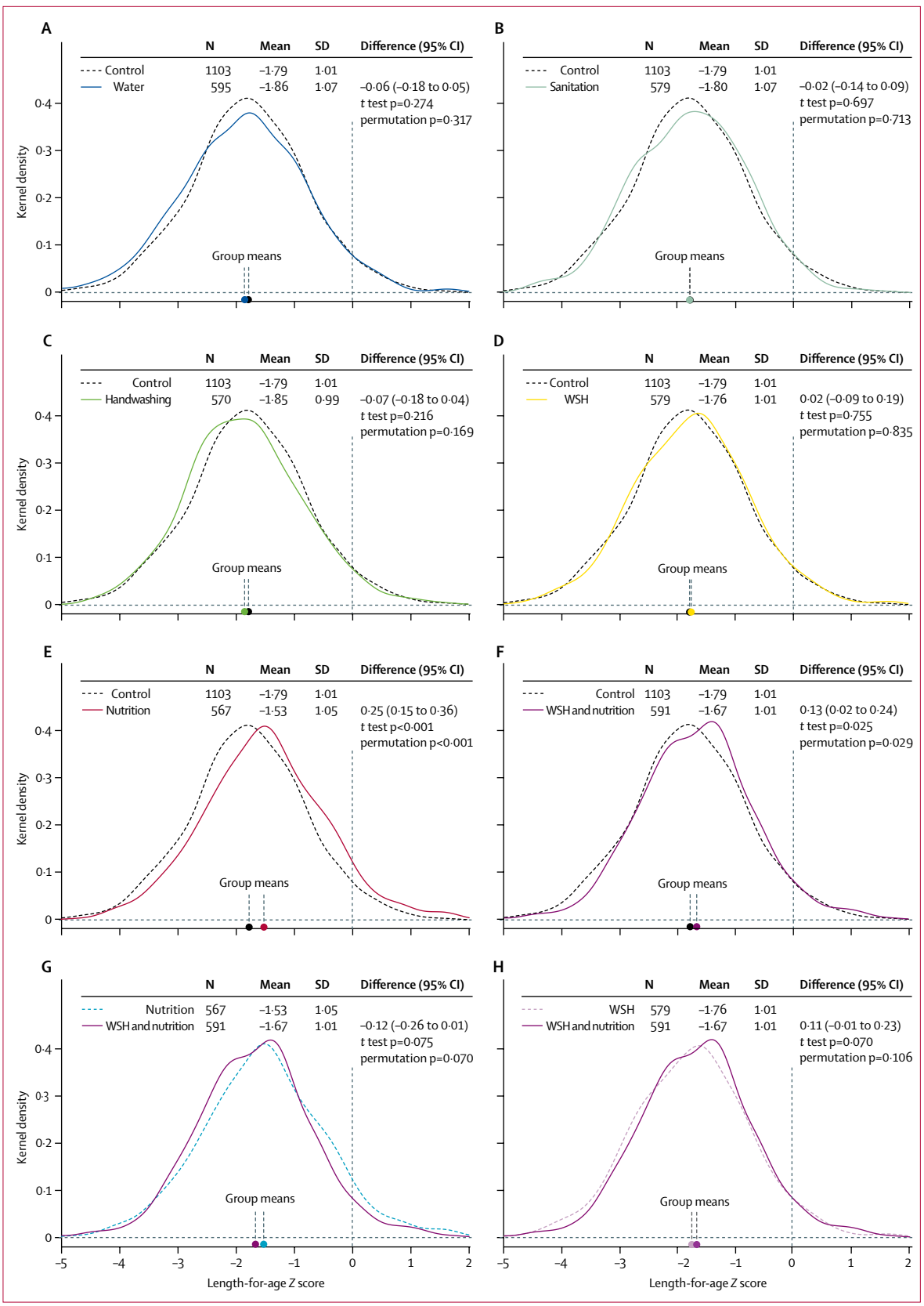


Figure 3: **Intervention effects on length-for-age *Z* scores in 4584 children after 2 years of intervention**

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

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

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

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

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

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

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

All three secondary outcomes were prespecified.

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

Discussion

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

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

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

*Prespecified secondary outcome. †Prespecified tertiary outcome.

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

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

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

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

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

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

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

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

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

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

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

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

Contributors

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

Declaration of interests

We declare no competing interests.

Acknowledgments

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

References

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

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

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

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

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

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

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

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

9 Null C, Stewart CP, Pickering AJ, et al. Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural Kenya: a cluster-randomised controlled trial. *Lancet Glob Health* 2018; published online Jan 29. http://dx.doi.org/10.1016/S2214-109X(18)30005-6.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

American Journal of Epidemiology

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

Original Contribution

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

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

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

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

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

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

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

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

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

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

METHODS

Setting

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

Study design and enrollment

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

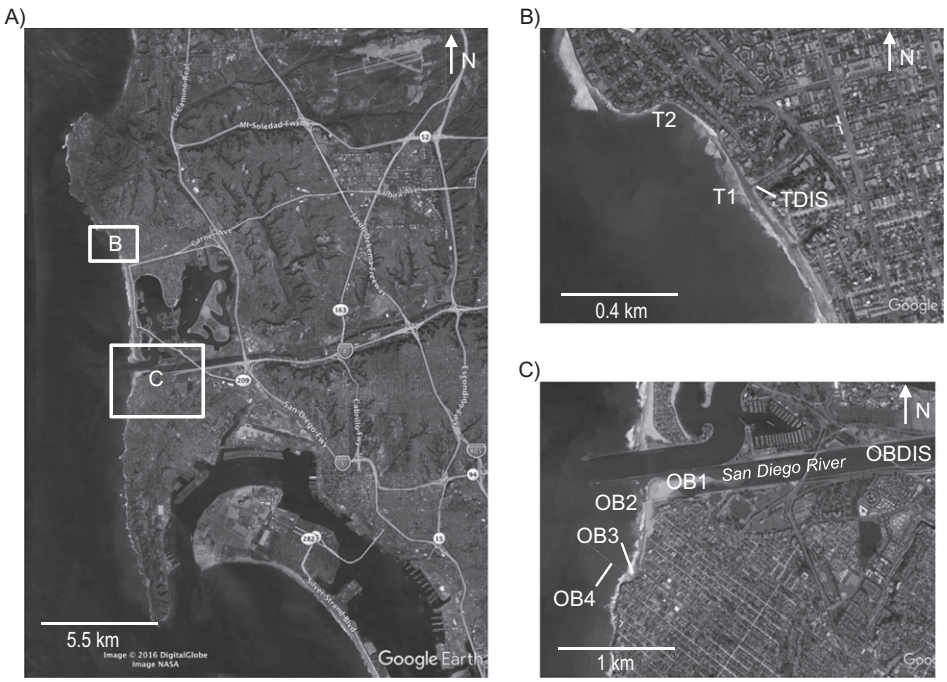


Figure 1. Monitoring beach water quality sampling locations in San Diego, California, winters of 2013–2014 and 2014–2015. Shown are the locations of the 2 monitored beaches along the San Diego coastline (A) and the water quality sampling sites at Tourmaline Surfing Park (B) and Ocean Beach (C). Samples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

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

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

[a] Beach enrollment only took place during the first winter (2013–2014); online enrollment spanned both winters (2013–2014 and 2014–2015). The study enrolled 73 individuals online during the first winter.
[b] Percentages within categories might not sum to 100 because of rounding.

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

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

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

Outcome definition and measurement

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

Exposure definition and measurement

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

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

Statistical analysis

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

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

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

RESULTS

Study population

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

Water quality and surfer exposure

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

Illness associated with seawater exposure

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

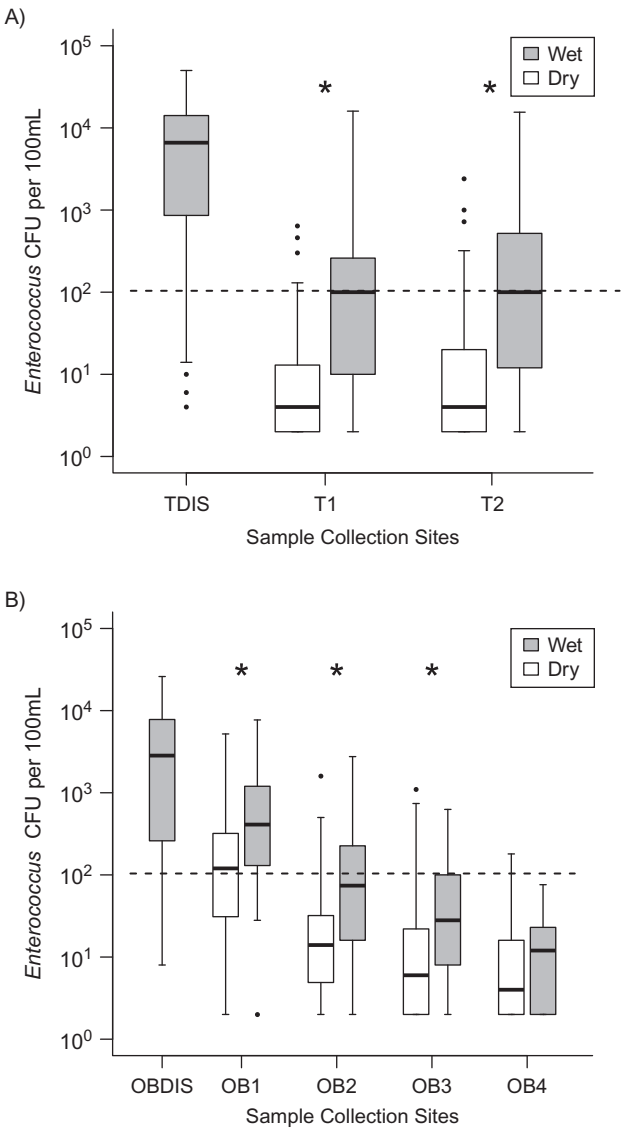


Figure 2. *Enterococcus* levels during dry and wet weather at the sampling locations at Tourmaline Surfing Park (A) and Ocean Beach (B) mapped in Figure 1. Boxes mark interquartile ranges, vertical lines mark 1.5 times the interquartile range, and points mark outliers. Horizontal dashed lines mark the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

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

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

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

Illness associated with fecal indicator bacteria levels

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

DISCUSSION

Key results

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

Interpretation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Limitations

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

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

Conclusions

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

ACKNOWLEDGMENTS

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

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

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

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

Conflict of interest: none declared.

REFERENCES

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

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