PHW250B Week 12 Reader

Topic 1: Matched Designs

Topic 2: Overmatching and Matched Analysis

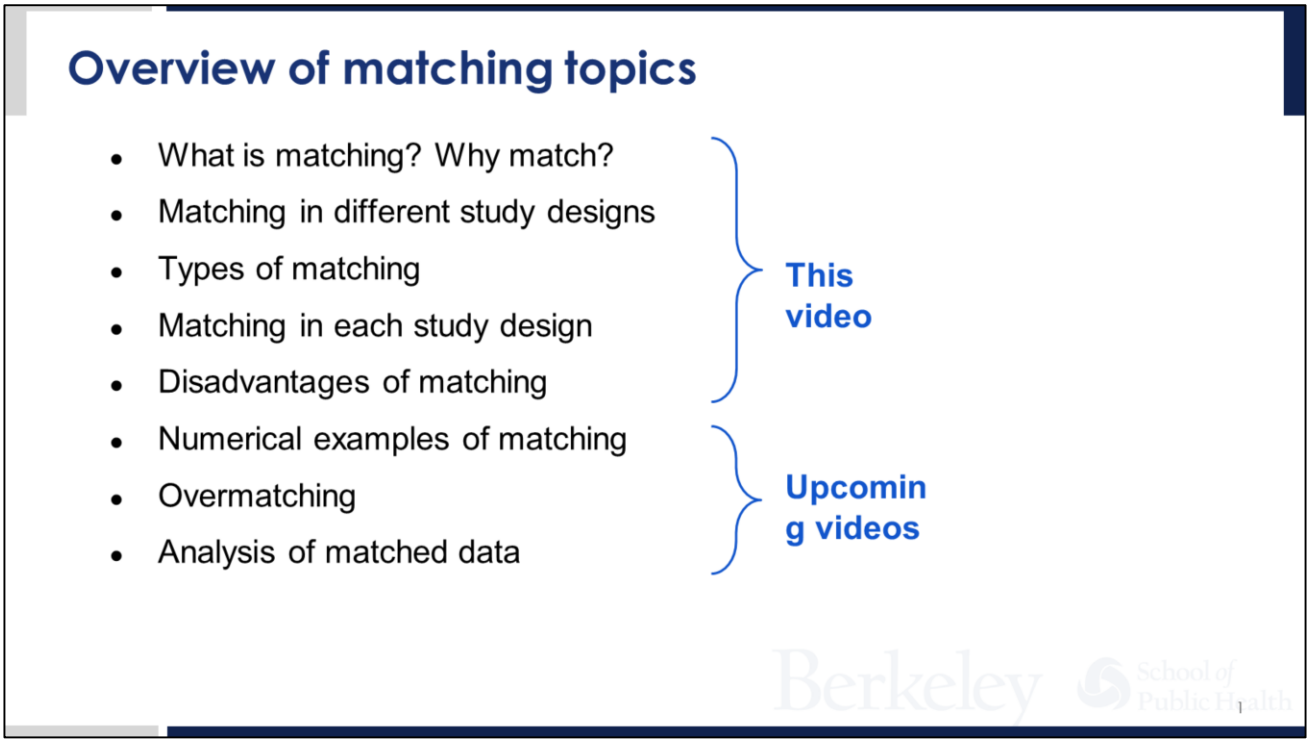
Topic 3: Screening

Podcast

Journal Club

Lecture: Introduction to Matching

JACK COLFORD: There are several ways to try to find a counterfactual group in epidemiology. One important tool is matching, and we're going to begin now a unit on matching to discuss the different applications and details of how epidemiologists use this powerful tool to create counterfactual comparison groups.



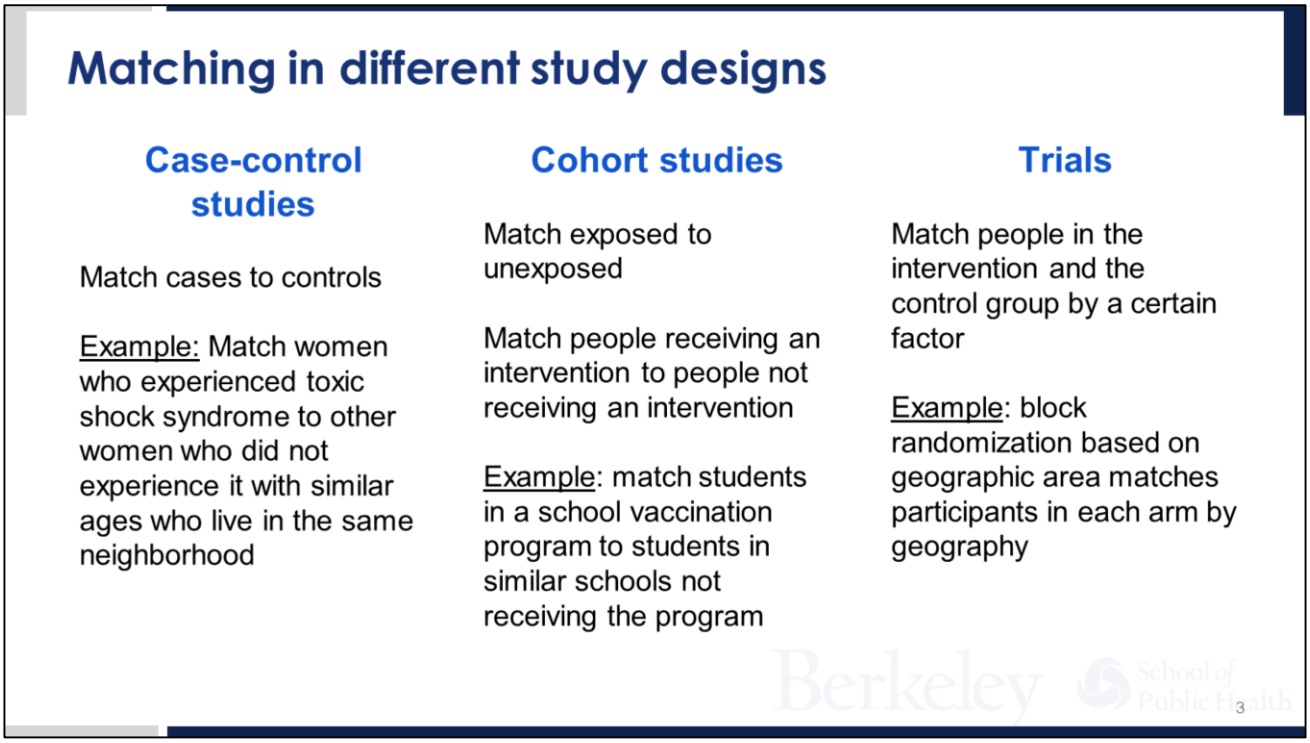
In this video, we'll be talking about a basic introduction to matching and why we match, the different study designs in which matching can be used. Some people think matching is only used in case control studies, but as you'll see, that's not true. There are different types of matching that can be applied to epi studies, and we'll talk about those and the specifics of matching in each of these study designs. And of course, there are some disadvantages to matching, which we'll discuss. All of these are in this video. In later videos, we'll talk about some numerical examples of matching, the concept of overmatching, and how to analyze match data.

What is matching? Why match?

- Selection of a reference series (controls in a case-control study or unexposed in a cohort study) that is nearly identical to the index series with respect to one or more potential confounders.
- Generally the goal is to improve the quality of your counterfactual, whether that is the:
  - Control arm in a trial
  - Unexposed group in a cohort study
  - Controls in a case-control study
- Matching helps reduce bias (and approximate a counterfactual) by making study groups more comparable.
  - (There are other reasons to match too)

Well, first, what is matching, and why should we match? The selection of a reference series-- for example, controls in a case control study as the reference for the cases or the unexposed persons in a cohort study as the reference for the exposed persons-- attempts to create a nearly identical series to the index series with respect to one or more potential confounders. In other words, the control series, whether it be in a case control study or a cohort study, the attempt here is to create a control or comparison group that's as much like the exposed group or the case group as possible.

So generally, the goal is to improve the quality of our counterfactual. And as I said, that could be a control arm in a trial or an unexposed group in a cohort study or the controls in a case control study. Matching helps us reduce bias and, thereby, approximate a counterfactual by attempting to make the study groups as comparable as possible. There are some other reasons to match in special circumstances as well.



So in different study designs, matched individuals or matched participants, matched study subjects are used. Let's look at the different study designs and how matching is used in each of these. So in a case control study, we match the cases to controls. So for example, we could match women who experience toxic shock syndrome to other women who did not experience it with similar ages who live in the same neighborhood. This would be matching on age and neighborhood. In a cohort study, we match the exposed participants, the exposed group, to the unexposed participants or the unexposed group. Here, we match people receiving an intervention to people not receiving an intervention.

Now recall here that the intervention hasn't been assigned by the investigator. The investigator is just observing what has naturally occurred in that one group has the intervention, and one group doesn't. An example here might be to match students in a school vaccination program to students in similar schools not receiving the program. And finally, in trials, particularly randomized trials, we match people in the intervention and then a control group by a certain factor. So this is beyond randomization. This matching is done in addition to the randomization set that's going to occur
.
For example, if we block randomized, based on geographic area matches, participants in each arm by geography, then in each of those areas, we randomize the clusters or the people. So for example, if I create blocks of villages in a country, where I'm doing a large study, I might have eight villages in each block. And if there are eight arms in my study, I would then randomly assign the eight different villages by randomization to the eight different interventions in each block. We'll talk more about this later.

Types of matching

- **Individual:** match subjects together based on individual characteristics (e.g., age, sex, neighborhood) (also called “pair matching”)
- **Frequency:** match the distribution of characteristics between subjects
- **Distance:** match multiple characteristics at once using algorithms that find the distance between these characteristics

There are several types of matching, and three broad categories are individual matching, frequency matching, and distance matching. In individual matching, we match subjects together based on individual characteristics, such as age or sex or neighborhood. Sometimes this is called pair matching, because essentially, we're matching one exposed individual in a cohort study or one case individual in a case control study to just one control or one unexposed individual. So if this is done in a ratio of one to one, that's called one to one pair matching. But we could match more controls to cases. We could have one case and two, three, four, or more controls. But this is still called pair matching, this idea of individual matching.

Frequency matching

- Match the distribution of characteristics between subjects
- The target population must be stratified into levels of the matching factor.
- The study is then conducted within each stratum of the matching factor.
  - E.g., in a case-control study this guarantees that there are both cases and controls within each stratum of the matching factor.
- In case-control studies, after frequency matching, it is no longer possible to estimate the association of the matched factor on the outcome.

So in frequency matching, we match the distribution of characteristics between subjects. The target population is first stratified into levels of the matching factor, and then the study is conducted within each stratum of the matching factor. For example, in a case control study, this would guarantee that there are both cases and controls within each stratum of the matching factor. In case control studies, after frequency matching is done, it's no longer possible to estimate the association of the matched factor on the outcome. And this is hopefully intuitively plausible or understandable to you, because if you've artificially created these strata, then you can no longer study the relationship between that factor and the risk factors of interest that you're studying.

Distance matching

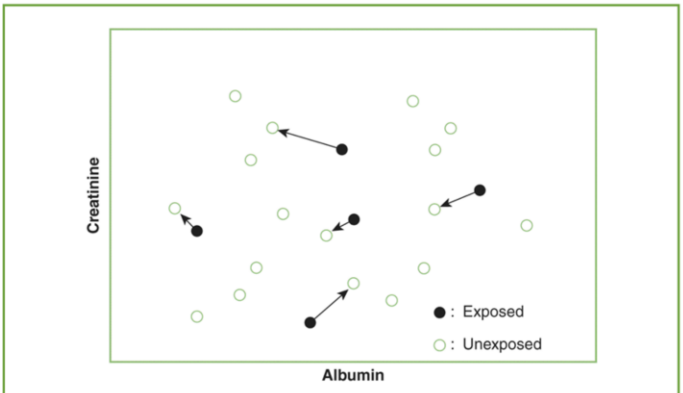


FIGURE 1-24 Matching according to minimal Euclidean distance measure method. Hypothetical example of a cohort study of survival after transplantation in multiple myeloma patients in which exposed individuals (e.g., older individuals) are matched to unexposed (younger) patients according to two prognostic factors: serum albumin and

- Distance measures identify people with the closest combination of multiple matching factors
- Particularly useful with continuous matching factors, and situations with many matching factors
- The figure shows only two potential matching variables (creatinine and albumin) but usually this method is used when there are multiple variables.
- In a case-control study for example, each case is matched to the control with the closest distance in bidimensional space.
- With multiple variables, individuals or groups would be matched in multidimensional space.

Szklo & Nieto, 3rd Ed.

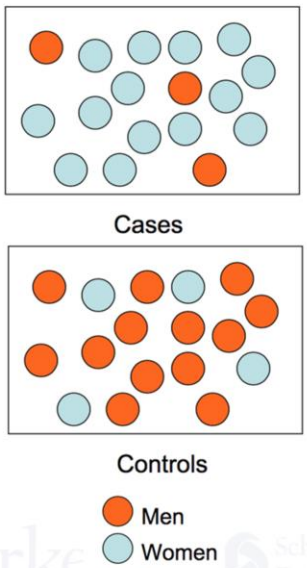
6

In distance matching, we identify people with the closest combination of multiple matching factors. This is particularly useful with continuous matching factors and in situations with many matching factors. In this figure, we're showing just two potential matching factors-- creatinine and albumin. But usually, this method is used when there are multiple variables. In this figure, the exposed individuals are paired up or matched to the individuals with the closest joint albumin creatinine level by plotting the albumin creatinine level of the exposed and unexposed individuals, and then just picking the physically closest or the closest distance on this figure here.

In a case control study, for example, each case is matched to the control with the closest bidimensional space, such as in this albumin creatinine example. And with multiple variables, individuals or groups would be matched in multi-dimensional space. So if you thought of this albumin creatinine figure, and then extend it into a third dimension, say, with hematocrit, that would be yet another axis on which each exposed and unexposed individual would be plotted. And then the matching would be done between the closest distance of each pair in three-dimensional space.

Other reasons to match (besides reducing confounding)

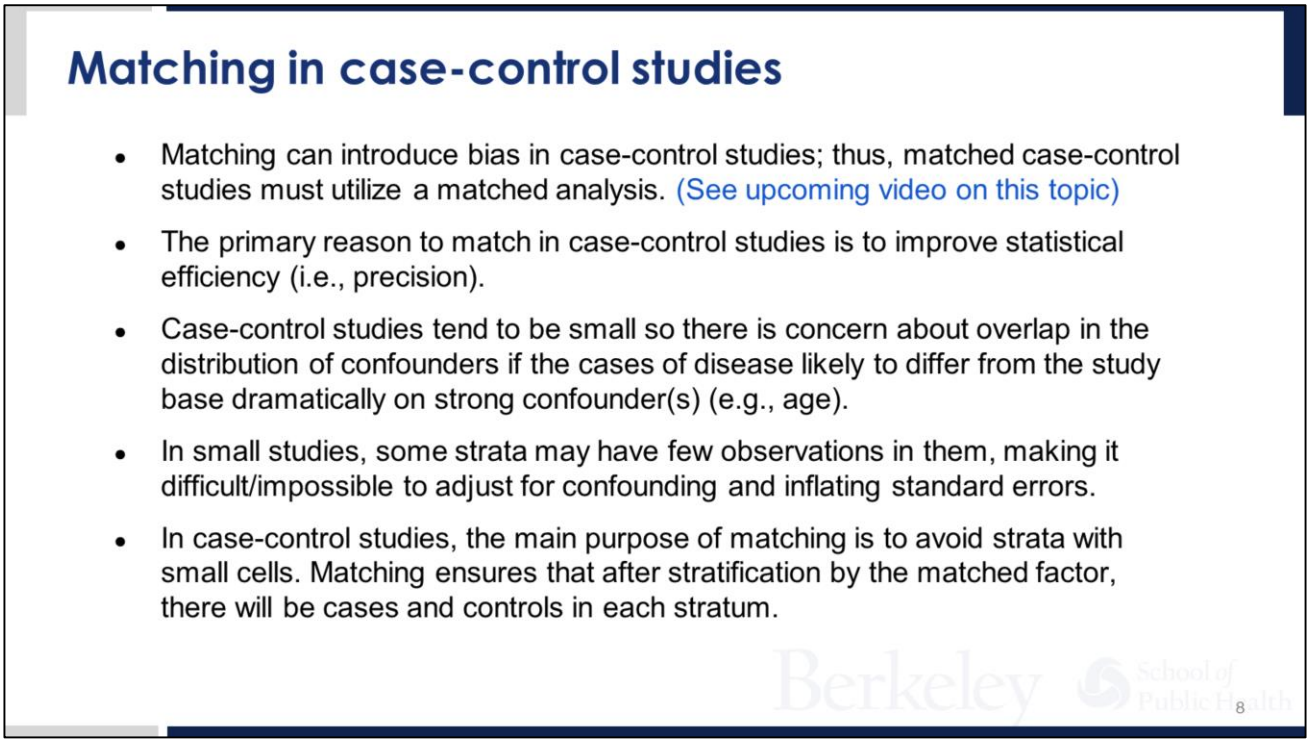
- Increase **statistical efficiency** (precision)
  - Prevent confounder distributions that are dramatically different between cases and controls (or exposed and unexposed), as shown in the figure
- Increase **statistical power**
  - Most common in case-control studies - matching can help there are sufficient controls to estimate associations within important subgroups (e.g., gender)
- **Feasibility** during study enrollment
  - Example: in a study of a perinatal outcome, next birth could be the control (waiting to take a random sample would delay selection etc.



7

There are other reasons to match besides the need to reduce confounding. Matching helps us to increase statistical efficiency or precision by preventing confounder distributions that are dramatically different between cases and controls or between exposed and unexposed. So we balance-- we balance the two groups with respect to the potential confounder. Matching can also increase statistical power. This is most common in a case control study. Matching here can help ensure that there are sufficient controls to estimate associations within important subgroups-- for example, within gender.

And finally, matching is done for feasibility reasons during study enrollment. So for example, in a study of a perinatal outcome, the next birth could be the control, because waiting to take a random sample might delay selection and slow down the study. So it just makes the study more feasible.



Let's talk about matching in case control studies. It's possible for matching to actually introduce bias in case control studies. Of course, that's not a good thing. So in order to use matching in a case control study, in many situations, we have to use a very specific analysis called a matched analysis. And we'll talk about this in a later video. The primary reason that we still do want a match often in case control studies is to improve statistical efficiency or precision.

Because case control studies tend to be small, there's concern about overlap in the distribution of confounders. The cases of disease are likely to differ from the study base dramatically on strong confounders, so that's a reason to use matching-- to make the case control study have cases that are more similar to the general population, create a better counterfactual. In small studies, some strata may have few observations in them, making them difficult or impossible to adjust for confounding and inflating standard errors. So that's another situation or reason in which matching is useful. And finally, in case control studies, the main purpose of matching is to avoid strata with small cells. Matching ensures that after stratification by the match factor, there will be cases and controls in each stratum.

What sparse data looks like

- Red cells in the table below are sparse - disease that is rare among people <18 years

Age < 18 years		
	Disease	No disease
Exposed	3	200
Unexposed	0	450

Age ≥ 18 years		
	Disease	No disease
Exposed	86	3824
Unexposed	351	8900

So a situation that can arise in studies is that in which we have sparse data or limited data. So in the example below, when we've stratified our population into individuals less than 18 years of age and those greater than or equal to 18 years of age, the red cells show us how sparse the data are among the disease population under 18 years of age.

Example of matched case-control study

- Study of toxic shock syndrome and tampon use. Controls were friends of cases and other women matched on age and neighborhood. Main purpose was to reduce confounding.

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

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

* Values given are mean ± SD (range).

Reingold et al., 1989

10

Here's an example of a matched case control study. This is a study of toxic shock syndrome and tampon use. The controls were friends of cases and other women matched on age and neighborhood. The main purpose of the matching was to reduce confounding. And you can see, after the matching is done, how balanced the control groups are in comparison to the patients with respect to age, the percent of white, the married percentage, and so forth. So matching was used to create balanced control groups that are good counterfactuals for the patient group.

Matching in cohort studies

- Studies tend to be large and intended to examine multiple exposures and outcomes so other approaches to controlling confounding are usually preferable
- **Prognostic cohort studies** use matching more often because they typically have one exposure and are smaller, so there is concern about overlap in distribution of confounders
- Example: study comparing prognosis of Indigenous Australians with cancer to other Australians with cancer in Queensland
- Indigenous people diagnosed with cancer identified through the cancer registry (n~800)
- Compared with randomly selected non-Indigenous patients who were frequency-matched for age, sex, place of residence, cancer site, and year of diagnosis

11

In cohort studies, matching is also done. Cohort studies tend to be large and intended to examine multiple exposures and outcomes. So other approaches to controlling and confounding are usually preferable, rather than matching. Prognostic cohort studies do use matching more often, because they typically have one exposure and are smaller, so there is concern about overlap in the distribution of confounders.

For example, a study comparing the prognosis of indigenous Australians with cancer to other Australians with cancer in Queensland-- indigenous people who were diagnosed with cancer were identified through the cancer registry, and there were about 800 of them. So to pick a counterfactual group for them, they were compared with randomly selected non-indigenous patients who were frequency matched for age, sex, place of residence, cancer site, and year of diagnosis. So all these potential confounding factors-- age, sex, place of residence, and so forth-- were matched so that those were balanced between the indigenous population in the exposed group and the non-indigenous population in the control or comparison group.

Matching in cohort studies

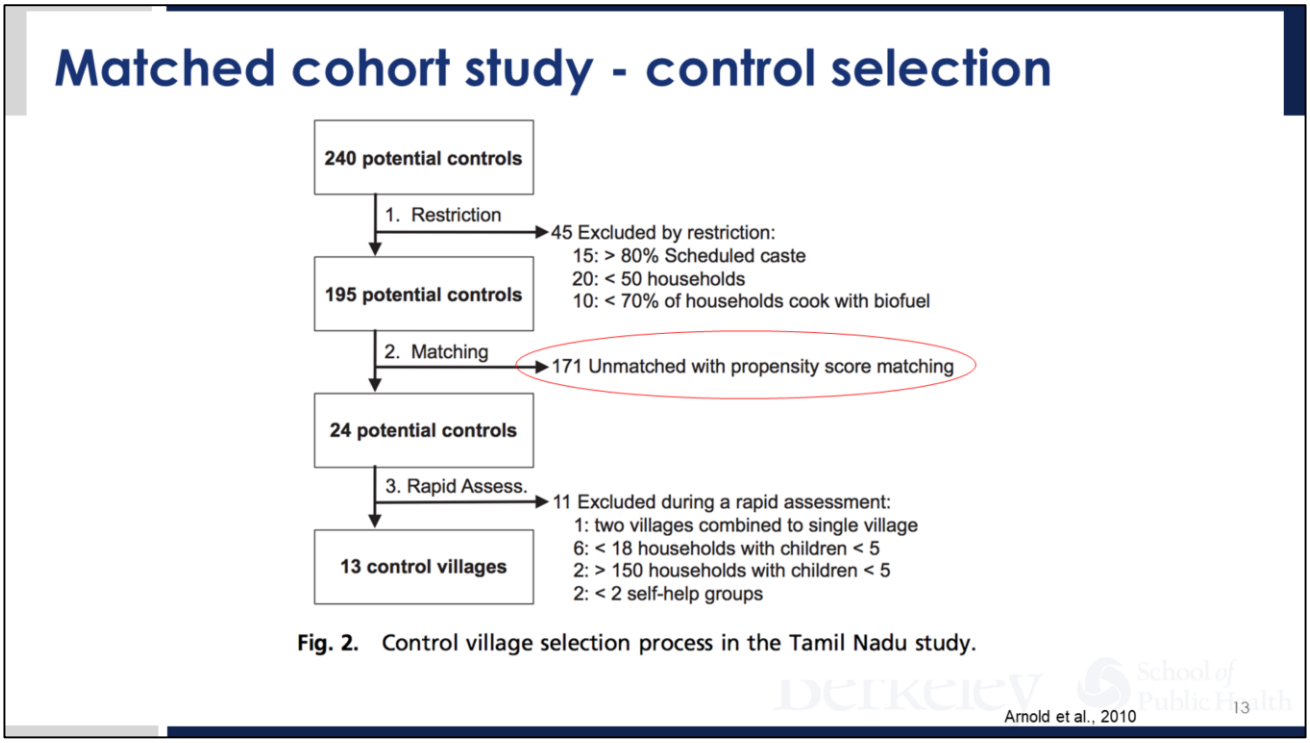
- **Impact evaluations** use matching when randomization is not feasible. Goal is to evaluate an existing, non-randomized intervention.
- Use matching with census (or other data) to identify individuals / communities who can serve as a control group who can approximate a control group in a trial
- Example: evaluation of a sanitation mobilization, water supply, and hygiene intervention in rural India
- Intervention in 12 villages
- Identified 13 control villages using pre-intervention census data and matching

Berkeley School of Public Health

Arnold et al., 2010 12

Impact evaluations are a type of study that use matching when randomization is not feasible. The goal is to evaluate an existing non-randomized intervention in this situation. In this type of design, this impact evaluation, we use matching with census or other data to identify individuals or communities who could serve as a control group and who can approximate a control group as if a trial were done. Again, a trial isn't being done, but can we create a control or comparison group with matching that's similar enough to properly serve as a control group?

So for example, in an evaluation of a sanitation mobilization, water supply, and hygiene intervention in rural India, there were 12 intervention villages, and 13 control villages were identified using pre-intervention census data and matching in this article from our group, led by Ben Arnold, in 2010.



In a matched cohort study, there are steps to the creation of the controller comparison group, and this is, again, from this Arnold 2010 article, where you can see there were 240 potential controls to choose from. The first step was a step called restriction, where villages were excluded by their caste status, because they weren't in the same levels of caste as the villages that had received the intervention. So that left 195 potential controls, and then matching was done, the process we're talking about in this video, where a multi-variate process called propensity score matching-- that is giving a score to each village based on its combination of factors-- was used to match the controls to the comparison group. That left us with 24 potential controls.

And then a rapid assessment was done to see which villages could still participate, and during the rapid assessment, there were some exclusions due to whether two villages had been combined into a single village or some villages were excluded with the lack of children in the right age group to participate and serve as a good comparison group for the intervention. So that resulted in 13 controlled villages that were then compared to the 13 villages that had received the prior intervention.

Matched cohort study - comparison of study groups following matching

- The "All villages" columns compare intervention and control villages prior to matching.
- The "Study sample" columns compare intervention and control villages after matching.
- Characteristics were very similar between groups after matching.

Table 2. Summary of preintervention characteristics before and after village selection

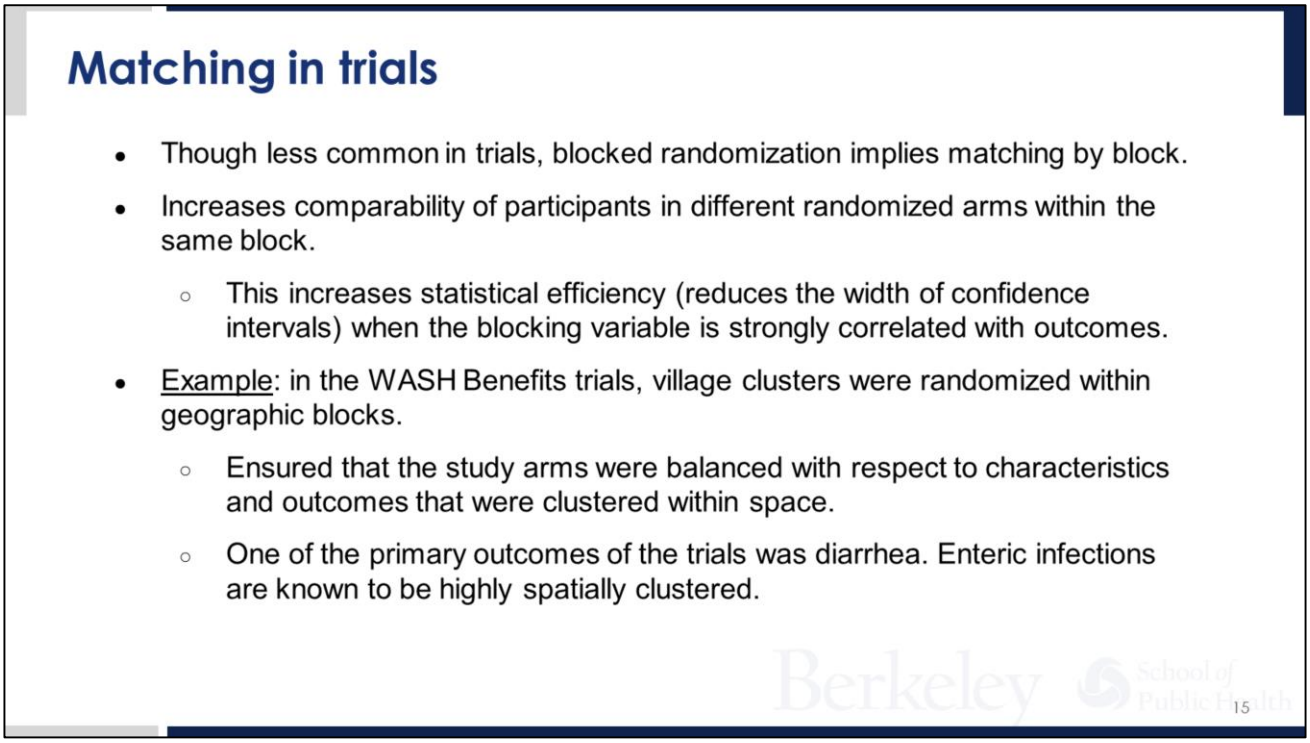
	All villages		Study sample	
Mean	Control	Intervention	Control	Intervention
Demographic				
Total households	170	161	181	161
Persons per household	5	5	5	5
Scheduled caste, %	19	12	15	12
Children ≤5 y old, %	12	12*	12	12
Female literacy, %	52	48	49	48
Socioeconomic				
Employment rate, %	81	78	79	78
Cultivators, %	27	28	31	28
Agricultural laborers, %	24	33	21	33
Marginal workers, %	19	22	21	22
Females work, %	74	69	71	69
Panchayat income (Rp/person)	12,255	7,470***	7,143	7,470
Per-capita cattle ownership	4	4	5	4
Use banking services, %	29	25	25	25
Use biofuel for cooking, %	91	97**	96	97
Own radio, %	43	43	38	43**
Own television, %	21	16	17	16
Own scooter/moped, %	10	10	9	10
Sanitation and water				
Private toilet/latrine, %	15	8**	9	8*
Open defecation, %	85	92**	91	92*
Tap water (private/public), %	75	76	75	76
Hand pump, %	12	14**	18	14
Other water source, %	13	10*	7	10
Persons per hand pump	260	302	240	302
Persons per deep bore well	437	679**	510	679
Water supply level (lpcd)	12	15**	14	15
No. of villages	240	12	13	12

Arnold et al., 2010

14

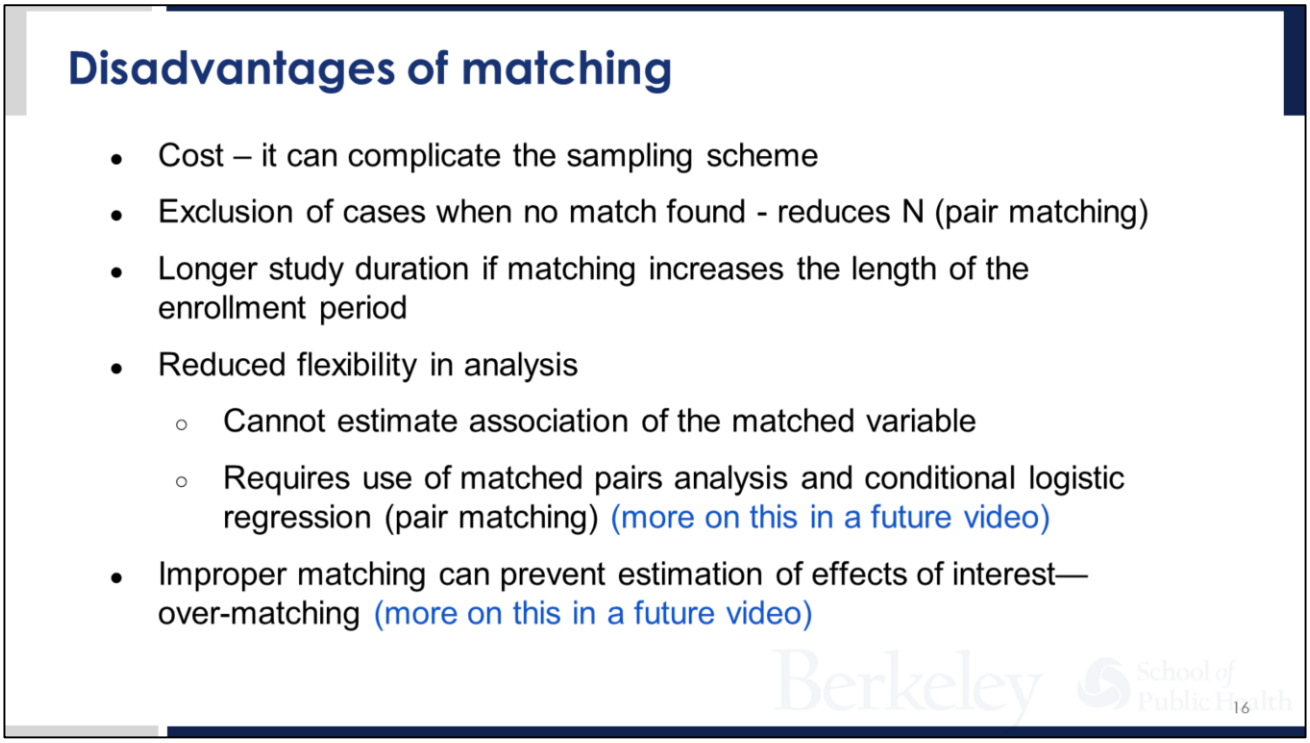
And this next figure from the same article is interesting, because it shows what matching additionally contributed to the study, as opposed to just looking at all villages as possible controls. So what's going on here in Table 2 is, essentially, two ways of conducting the study with different comparison groups. And the first set of columns, under All Villages, all the possible villages in the control group are included. And we're trying to find the best balance for control and intervention groups so that we have a good counterfactual. We'd like the control group to look as much like the intervention group as possible, and you see in the first two columns, under All Villages, the asterisks mark statistically significant difference.

So on many different factors, the control groups are different than the intervention groups. So that's not a good counterfactual group. But look what happens when the study sample, after matching is applied-- this is the situation where 13 control villages are shown in the prior slide were selected to compare to the 13 intervention villages. Now there is a great reduction in the number of characteristics that differ between the two groups. So we've improved the comparison of the two groups, which is our basic goal, because we want a good counterfactual after the matching process.

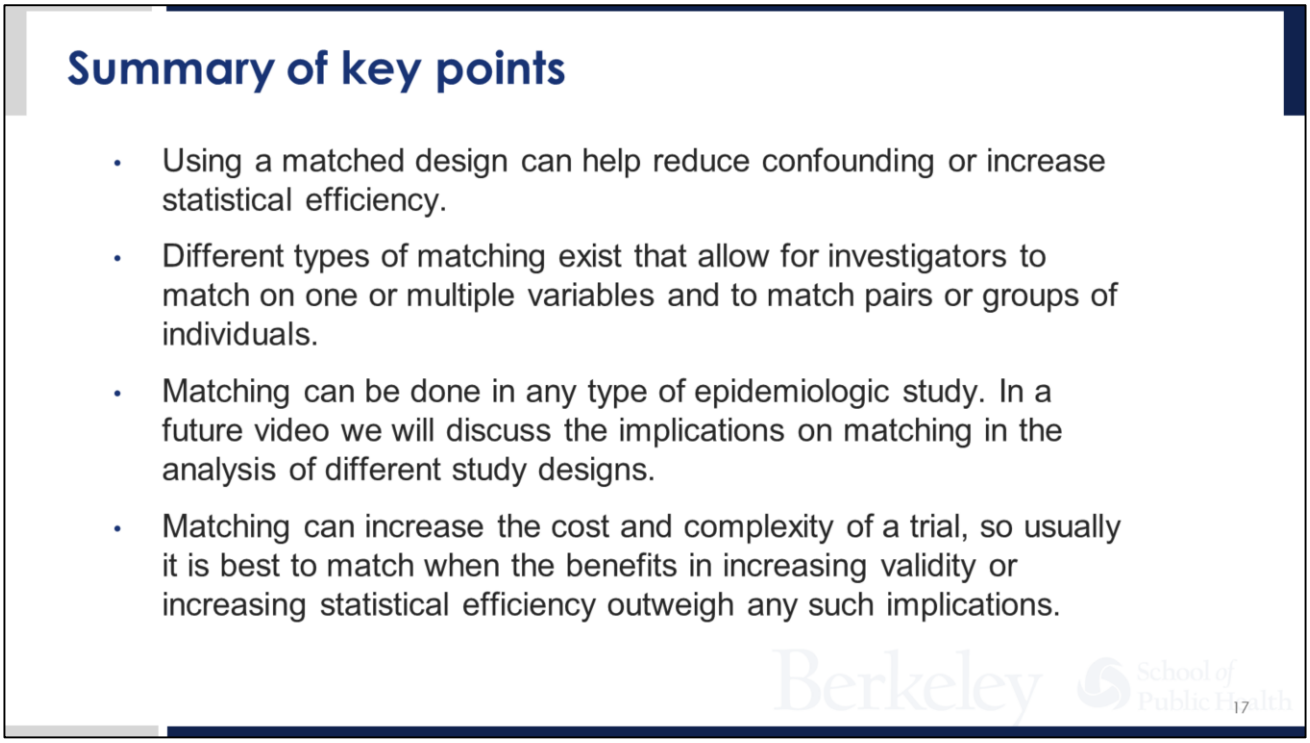


Matching is less commonly used in trials, but when you use blocked randomization, that we've discussed before, that implies matching by block. So if we create-- in a randomized trial, we create blocks of eight participants, for example, at a time. Each block is essentially using matching to create a stratum within which to then do randomization. This process increases the comparability of participants in different randomized arms within the same block and helps to increase statistical efficiency. This reduces the width of the confidence intervals when the blocking variable is strongly correlated with outcomes.

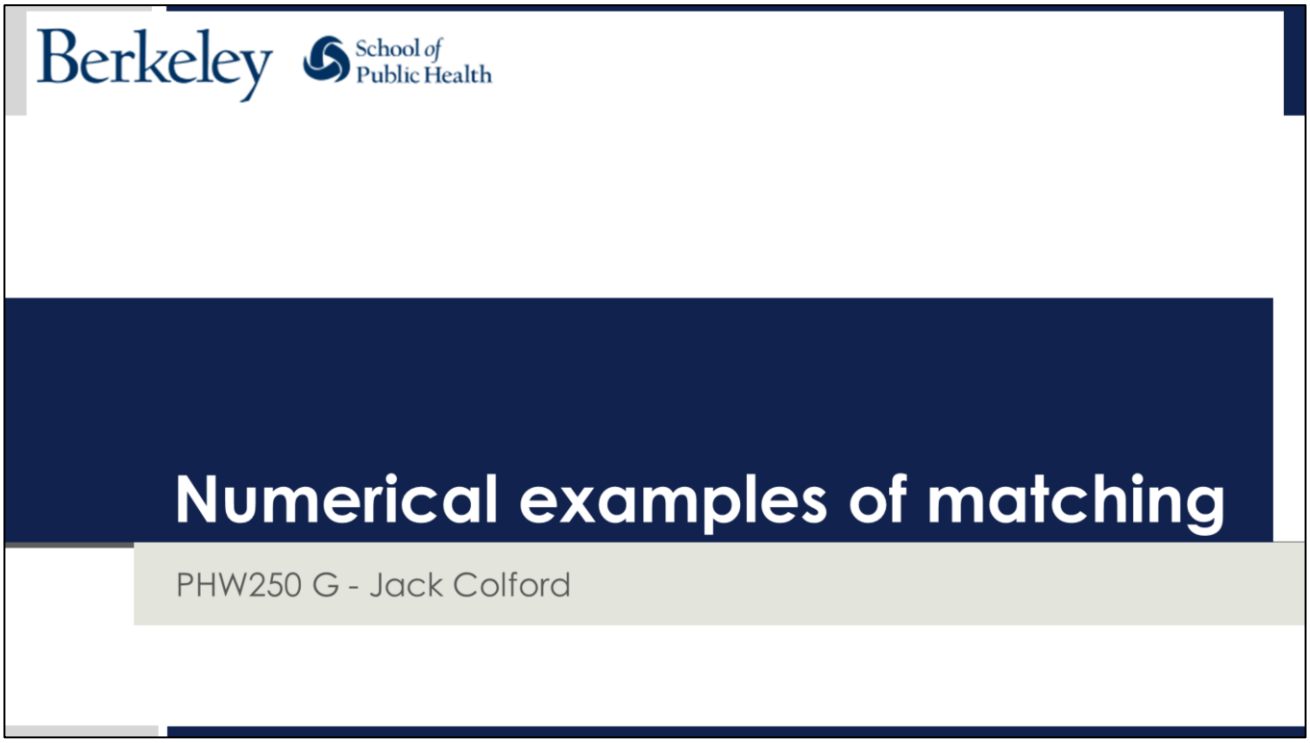
For example, in the WASH Benefits trials that we've discussed before, village clusters were randomized within geographic blocks. And this helped to ensure that the study arms were balanced with respect to characteristics and outcomes that were clustered within space. One of the primary outcomes in the trials was diarrhea. Enteric infections are known to be highly spatially clustered, so that's why the clusters, the groups created by geographic areas, helped to form comparable groups to which randomization was applied. So that's a type of matching.



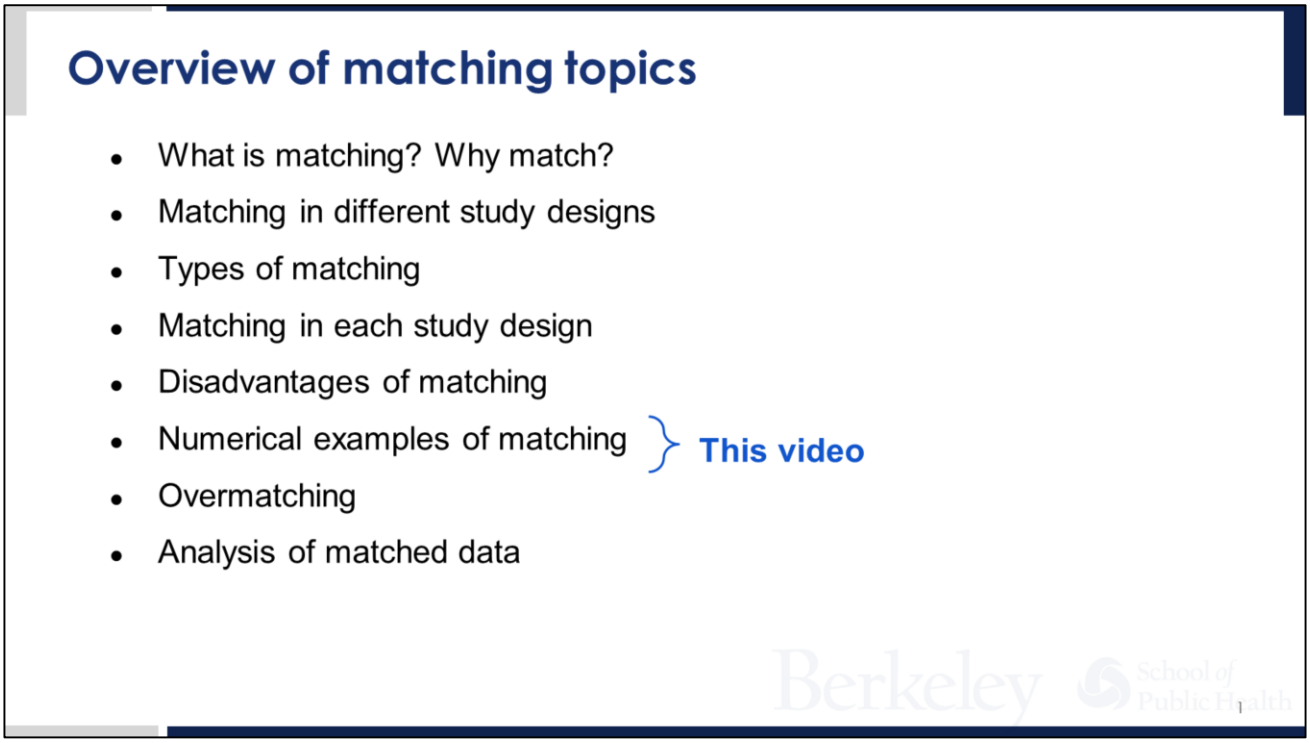
There are some disadvantages to matching. It can be costly and complicate the sampling scheme. It can lead to the exclusion of cases when no match is found, so if you can't find a control for a case, then you aren't able to use that case. So that's not ideal. It can create a longer study duration if the matching process increases the length of the enrollment period. It can result in reduced flexibility and analysis. And we can't estimate the association of the matched variable, so if the matched variable's relationship with the outcome is itself of interest, we can't do that if we've matched on a certain variable. If we've matched on age in a study of smoking and cancer, we can no longer study the relationship between age and cancer, for example. And matching requires the use of matched pairs analysis and conditional logistic regression in the pair matching situation, and we'll talk more about this in a future video. Improper matching can prevent estimation of effects of interest, and this is called overmatching, and we'll talk about this in a later video as well.



So to summarize some of the key points we've covered in this Introduction to Matching module, we can use a matched designed to help us reduce confounding or to increase statistical efficiency. Both are reasons to consider matching. Different types of matching exist that allow for investigators to match on one or multiple variables and to match pairs or groups of individuals. Matching can be done in any type of epi study. In a future video, we will discuss the implications on matching and the analysis of specific different study designs. And finally, matching can increase the cost and complexity of a trial, so usually, it's best to match when the benefits in increasing validity or increasing statistical efficiency outweigh any of these implications.



I think that matching is easier to understand with some actual examples that use numbers to make calculations.



We covered several topics in the earlier video about why we matched and how matching was used in different study designs. This video is going to focus just on some numerical examples before we move on to other topics such as overmatching and the analysis of match data.

Numerical examples of matching

Examples from Rothman et al. *Modern Epidemiology*

- Cohort study
  - Does not necessitate control of the matched factor in the analysis
- Case-control study
  - Necessitates control of the matched factor in the analysis

So these numerical examples of matching are taken from your Rothman Modern Epidemiology textbook. In a cohort study, there is no special control needed for the match factor in the analysis. But in a case control study, in most situations, there's a necessity to control for the match factor in the analysis itself. And we'll go through some examples of this.

Matching in a cohort study

Example of data from target population

	Men		Women	
	Disease	**Total**	**Disease**	**Total**
Exposed	4,500	900,000	100	100,000
Unexposed	50	100,000	90	900,000

Crude CIR pooling across gender: $\frac{(4500 + 100) / (900{,}000 + 100{,}000)}{(50 + 90) / (100{,}000 + 900{,}000)} = 33$

CIR for Men: $\frac{4500 / 900{,}000}{50 / 100{,}000} = 10$

CIR for Women: $\frac{100 / 100{,}000}{90 / 900{,}000} = 10$

There is strong confounding by gender.

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

3

So here's some example from a cohort study in which data is drawn from a target population. And we've stratified the population into men and women and divided them by their exposed and unexposed status, and tabulated how much disease there is in the total population.

So if we calculate a crude incidence ratio across all the genders-- that is, we don't take account of gender, we just add up all the cases divided by the population in the exposed. And then we add up all the cases divided by the population in the unexposed. You see the calculations here.
There are 4,500 plus 100 exposed individuals who developed disease. So that's a total of 4,600 over a total exposed population of 900,000 plus 100,000. And for our denominator for the cumulative incidence ratio, we have 50 cases in the men plus 90 cases in the women among the unexposed. So that's a total of 140, again over 1 million, the sum of all the unexposed individuals.

So with that numerator and that denominator, we have a cumulative incidence ratio of 33, a suggestion of a very high relationship between exposure and disease. But let's do the cumulative incidence ratio stratified. That is, for each of the two strata men and women.
So you see the calculation here for men. The cumulative incidence ratio is 10. And for women, it's also 10. So the two stratified estimates give a value that's quite different than the unadjusted or the crude ratio because the average of 10 and 10 would be 10, of course. So the adjusted estimate would be 10, and that's quite different than the crude cumulative incidence ratio of 33. So there's strong confounding by gender of the relationship between exposure and disease in this example.

Matching in a cohort study

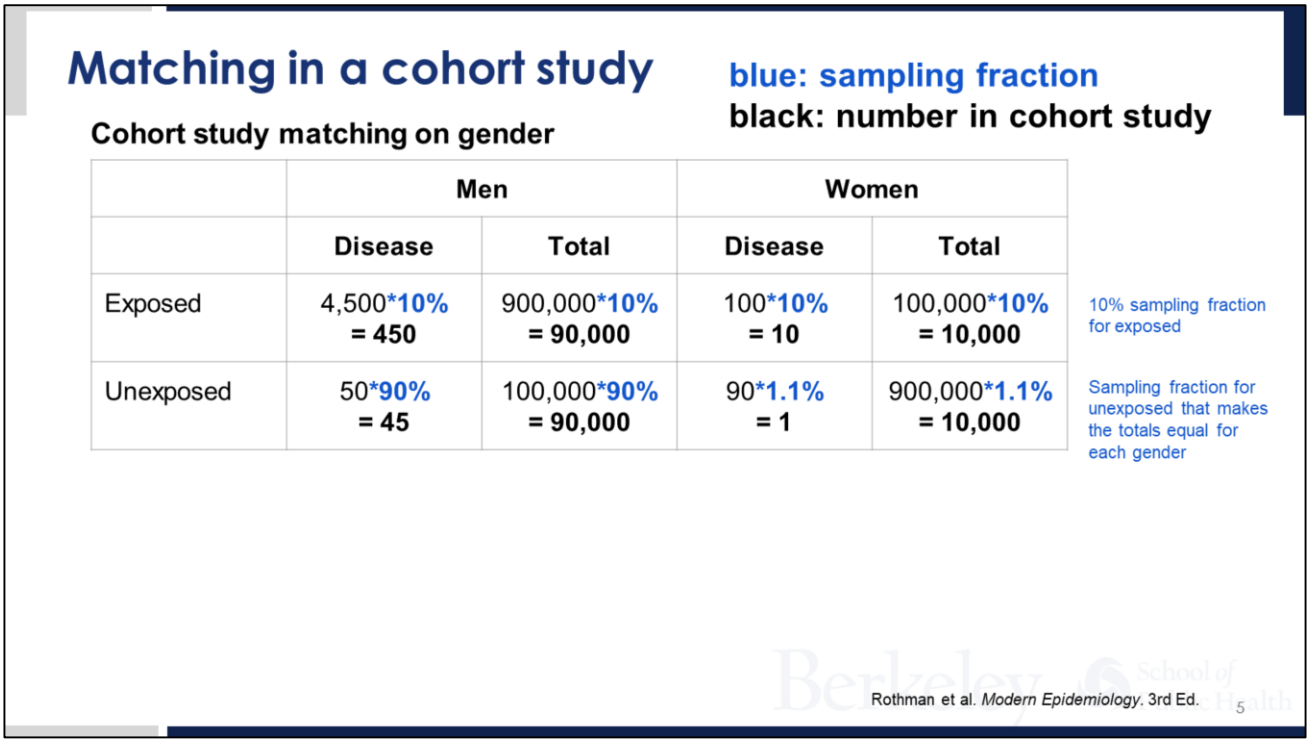
Designing a cohort study to control for confounding by gender

	Men		Women	
	Disease	**Total**	**Disease**	**Total**
Exposed	4,500	900,000	100	100,000
Unexposed	50	100,000	90	900,000

- Notice that 90% of those exposed are male and 10% are female.
- We want to conduct a cohort study using a 10% sample of the exposed.
- If we draw a random sample of the unexposed, the same confounding will occur.
- If we draw a sample so that the proportion of men matches that in the exposed cohort, how much confounding remains?

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

Now let's design a cohort study to control for confounding by gender. And notice that 90% of those that were exposed were male and 10% are female. So we want to conduct a cohort study where we're going to use a 10% sample of the exposed group. If we draw a random sample of the unexposed group, the same confounding will occur. If we draw a sample so that the proportion of men matches that in the exposed cohort, how much confounding remains? In other words, will matching remove the confounding that we know is present?



So in this table, we're showing in blue the sampling fraction when we sample the same in each of the exposed groups based on keeping the sampling fraction the same. And the black shows the numbers in the cohort study. So in the men, if we sample 10% of the 4,500 exposed disease men, we get a sample of 450. And if we sample 10% of the total population of exposed men, a 10% sample of 900,000 is 90,000.

And similarly, applying the same sampling fractions to the women, we come up with a 10% sampling fraction for the exposed group. So for the 100 disease-exposed women, 10% sampling fraction would give us 10. And a 10% sampling fraction of the total exposed women would be 10% of 100,000, or 10,000. The sampling fraction for the unexposed, it makes the totals equal for each gender is then applied.

In the unexposed group, 90% of the population of 100,000, that gives us 90,000 for the unexposed men. We sample 90% of that population of 50, that gives us 45. And a 1.1% sample of the 900,000 unexposed women gives us 10,000 women. And so a 1.1% sample of the 90 unexposed diseased women gives us one woman.

Matching in a cohort study

blue: sampling fraction
black: number in cohort study

Cohort study matching on gender

	Men		Women		
	Disease	**Total**	**Disease**	**Total**	
Exposed	4,500***10%** **= 450**	900,000***10%** **= 90,000**	100***10%** **= 10**	100,000***10%** **= 10,000**	10% sampling fraction for exposed
Unexposed	50***90%** **= 45**	100,000***90%** **= 90,000**	90***1.1%** **= 1**	900,000***1.1%** **= 10,000**	Sampling fraction for unexposed that makes the totals equal for each gender

Crude CIR pooling across gender: $\frac{(450 + 10) / (90{,}000 + 10{,}000)}{(45 + 1) / (90{,}000 + 10{,}000)} = 10$

CIR for Men: $\frac{450 / 90{,}000}{45 / 10{,}000} = 10$

CIR for Women: $\frac{10 / 10{,}000}{1 / 10{,}000} = 10$

Confounding by gender was removed by matching on gender.

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

6

If we now conduct our same estimates of the crude cumulative incidence ratio using the new numbers we have created by our sampling, we get a crude estimate of the cumulative incidence ratio of 10. And of course, the cumulative incidence ratio for men and women is still 10 in each of the strata.

So what has happened here is that the confounding that we saw before-- remember that the crude ratio was 33? That was removed by matching on gender. So matching in the cohort study removed the confounding by gender that we saw previously.

Take home message: matching in cohort studies

- Matching on a confounder in a cohort study can remove confounding.
- When a cohort study is matched in the design phase, it is not necessary to adjust for the confounder that was matched on in the analysis because the matching already removed confounding.

So what's the take home message about when we apply matching in a cohort study? Well, matching on a confounder in a cohort study can remove confounding. And when a cohort study is matched in the design phase, it's not necessary to adjust the confounder that was matched on in the analysis because the matching already removed the confounding.

Matching in a case-control study

Example of data from target population

	Men		Women	
	Disease	**Total**	**Disease**	**Total**
Exposed	4,500	900,000	100	100,000
Unexposed	50	100,000	90	900,000

	Cases	**Controls**
Exposed	4,600	4,600
Unexposed	140	140

Suppose that we sample 4,740 controls from the target population that are matched on sex.

Crude OR = 6,400* 626 / 4,114*140= 5.0

This OR is lower than the true RR of 10.

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

Now let's contrast that with a case control study. Here's a table from the same population where we're going to conduct a case control study. So suppose that we sample 4,740 controls from the target population, and we've matched them on sex. Well, in our smaller table at the bottom here, we see the crude odds ratio is 5.0. And this odds ratio is lower than the true relative risk of 10.

Matching in a case-control study

Example of sampling the target population for the case-control study

	Men		Women	
	Cases	**Controls**	**Cases**	**Controls**
Exposed	4,500		100	
Unexposed	50		90	
Total	4,550	4,550	190	190

- All 4740 individuals with disease included as cases.
- Equal number of controls matched on gender are sampled from the target population.

Rothman et al. *Modern Epidemiology*. 3rd Ed. 9

Now let's stratify and do our analysis after the matching. All 4,740 individuals with disease are included as cases. And an equal number of controls matched on gender are then sampled from the target population. So let's fill this table in and see where individuals would fall into each of the boxes.

Matching in a case-control study

Example of sampling the target population for the case-control study

	Men		Women	
	Cases	**Controls**	**Cases**	**Controls**
Exposed	4,500	4,095	100	
Unexposed	50	455	90	
Total	4,550	4,550	190	190

- All 4740 individuals with disease included as cases.
- Equal number of controls matched on gender are sampled from the target population.
- Of the 4550 male controls, assume 90% are exposed (N=4095) and 10% unexposed (N=455).

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

10

So we see that for male controls, we assume that 90% of them are exposed. So that's 4,095. And 10% are unexposed, which would be 455. And this is just using the same ratios as before.

Matching in a case-control study

Example of sampling the target population for the case-control study

	Men		Women	
	Cases	**Controls**	**Cases**	**Controls**
Exposed	4,500	4,095	100	19
Unexposed	50	455	90	171
Total	4,550	4,550	190	190

- All 4740 individuals with disease included as cases.
- Equal number of controls matched on gender are sampled from the target population.
- Of the 4550 male controls, assume 90% are exposed (N=4095) and 10% unexposed (N=455).
- Of the 190 female controls, assume 10% are exposed (N=19) and 90% unexposed (N=171).

Rothman et al. *Modern Epidemiology*. 3rd Ed. 11

And of the 190 female controls, if we assume that 10% are exposed, that would give us 19 female controls. And 90% are unexposed, that would give us 171 unexposed female controls. So now when we do our calculations of the crude odds ratio as we saw, that's 5.0, which is much lower than the true relative risk, which we know is 10 from the cohort data.

Matching in a case-control study

Example of sampling the target population for the case-control study

	Men		Women	
	Cases	**Controls**	**Cases**	**Controls**
Exposed	4,500	4,095	100	19
Unexposed	50	455	90	171
Total	4,550	4,550	190	190

Crude OR = (4,500 + 100) * (455 + 171) / (4,095 + 19) * (50 + 90)= 5.0 (Much lower than true RR=10)

Male OR = 4500 * 455 / 4,095* 50 = 10

Female OR = 100 * 171 / 90* 19 = 10

Confounding of the pooled OR was not removed by matching on gender.

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

12

So among the males, the odds ratio, though, when we stratify and do our analysis after the matching, is 10, and among the females, it's 10. So the adjusted estimate would also be 10. So all of this is to say that the confounding of the pooled odds ratio wasn't removed by matching on gender. We have to do something extra to remove that confounding. And that was the stratified analysis we did.

Take home message: matching in case-control studies

- Matching on a confounder in a case-control study can remove confounding but only if you stratify on the confounder in the analysis.
- When a case-control study is matched in the design phase, it is necessary to adjust for the confounder that was matched on in the analysis because the matching already removed confounding.
- If you do not stratify on the confounder in the analysis, there will be bias towards the null.

So matching on a confounder in a case control study can remove confounding, but only if you stratify on the confounder in the analysis. When a case control study is matched in the design phase, it's necessary to adjust for the confounder that was matched on in the analysis. If you don't stratify on the confounder in the analysis, there will be a bias toward the null.

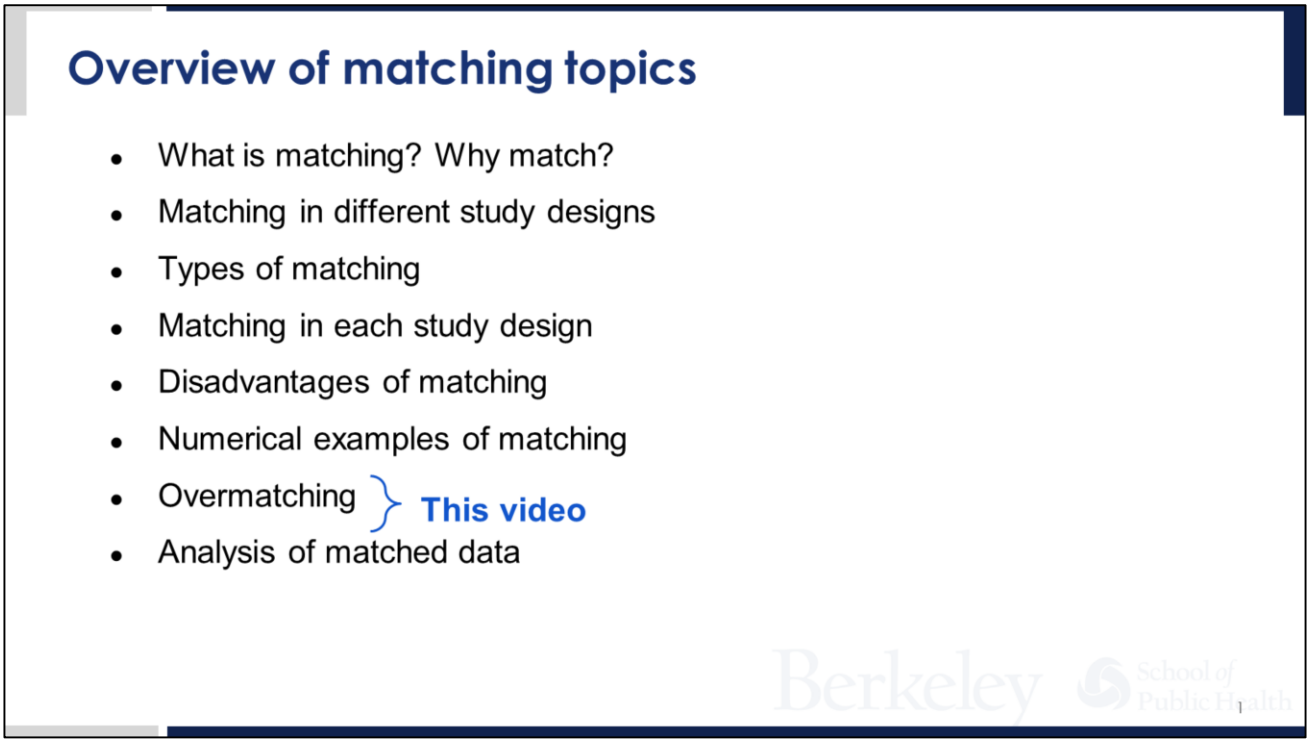
Summary of key points

- In a matched cohort study, **it is not necessary** to adjust for the confounder that was matched on in the analysis.
- In a matched case-control study, **it is necessary** to adjust for the confounder that was matched on in the analysis.

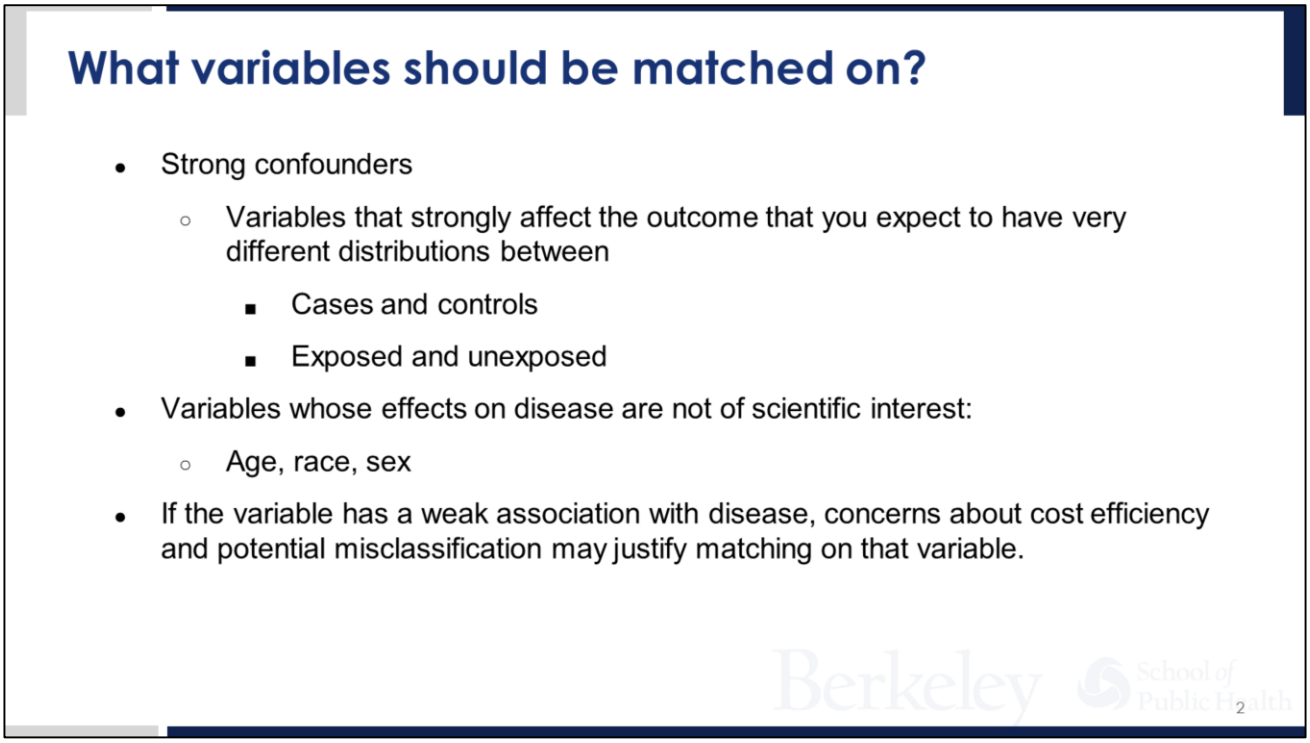
So in summary, in a matched cohort study, it's not necessary to adjust for the confounder that was matched on in the analysis. In a matched case control study, it is necessary to adjust for the confounder that was matched on the analysis. So it's two very different situations in the cohort versus case control study.

Lecture: Diagnosing Overmatching with DAGs

In this video, I'm going to introduce the concept of overmatching. And then I'll tell you how you can assess the risk of overmatching in the design phase of a study using Directed Acyclic Graph, or DAGs.



We've talked about what matching is, why it can be advantageous to match, as well as a bunch of details related to matched study designs and numerical examples of matching. And this video focuses on the issue of overmatching specifically. And again, this is something that is valuable to think about at the design stage. Because as we're about to see, some types of overmatching can cause irreparable harm on our ability to estimate valid or accurate or precise measures of association.



In a matched study, what variables should we consider as potential matching factors? Strong confounders are great candidates. These are variables that strongly affect our outcome or disease that we expect to have very different distributions between cases and controls and the exposed and the unexposed. And the reason that we want to have these very different distributions is that matching helps us ensure that we have sufficient observations within the strata of our two by two table to be able to do a statistically efficient analysis.

We also want to pick variables whose effects on our disease or outcome are not of our own scientific interest. Examples include age, race, and sex. And the reason for this is that if we match on a certain factor depending on our study design, we might not be able to analyze the association between that matched factor and our outcome in a matched design once the data is collected.

So I mentioned strong confounders at the top of this slide. And there are some situations where a variable might have a weak or moderate association with disease. So it's a weak or moderate confounder. And in those situations, we may choose to match even though it's not a strong confounder because there's concerns about cost efficiency or misclassification. And I'm not going to go into detail about this. But your reading in the Rothman textbook provides some more examples of this kind of situation.

What is overmatching?

- Overmatching occurs when matching on a non-confounder.
- It can occur when matching on an intermediate between exposure and disease, or a factor that is affected by both exposure and disease can lead to bias
- **Types of overmatching:**

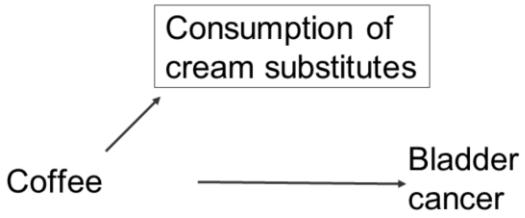
1) Overmatching that harms statistical efficiency (precision)
2) Overmatching that harms validity (accuracy)

Rothman et al., 3rd Ed.

When we match on a non-confounder, something called "overmatching" occurs. This can also occur when we match on an intermediate between the exposure and disease or a factor that's affected by both the exposure and disease, which we could call a "collider."
So we have two different types of overmatching we'll talk about-- overmatching that harms our statistical efficiency-- in other words, our precision-- and overmatching that harms our validity or our accuracy of our estimates. Now, ideally, we would maximize both precision and accuracy. So we want to try to design our study in a way that minimizes any potential effect of overmatching on statistical efficiency or validity.

Overmatching that harms statistical efficiency

- Results from matching on a non-confounder associated with exposure but not disease
- Causes a loss of information in the analysis because the stratified analysis would have been unnecessary without matching.
- Worst candidate for matching: variable strongly correlated with exposure but not disease.



In this example, many controls matched to cases will be classified identically to the case with regard to coffee drinking merely because they also consume cream substitutes.

Rothman et al., 3rd Ed.

4

Let's start with overmatching that harms statistical efficiency. So this occurs when we match on a non-confounder that's associated with the exposure, but not the disease. Let's take a look at this DAG, Directed Acyclic Graph, on the bottom left. Coffee is our exposure. So we're interested in whether coffee consumption increases the risk of bladder cancer, our outcome.

Let's say we matched in our study design phase on whether people consumed cream substitutes. So these are products that people can put into their coffee instead of cream. And we have a box around consumption of cream substitutes because by matching, we're essentially stratifying our data in a specific way based on this variable.

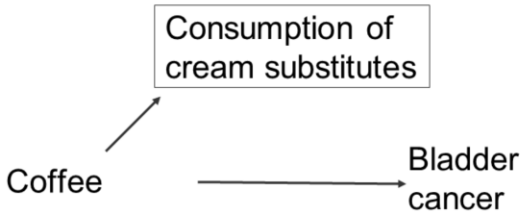
So this is not a confounder because we have an arrow from our exposure into this variable. And also, this variable is not associated with bladder cancer. When we do this kind of matching, we lose information during our statistical analysis because the stratified analysis would not have been necessary without doing this matching.

So in a upcoming video, you'll see how we analyze matched data. And we'll take a closer look at this. But for our purposes, you just need to be able to recognize this particular DAG structure and identify that this DAG structure harms statistical efficiency in a matching context.

So in our example, we may have many controls matched to cases that will be

Overmatching that harms statistical efficiency

- Results from matching on a non-confounder associated with exposure but not disease
- Causes a loss of information in the analysis because the stratified analysis would have been unnecessary without matching.
- Worst candidate for matching: variable strongly correlated with exposure but not disease.



In this example, many controls matched to cases will be classified identically to the case with regard to coffee drinking merely because they also consume cream substitutes.

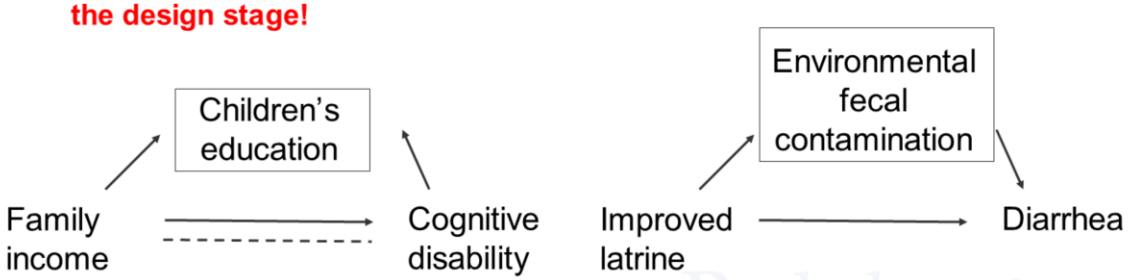
Rothman et al., 3rd Ed.

classified identically to the case regarding their coffee drinking, the exposure. And that's only because they consumed cream substitutes. It's not for any other reason. But this particular exposure isn't associated with the disease. So it's introducing this really unnecessary form of stratification that can harm our statistical efficiency.

And the worst kind of candidate for matching would be a variable that's strongly correlated with our exposure, but not with our disease at all. So this DAG is not really showing how strong the correlations are. But the fact that there's no arrow from the consumption of cream substitutes to bladder cancer and the fact that we have an arrow from coffee into consumption of cream substitutes suggests that, in this study, matching on consumption of cream substitutes is not a good idea.

Overmatching that harms validity

- Results from matching on a variable that is affected by the exposure or disease or both (a collider)
- Matching on (i.e. conditioning on) a collider of the exposure and outcome opens a backdoor pathway between them, introducing bias.
- Matching on an intermediate also introduces bias.
- **This bias cannot be fixed in the analysis! It is very important to avoid in the design stage!**



Rothman et al., 3rd Ed.

The next kind of overmatching can harm validity. And this is arguably a larger concern, something we really want to avoid. So this occurs when we match on a variable that's affected by the exposure or disease or both. And if it's both, we would call that a "collider."

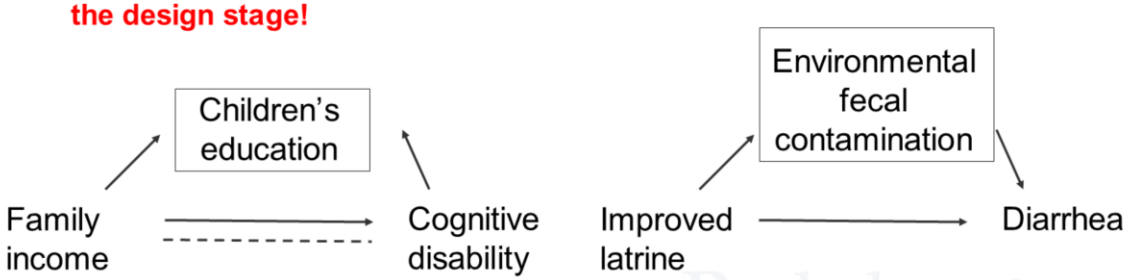
Now let's take a look at our DAGs. So the bottom-left DAG shows us an exposure of family income, an outcome of interest-- cognitive disability. And then children's education is a collider in this relationship. And if we match on children's education, because it's a collider, we're essentially stratifying on this variable, which opens a backdoor pathway between family income and cognitive disability.

Our DAG on the right shows us a slightly different example. We have improved latrines as our exposure of interest, diarrhea as our outcome of interest, and then environmental fecal contamination as our potential matching factor. The improved latrine will affect environmental fecal contamination. And environmental fecal contamination will affect diarrhea.

If we match on environmental fecal contamination, this is matching on an intermediate. It doesn't open a backdoor pathway. But it does introduce bias. And we haven't formally shown you that statistically in this course. It's a little bit more of an advanced topic. But take our word for it. These are our two DAG structures that we want to avoid when matching.

Overmatching that harms validity

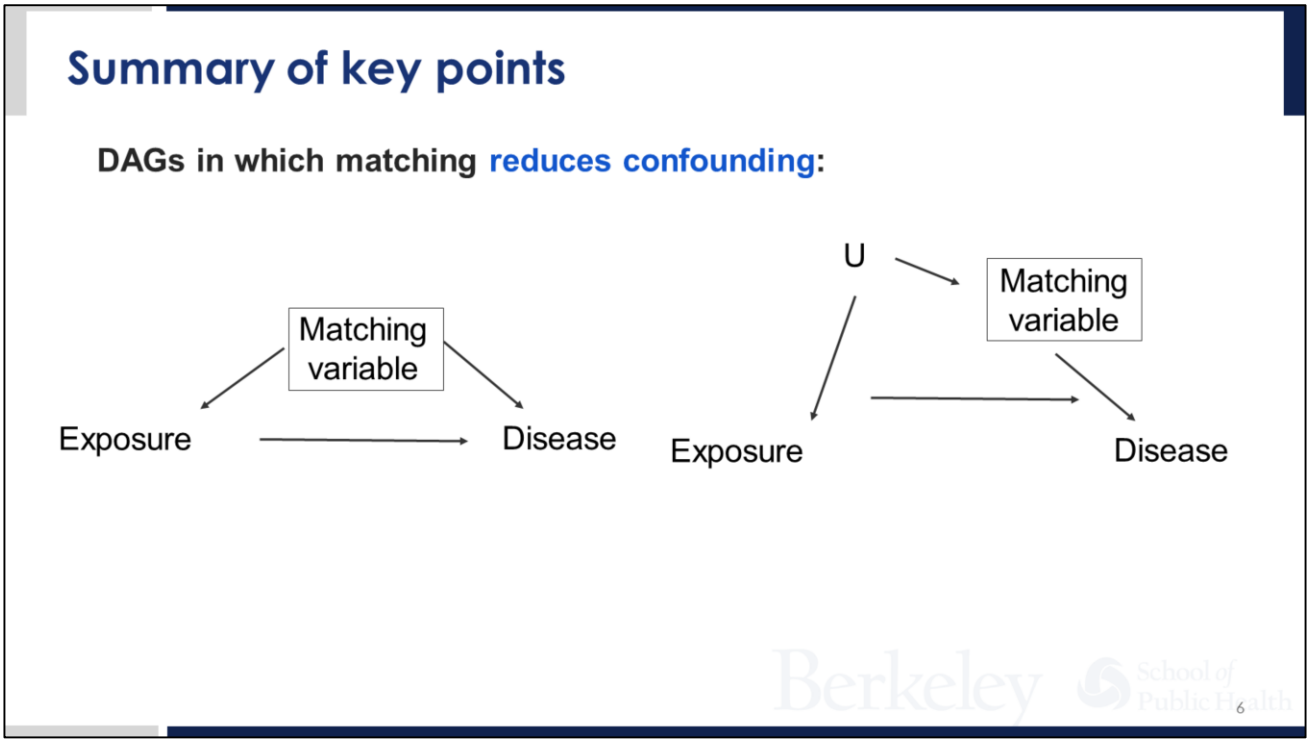
- Results from matching on a variable that is affected by the exposure or disease or both (a collider)
- Matching on (i.e. conditioning on) a collider of the exposure and outcome opens a backdoor pathway between them, introducing bias.
- Matching on an intermediate also introduces bias.
- **This bias cannot be fixed in the analysis! It is very important to avoid in the design stage!**



Rothman et al., 3rd Ed.

And the reason is if we do this in the design phase, we can't fix it in the analysis because it's affected the way we've sampled our population. And that's the only data that we'll have. So it's really, really important that we draw a DAG at the design stage to try to assess if we're going to match if any of these structures are present-- either the structures on this slide or the previous slide-- and try to avoid them as we think about our potential matching factors.

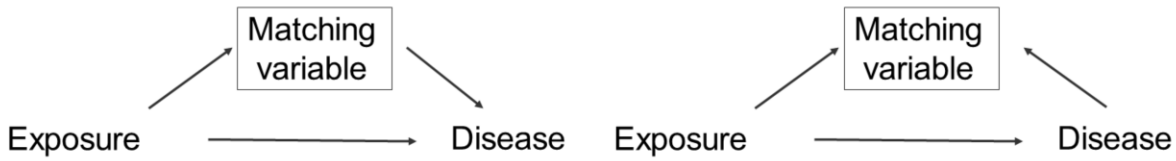
Now, I want to also point out that in the DAG on the right-hand side of this slide, another reason you might not want to match on environmental fecal contamination is that you might be interested in the effect of environmental fecal contamination on diarrhea. And if we match on it again, we're not able to explicitly look at that effect. So that's yet another reason that that DAG would indicate this is not a good design strategy.



So to summarize, here are DAGs in which matching would reduce confounding. So this is a good thing. So on the left, we have our classical DAG structure for confounding, where the matching variable affects the exposure and also affects the disease. In the DAG on the right, we have a similar structure. But we also have an unmeasured variable, as indicated by the U, that affects the exposure and the matched variable. And in this situation, it's still desirable to match on the matching variable because it's closing the pathway from exposure to disease through that matched variable and through the unmeasured variable.

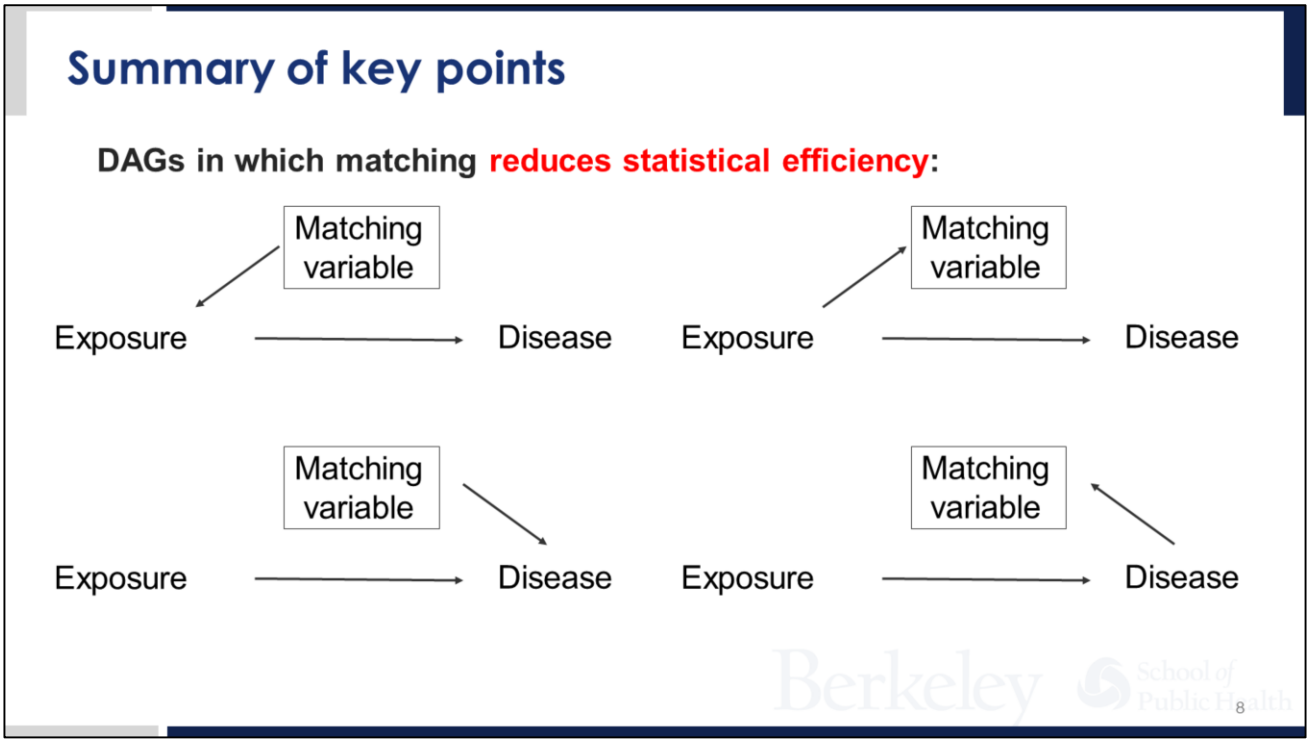
Summary of key points

DAGs in which matching introduces bias:

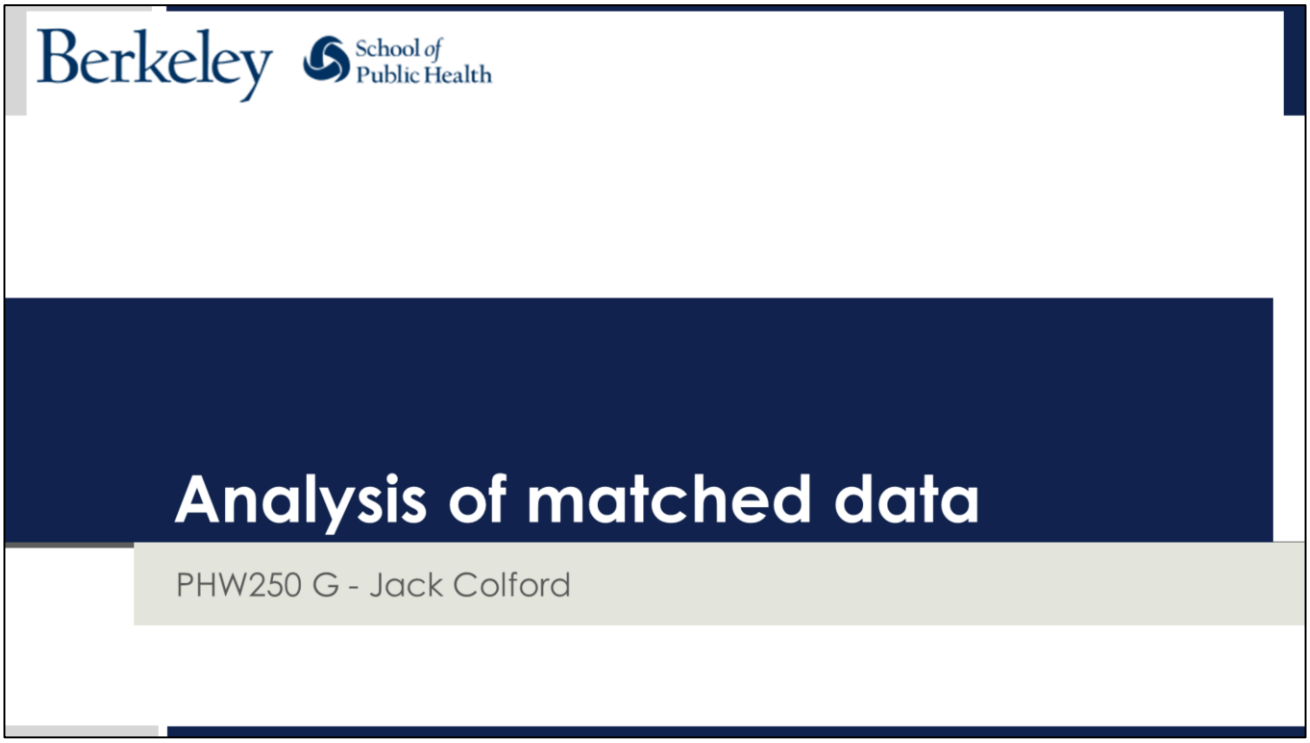


7

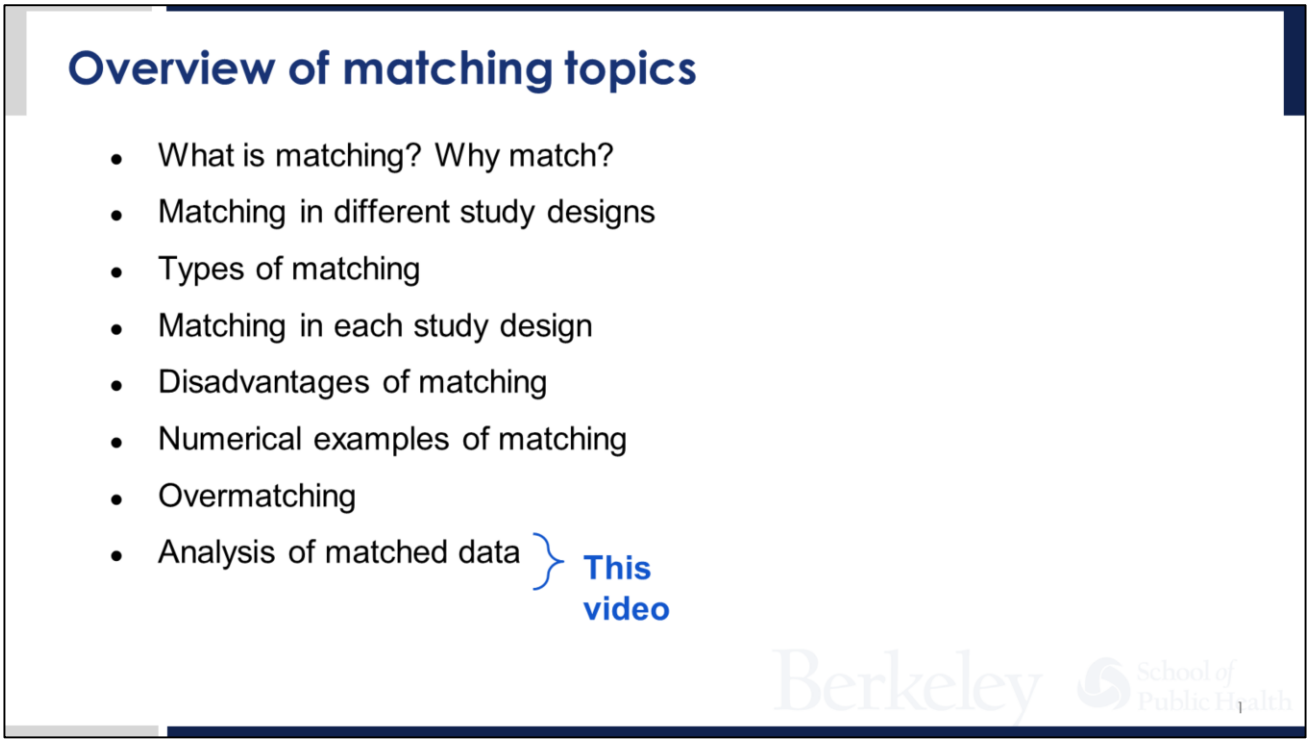
Here are DAGs that introduce bias. And these are the ones we really, really want to avoid. So this is a structure where we match on an intermediate of the exposure-disease relationship is shown on the left and a DAG where the matching variable is a collider of the exposure-disease relationship on the right.



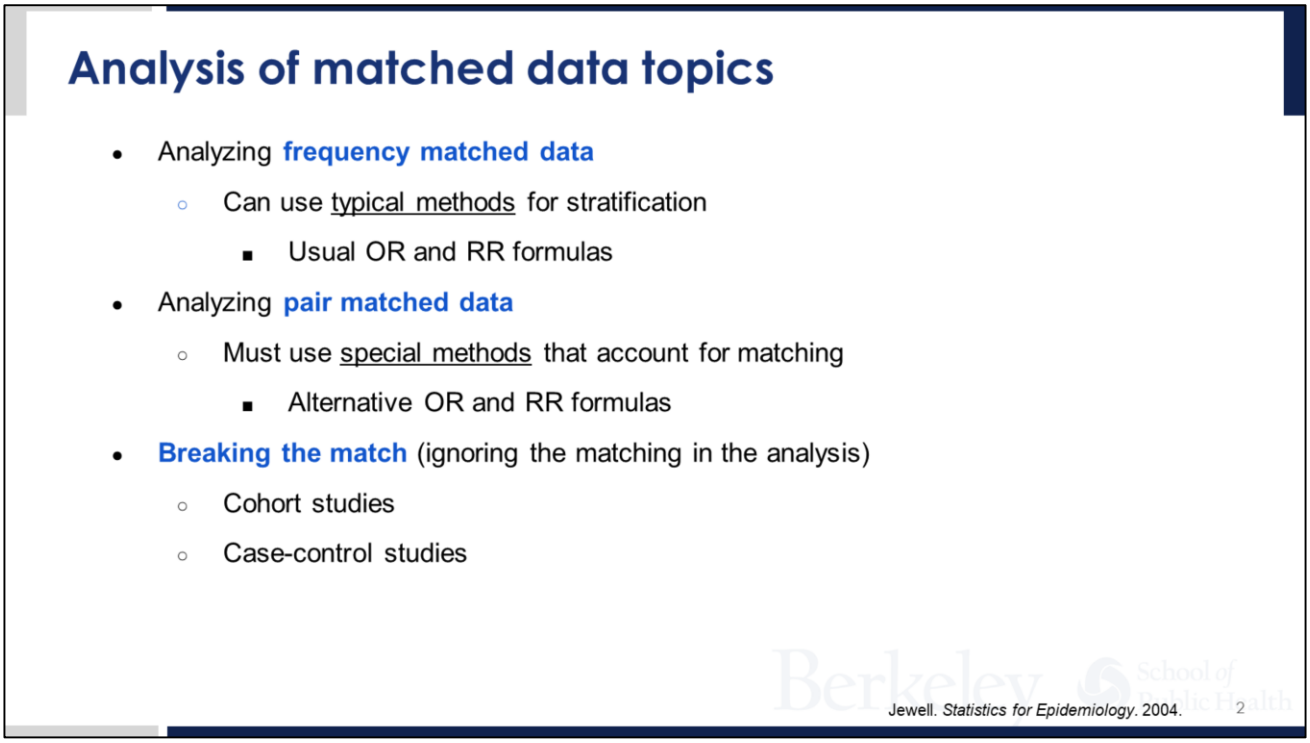
And finally, here are DAGs where matching reduces statistical efficiency. So this is also not desirable. And there is a lot of different DAGs on this slide. But basically all of them are showing you a situation where the matching variable is only associated with the exposure or disease. And the arrow can go in either direction. So we really, really don't want to match on something that doesn't have an arrow to exposure or to disease or from exposure or from disease.
So again, make sure to draw a DAG in your design phase if you're interested in matching. And look out for these potentially harmful DAG structures that could affect your statistical efficiency and the validity of your measures of association.



Now that we've been introduced to matching and done some numerical examples of matching, let's actually do some analysis of match data.



So this final video in our March Through Matching is just going to focus on some analysis examples.



The various topics to consider when we talk about the analysis of match data include three main areas.

First, how do we analyze frequency matched data? For these types of data, we can use typical methods for stratification that we've seen before. And then use adjusted calculations for the odds ratio and the relative risk formula like we've seen in prior videos when we calculate adjusted or weighted averages using the examples and formulas from earlier in the course.

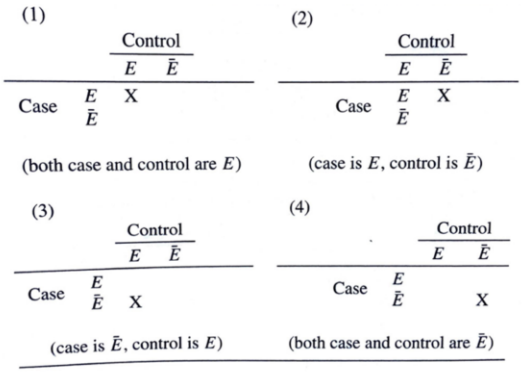
Second, we can analyze pair matched data, but here we have to use special methods that account for matching. So when pair matching has been used specifically, we have to use these special methods. And there are alternative formula here for developing the odds ratio and relative risk estimates.

And finally a third approach we can take is to break the match and ignore the match in the analysis. And we'll talk about this situation for cohort studies and case control studies.

Analyzing pair matched data

This example focuses on case-control data, but these points also apply to cohort data.

- In pair-matched studies, there are only 2 observations in each stratum of the matched factor — a case and a control.
  - Type 1: Both case and control are exposed
  - Type 2: Case is exposed, control is unexposed
  - Type 3: Case is unexposed, control is exposed
  - Type 4: Both case and control are unexposed
- We can use this classification to summarize pair matched data.



Jewell. *Statistics for Epidemiology*. 2004. 3

Let's first talk about how to analyze pair matched data.

Well, pair matched data are a bit special. There are really two observations in each stratum of the match factor, a case and a control. So if you think about a case and a control, the case can either be exposed or unexposed, and the control can be exposed or unexposed.

So that creates four different possible strata. So in stratum type one, both the case and the control are exposed. And you see there the x where both the case and the control are exposed in the figure, the top left figure.

In type two, the case is exposed and the control is unexposed. And you see that in the labeled number two in the figure to the right as well. So the case is exposed and the control is unexposed here.

In the third type, the case is unexposed and the control is exposed. And in the fourth type, both the case and the control are unexposed. We can use this classification to summarize pair matched data.

Analyzing pair matched data

- The lower table shows a new type of 2x2 table with the number of exposed vs. unexposed cases on the left and the same layout for controls at the top.
- The counts A, B, C, and D are equivalent to the counts of each "type" of pair.
- N is the number of pairs.
- The total number of people sampled = 2*N

(1)

		Control E	Control $\bar{E}$
Case	E	X	
	$\bar{E}$		

(both case and control are E)

(2)

		Control E	Control $\bar{E}$
Case	E	X	
	$\bar{E}$		

(case is E, control is $\bar{E}$)

(3)

		Control E	Control $\bar{E}$
Case	E		
	$\bar{E}$	X	

(case is $\bar{E}$, control is E)

(4)

		Control E	Control $\bar{E}$
Case	E		
	$\bar{E}$		X

(both case and control are $\bar{E}$)

Organization of matched pair data

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

N

Jewell. *Statistics for Epidemiology*. 2004. 4

And then to analyze the pair matched data, we sort of sum up and calculate how many pairs are in each of these four different strata. So the lower table shows a new type of two by two table. At the very bottom here on the right, with the number of exposed versus unexposed cases on the left and the same layout for controls at the top. So notice that this is a very different two by two table. I can't emphasize that enough.

Now, the counts of a, b, c, and d here are equivalent to the counts of each type of pair from above. So there were the four types of situations with the case exposed, control exposed, case unexposed, control unexposed, and so forth. Those are four different types of pairings. And those pairings are represented in the four cells a, b, c, and d here, in the bottom table.

So the total number of people in the study, since each of these letters represents a pair or two people, would be 2 times the sum of a plus b plus c plus d. So a plus b plus c plus d is often written as n. That would be the number of pairs. The number of people would be 2 times n.

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

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

Concordant pairs
Discordant pairs

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

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

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

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

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

Organization of matched pair data

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

N

Jewell. *Statistics for Epidemiology*. 2004. 5

So let's talk about how to analyze data when they're presented in this format, in this special matched pair data layout for a case control study. So down at the bottom again, we see the organization of the matched pair data where all the pairs are represented. But one thing to note is in our four different types of pairs, which we see above in table 16.3, type one, type two, type three, and type four.

In types one and four, the case and the control are the same with respect to exposure. In type one, both the case and the control, that's the d in the d bar, are exposed. And in case four, both the case and the control, the d and the d bar, are unexposed.

These are said to be concordant cells and they provide us with no information because there is no difference between case and control, so we don't learn anything about whether exposure affected case and control differently. What we want to look at are cells b and c where we have the discordant pairs. And in these pairs, you can see that the case and the control differ with respect to their exposure status.

So in table number two, the case is exposed and the control is unexposed. And in table number three, the control is exposed but the case is unexposed. These cells are called discordant cells and they provide us with information to help us distinguish the effect of the exposure differentially on the case and the control.

And you can see the Jewell textbook for this on page 261, but the odds ratio turns out

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

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

Concordant pairs
Discordant pairs

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

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

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

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

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

Organization of matched pair data

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

N

Jewell. *Statistics for Epidemiology*. 2004. 5

to be represented by the ratio of cell b over cell c. So if we divide b by c, this gives us the estimate of the odds ratio. And the intuition about this is that b and c are the pairs in which we have some variability in the exposure.

Because if there's no variation exposure, as I was saying, there's no way to look at the relationship between exposure and disease in terms of seeing how they differ.

Example

- Study of whether spontaneous abortion history is related to coronary heart disease, possibly due to endocrine effects.
- Matched case-control study
  - **Cases**: Had coronary heart disease (CHD)
  - **Controls**: Did not have CHD
  - **Exposure**: At least one spontaneous abortion
  - **Matching factors**: age and location of residence

Jewell. *Statistics for Epidemiology*. 2004. 6

So let's go through an example. I think this will make it very clear.

This is a study of whether spontaneous abortion history is related to coronary heart disease possibly due to endocrine effects. So this is a matched case control study in which the cases were defined as people who had coronary heart disease, we'll abbreviate as CHD. The controls did not have coronary heart disease.

The exposure was at least one spontaneous abortion. And the factors on which the cases were matched to controls were age and location of residence. So for a given case with a certain age and location of residence, a control was found with the same age and location of residence. So this creates matched pairs.

Example

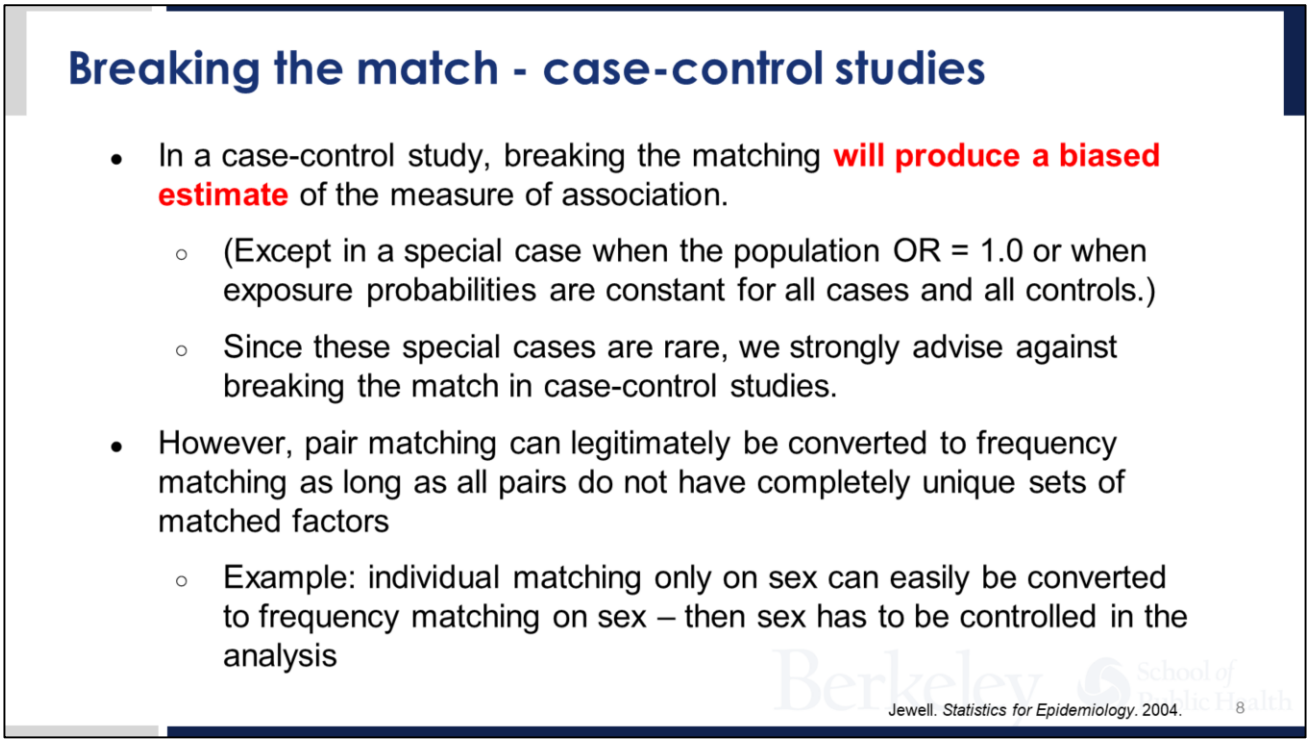
		Control	
		≥1 SA	No SA
Case	≥1 SA	7	18
	No SA	5	20
			50

- "SA": spontaneous abortions
- Odds Ratio = B/C = 18/5 = 3.6
- This finding suggests that a history of at least one spontaneous abortion increases the odds of coronary heart disease.

Jewell. *Statistics for Epidemiology*. 2004. 7

Here's the table displaying the 50 pairs. Now remember, each of these numbers in each of the cell, 7, 18, 5, and 20, represents pairs of people. So in total, there are a hundred individuals in this study. We're just showing the pairs in the table. So the abbreviation here SA is for spontaneous abortions. And the odds ratio then, of course, is calculated as b over c, or 18 over 5, which is 3.6.

So this is suggesting that having a history of at least one spontaneous abortion increases the odds of coronary heart disease compared to not having a history of spontaneous abortion. So the odds is 3.6 times higher in those who've had a spontaneous abortion compared to those who did not.



Finally, let's talk about a situation we call breaking the match. That is, if in a case control study we break the match, we're going to find that we have a biased estimate of the measure of association. The only time this isn't true is a very special case when the population odds ratio is 1, or when the exposure probabilities are constant for all cases and all controls. And those are very unusual circumstances.

So since these are rare, our advice is not to break the match in case control studies. In other words, breaking the match means ignoring the matching and just analyzing the individuals in the study as individuals, not analyzing them in a matched pair way.

However, you can legitimately convert pair matching to frequency matching as long as all the pairs do not have completely unique sets of match factors. So for example, individual matching only on sex can easily be converted to frequency matching on sex. But then sex has to be controlled in the analysis. So if you break the analysis, you then have to control for it.

Risk ratio formula for matched pair data - cohort study

Note the different 2x2 table layout than for case-control studies.

Risk Ratio:

$$\frac{\text{Proportion of exposed with disease}}{\text{Proportion of unexposed with disease}}$$

$$= \frac{(A + B) / (A + B + C + D)}{(A + C) / (A + B + C + D)}$$

$$\mathbf{= (A + B) / (A + C)}$$

		Unexposed	
		Disease	No disease
Exposed	Disease	A	B
	No disease	C	D

Here's a table describing the approach to estimating the risk ratio with matched pair data in a cohort study. And note that the table setup for a matched pair cohort study is different than the two by two table for a matched pair case control study. In the case control study, our controls were in the columns and our cases were in the rows.

It was set up that way because this is a key feature of case control studies. We are setting up our groups based on outcome status. By contrast, in a cohort study, our groups are determined by their exposure status. The tables for our matched pairs reflect that difference.

So the matched pair table for the cohort study has unexposed people in the columns and exposed peoples in the rows. Because the table setup is different, our formula for the risk ratio is also different. It's no longer cell b over cell c. It's rather the sum of a plus b divided by the sum of a plus c.

And here, the risk ratio is representing for us the proportion of those exposed with the disease divided by the proportion of those unexposed with the disease. So this is just a formula to learn and know, and it is different than the formula that's used in the matched pair case control situation.

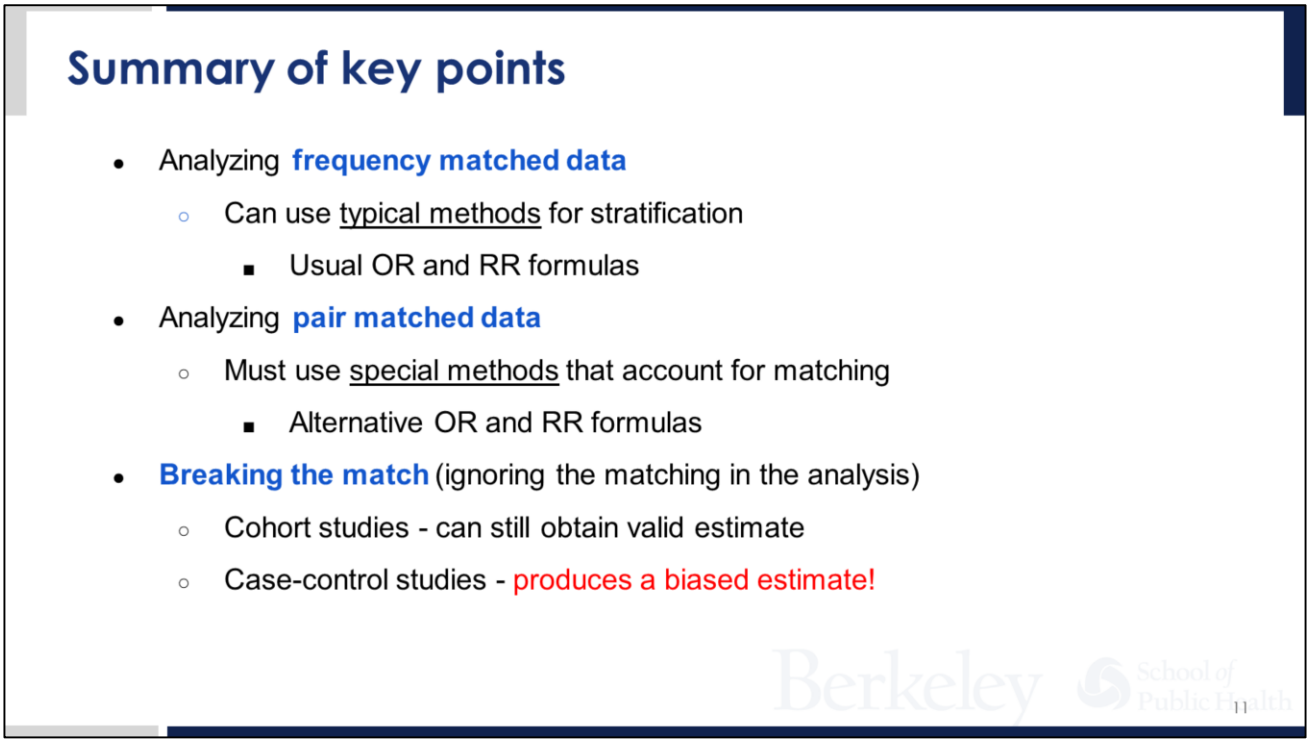
Breaking the match - cohort studies

- Sometimes we conduct a matched study and then are tempted to ignore the matching during statistical analyses.
- In a cohort study, breaking the matching does not prevent us from estimating **a valid estimate** of the measure of association.
  - However, the variance estimates for the measure of association will be incorrect if they ignore the matching.
  - If there is differential misclassification of the outcome, we need to control for the matched factor to obtain a valid effect estimate.

Jewell. *Statistics for Epidemiology*. 2004. 10

Sometimes we conduct a match study and then were tempted to ignore the matching during statistical analyses. In a cohort study, breaking the match doesn't prevent us from estimating a valid estimate of the measure of association. However, the variance estimates for the measure of association will be incorrect if they ignore the matching.

That is, the confidence intervals won't be done correctly. We're not showing you the full derivation of that here, but just so you know that. And if there's differential misclassification of the outcome, we need to control for the match factor in order to obtain a valid effect estimate. So if we break the match in a cohort study, we need to control for that matched factor to obtain a valid effect estimate.



So to summarize, we talked about three different topics related to matched analysis. First we talked about analyzing frequency match data, where we could use typical methods and stratify by our matching variables using traditional odds ratio or relative risk formulae. We talked about analyzing pair matched data, where we have to use special methods that account for the matching. And we went over some specific formulas to do that that gave us the odds ratio and relative risk.

And finally, we talked about breaking the match, or ignoring the matching in the analysis. So in cohort studies, we can still obtain a valid estimate. But in case control studies, we end up having a biased estimate if we break the match.

Causal inference methods to study nonrandomized, preexisting development interventions

Benjamin F. Arnold[a,1], Ranjiv S. Khush[b], Padmavathi Ramaswamy[c], Alicia G. London[b], Paramasivan Rajkumar[c], Prabhakar Ramaprabha[c], Natesan Durairaj[c], Alan E. Hubbard[a], Kalpana Balakrishnan[c], and John M. Colford, Jr.[a]

[a]School of Public Health, University of California, Berkeley, CA 94720-7358; [b]Aquaya Institute, San Francisco, CA 94129; and [c]Department of Environmental and Health Engineering, Sri Ramachandra Medical College and Research Institute, Porur, Chennai 600116, Tamil Nadu, India

Edited* by Kirk R. Smith, University of California, Berkeley, CA, and approved November 2, 2010 (received for review July 2, 2010)

Empirical measurement of interventions to address significant global health and development problems is necessary to ensure that resources are applied appropriately. Such intervention programs are often deployed at the group or community level. The gold standard design to measure the effectiveness of community-level interventions is the community-randomized trial, but the conditions of these trials often make it difficult to assess their external validity and sustainability. The sheer number of community interventions, relative to randomized studies, speaks to a need for rigorous observational methods to measure their impact. In this article, we use the potential outcomes model for causal inference to motivate a matched cohort design to study the impact and sustainability of nonrandomized, preexisting interventions. We illustrate the method using a sanitation mobilization, water supply, and hygiene intervention in rural India. In a matched sample of 25 villages, we enrolled 1,284 children <5 y old and measured outcomes over 12 mo. Although we found a 33 percentage point difference in new toilet construction [95% confidence interval (CI) = 28%, 39%], we found no impacts on height-for-age *Z* scores (adjusted difference = 0.01, 95% CI = −0.15, 0.19) or diarrhea (adjusted longitudinal prevalence difference = 0.003, 95% CI = −0.001, 0.008) among children <5 y old. This study demonstrates that matched cohort designs can estimate impacts from nonrandomized, preexisting interventions that are used widely in development efforts. Interpreting the impacts as causal, however, requires stronger assumptions than prospective, randomized studies.

impact evaluation | study design | propensity score matching | community-led total sanitation | open defecation

In 2000 the United Nations member states agreed upon the Millennium Development Goals (MDGs), which formalized the global community's renewed commitment to solve some of the world's most intractable health and development problems. The MDGs set aggressive targets for 2015 in core metrics, such as reducing by two-thirds the under 5 y population mortality rate and reducing by half the population without access to safe drinking water and basic sanitation. Governments, foundations, and nongovernmental organizations (NGOs) have subsequently increased investment in global health and development programs and have relied on the scientific community to help rigorously measure the impact and cost effectiveness of the interventions. Such empirical measurement is necessary to guarantee that resources are applied in the best possible way (1).

Many development programs use community interventions that deploy treatments at the group level, because they change the physical or social environment, because they cannot be delivered to individuals, or because they wish to capture group-level dynamics. The gold standard for inference in community interventions is a community-randomized trial because the design eliminates confounding bias (2). Bias from other sources can result from frequent measurement (3) or lack of blinding treatment (blinding is rarely possible for community interventions) (4, 5). Even if unbiased, trials must evaluate treatments that are amenable to randomization and typically estimate the average effect of an intervention under ideal conditions (delivery and compliance) in populations most likely to benefit; it is widely acknowledged that treatment effects estimated in such trials can differ from those obtained when the intervention is deployed in the general population (6). Measuring intervention sustainability using prospective trials can also be difficult due to logistical complexity, short funding cycles, and rare sequential awards (7).

The gap between the evidence generated by most community-randomized trials and information that is directly useful to policy makers suggests that studies of nonrandomized, preexisting community interventions implemented by governments and NGOs could contribute both unique and complementary data to inform evidence-based decisions. The sheer number of such community interventions, relative to randomized studies addressing the same issues, speaks to a need for a more rigorous methodology with which to evaluate the impact of the interventions used. We define "nonrandomized, preexisting" interventions as those that were designed and deployed before a structured scientific study.

In this article we draw on the potential outcomes model for causal inference (8–11) to motivate a matched cohort design that enables scientific learning from preexisting, real-world implementation programs under a reasonable set of conditions (described below). Causal inference methods have been well articulated in the statistics and economics literature for decades, but have only recently gained popularity in epidemiology. Here, we frame the matched cohort design in terms of potential outcomes—an extension of prior epidemiologic literature on the design (4, 12)—and tailor it to studies of preexisting, community interventions. The design we propose is most relevant for evaluations with a nonrandomized, predefined intervention group of communities, baseline (pretreatment) data on key confounding variables, and finite resources so that outcome measurement is not possible in all communities. The design naturally estimates the average treatment effect among those most likely to receive an intervention from providers who will actually deliver it—a policy-relevant quantity (1)—and the design enables rapid collection of data about intervention sustainability. We illustrate the usefulness and limitations of the approach with a village-level sanitation mobilization, water supply, and hygiene education intervention conducted in rural Tamil Nadu, India. We believe this general methodology will be useful to study a wide range of preexisting, development interventions beyond the sanitation, water, and hygiene sector.

Materials and Methods

Potential Outcomes Model for Causal Effects. Our approach to evaluating preexisting interventions is grounded in the Neyman–Rubin potential outcomes model (8–11). Let $Y_{i,1}$ denote the potential outcome for community i if the community receives an intervention (treatment), and let $Y_{i,0}$ denote its

Author contributions: B.F.A., R.S.K., P. Ramaswamy, A.E.H., K.B., and J.M.C. designed research; R.S.K., P. Ramaswamy, A.G.L., P. Rajkumar, P. Ramaprabha, and N.D. performed research; B.F.A., A.E.H., and J.M.C. analyzed data; and B.F.A., R.S.K., P. Ramaswamy, A.E.H., K.B., and J.M.C. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence should be addressed. E-mail: benarnold@berkeley.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1008944107/-/DCSupplemental.

potential outcome if it does not receive treatment. The treatment effect for community i is $\psi_i = Y_{i,1} - Y_{i,0}$, but only one potential outcome can ever be observed at the same time. If treatment is randomized, then the treatment assignment (A) is independent of the potential outcomes ($A \perp\!\!\!\perp Y_{i,1}, Y_{i,0}$), and the average difference between treatment groups in observed outcomes, Y_i, is an unbiased estimate of the individual community treatment effect: $\widehat{\psi} = E[Y_{i,1} - Y_{i,0}] = E[Y_{i,1}] - E[Y_{i,0}] = E[Y_i \mid A_i = 1] - E[Y_i \mid A_i = 0]$. Valid inference further requires that the units of intervention (communities) are independent—the treatment of one unit does not influence outcomes in another—and that those treated receive the same treatment (10).

In observational studies treatment assignment A is not random, and so the design does not guarantee that treatment is independent of the potential outcomes. There are usually characteristics (covariates) W that are common causes of both receiving treatment and the outcome and confound the unadjusted comparison of means. The potential outcomes model requires a "strong ignorability" assumption to identify unbiased treatment effects in observational studies (13). Strong ignorability states that all W are measured (no unmeasured confounding) and the treatment and control groups overlap for all combinations of W ($0 < P[A = 1 \mid W] < 1$). The assumption of no unmeasured confounding cannot be evaluated empirically and is a central problem for observational studies.

One approach to weaken the assumption slightly is to target a conditional average treatment parameter: The average treatment effect among the treated (ATT), $\psi_i^{\text{ATT}} = Y_{i,1} - Y_{i,0} \mid A_i = 1$. Estimating the ATT weakens the covariate overlap assumption because it estimates the average effect in the subsample of treated units and thus requires that overlap exist between treatment and control groups at levels of $W \mid A = 1$ rather than the full distribution of W. Observational studies of preexisting interventions are usually constrained to estimating the conditional ATT parameter because without randomization interventions are usually targeted to, or adopted by, a self-selected group that is a nonrandom subset of the total population. However, the ATT is still a very policy-relevant parameter: when estimated for preexisting interventions, it is the average effect in the population most likely to receive the intervention given the providers who would actually deliver it.

Matching in the Design to Approximate a Randomized Experiment. Epidemiology and the social sciences have a long history of using matched cohort designs to study interventions and exposures that are not randomly assigned (14, 15). Recent efforts have used the design in prospective group-level intervention studies (16–18) and in preexisting intervention studies (19–21). In a typical scenario, investigators define a study population, and a subset of the population is selected to receive the intervention by a known or unknown process that is not random. Investigators have information about important confounders, but have not measured outcomes. Matched cohort studies incorporate nonrandom sampling from the study population so that the observed covariate distribution in the control group overlaps and closely matches the covariate distribution in the treatment group. In practice, the design naturally estimates the ATT and is consistent with a general approach of first assembling a control group that is as similar as possible to the treatment group using matching methods and then adjusting for any residual confounding using some form of regression (10, 11).

It is well established that exact matching methods fail to find matches for many treated units in finite samples because the dimension of the joint covariate distribution is too large for nonparametric inference (10). There are many multivariate matching approaches to help address this problem (11, 15, 22). One of the most common is propensity score matching, which simplifies the problem of matching on large numbers of covariates by collapsing the covariates into a single scalar—the propensity score—and then matching treatment and control communities using a one-dimensional match on the propensity score (23, 24). The propensity score is the probability of receiving treatment given a set of baseline covariates, $P(A = 1 \mid W)$, which is unknown for observational studies and must be estimated, usually with a logistic regression. There are numerous ways to match treatment and control units using functions of the propensity score, including nearest-neighbor matching, Mahalanobis distance matching, and optimal matching (11). Sekhon (15) discusses the limitations of propensity score matching in realistic scenarios, where covariates are poorly behaved and not linearly related to the outcome, and proposes a genetic matching algorithm that searches for a matched sample with optimal balance in W.

Whatever matching technique is used, matching does not solve the fundamental problem of unmeasured confounding in observational studies; however, selecting a matched control group in the design stage before measuring outcomes has important advantages (11). First, restricting field data collection to matched treatment and control communities is cost effective because it prevents outcome measurement in extraneous control communities that will not help estimate the ATT parameter. Second, matching helps guarantee that the observed covariate distributions in the treated group overlap with the control group, which enables the analysis to rely less on parametric statistical models and the assumptions they require (25). Matching also accounts for arbitrarily complex relationships between the treatment and covariates, which would need to be modeled explicitly if using regression alone (10). Finally, compared with post hoc statistical adjustment, matching can increase the statistical efficiency of difference parameters, which are useful contrasts for intervention studies (4, 12).

Matched Cohort Designs for Preexisting, Community Interventions. Fig. 1 provides an overview of the design. The innovative components of the design are its use of retrospective, baseline (preintervention) data at the community level to match intervention communities to control communities and its use of propensity score matching—or another alternative multivariate matching approach—to overcome the practical limitations of exact matching in finite samples.

A challenge of studying preexisting interventions is that investigators do not control the intervention, and many community-level interventions that are planned outside of the scientific process have characteristics that make them impossible to evaluate. Before evaluating a nonrandomized, preexisting intervention, investigators should confirm that the intervention meets basic conditions that will enable a valid study (Table 1). In this article we focus on community interventions that are deployed to known geographical units, such as rural villages or neighborhoods in urban areas. We make this restriction because the availability of baseline data collected for purposes other than the study at hand is a core component of the design. These data are typically available for administrative units with known geography (as in a national census), but in theory the design applies to any unit of intervention.

Threats to Validity. ***Unmeasured confounding.*** Because the matched design for preexisting interventions relies on data collected in the past—often independent from the study—it is likely that the data available to match will be incomplete or poorly measured. Matching will improve the balance for measurable characteristics, but is unlikely to remove all differences between treated and control communities so the strong ignorability assumption is unlikely to hold. Matching in the design does not preempt subsequent data analysis (11). Investigators can conduct additional statistical adjustment using data collected in the field study, but they must make a reasoned argument that adjustment covariates could not fall on the causal path between the intervention treatment and outcome of interest (10). If pretreatment outcomes can be measured retrospectively, then the change in the outcome can be compared between treatment and control groups. This "difference-in-differences" parameter removes time-invariant unmeasured confounding assuming the two groups would have had parallel outcome trajectories absent treatment (11). As a robustness check, we recommend falsification tests, where the analysis is repeated for outcomes that could not be influenced by the treatment to investigate whether other interventions or characteristics correlated with treatment could account for the results.

Informative censoring. In the time that elapses between the baseline measurement used to define the study population and the postintervention out-

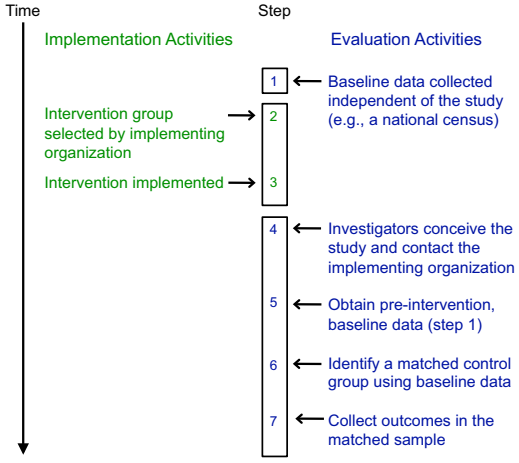


Fig. 1. Overview of the matched cohort design for preexisting interventions.

Table 1. Necessary conditions for matched cohort studies of nonrandomized, preexisting community interventions

	Condition, rationale, and example from this evaluation
1	A partnership with the implementing organization. The implementing organization is the key provider of information about the intervention components, how the intervention beneficiaries were selected, and the timeline and location of activities. Example: We partnered with Water.org and Gramalaya through their funding organization.
2	Sufficient intervention scale. Each community is the independent unit of intervention and it is unusual to have adequate power without at least 8–10 communities per group (2). Example: The intervention included 12 independent communities.
3	Uniformity of the intervention across communities. A relatively uniform intervention is necessary to define and estimate a common treatment effect across communities (in practice, implementation will often vary slightly across communities). Example: The NGOs implemented sanitation, water supply, and hygiene education improvements to raise all communities to a high level of coverage for all three components.
4	Availability of control communities. Control communities are necessary to provide a counterfactual comparison group. Ideally, there should be at least 2 potential control communities for every treatment community. Example: We started with 240 potential control communities from neighboring blocks.
5	Community independence. As with community randomized trials, all units of intervention must be independent with respect to effect of the intervention on outcomes (i.e., no spillover effects). Example: We selected control communities from separate administrative blocks to prevent spillover. We also ensured that communities were qualitatively independent during a rapid assessment following the match but before data collection.
6	Availability of baseline (preintervention) data. Baseline data that include key confounding covariates are used for matched sampling of communities. Baseline data provide a basis for judging baseline comparability of groups in the matched sample. They should reflect conditions at the time of intervention community selection. Example: Census 2001 and Tamil Nadu Water Supply and Drainage board 2003 data collected in the 2 y before the program started included key sanitation, water, and socioeconomic covariates at the community level.

come measurement, individuals within communities will exit the study population (commonly referred to as censoring). If censoring is a common effect of both the intervention treatment and the outcome, then it is informative and will cause bias (26). Informative censoring is a potential source of bias in all study designs, but in prospective designs, characteristics of individuals who exit the study population are available to assess whether censoring is informative. In studies of preexisting interventions, the censored individuals are never measured and so investigators have no direct information about the magnitude of censoring or characteristics of those censored.

Measurement error. If outcomes or exposures are measured retrospectively in the postintervention survey, then they will likely be measured with more error than if they had been measured contemporaneously. Measurement error will cause bias unless it is independent of both treatment status and outcomes (27). Limiting the recall period over which outcomes are measured and using objective outcomes rather than those that rely on self-report can reduce measurement error.

Sampling bias. Sampling bias is possible during community selection or, if outcomes are measured below the community level, in the selection of units below the community level. Investigators should evaluate the completeness of baseline data used to select communities, as incomplete sampling frames could lead to systematic bias. If outcomes are sampled from within the community, then they should be collected from a random sample.

Application of the Design

Sanitation, Water Supply, and Hygiene Intervention in India. Between 2003 and 2007 two NGOs, Water.org and Gramalaya, implemented a combined environmental intervention in 12 rural villages near the city of Tiruchirappalli in Tamil Nadu, India. The intervention combined water supply improvements and repairs with sanitation and hygiene behavior change campaigns that used similar demand mobilization to India's Total Sanitation Campaign. Intervention details varied slightly by village (Table S1), and its intent was to bring all villages to a high level of water supply, sanitation access, and hygiene knowledge. *SI Materials and Methods* includes details of the intervention and study location. The primary objective of the field study was to revisit households after the conclusion of intervention activities to assess outcomes compared with a control group matched on preintervention characteristics. Outcomes included sanitation, water and hygiene conditions and behavior, and health in children <5 y old measured by caregiver-reported diarrhea and anthropometric growth. Diarrhea and child weight measure acute illness in young children, whereas height measures cumulative effects of acute diarrheal illness and chronic intestinal enteropathy caused by repeated exposure to gastrointestinal pathogens (28–30).

Control Selection and Outcome Measurement. The intervention was not randomized and was deployed in villages that were purposely selected by the NGOs. To help reduce potential bias due to differences between intervention and control villages at baseline, we selected control villages with a combination of restriction, propensity score matching on baseline characteristics, and rapid assessment in late 2007. Fig. 2 summarizes the selection process, and *SI Materials and Methods* includes a more detailed description. We enrolled a random sample of up to 50 households per village with children <5 y old. Between January 2008 and April 2009 we visited each participating household once per month for a total of 12 visits. All data collection followed protocols approved by the institutional review boards at the University of California, Berkeley, and Sri Ramachandra Medical College, Chennai, India, and all participants provided informed consent. *SI Materials and Methods* includes details of our exposure and outcome measurement methodology, as well as our statistical analyses using the matched cohort sample. Briefly, we collected detailed information about sanitation conditions and practices, water sources and water quality, and hygiene indicators and handwashing knowledge. In each visit, we collected symptoms of diarrhea, respiratory illness, and general illness in children <5 y old (7 d recall). In the first and last surveys we collected anthropometric growth measurements for children <5 y old. We measured the change in private toilet ownership and water supply between 2003 and 2008 on the basis of household reports in 2008 (retrospective recall for 2003). For all other outcomes we compared groups using postintervention outcomes measured in the 2008–2009 field visits. All estimates are conditional on the matching process using baseline confounders (Fig. 2 and *SI Materials and Methods*). We conducted adjusted

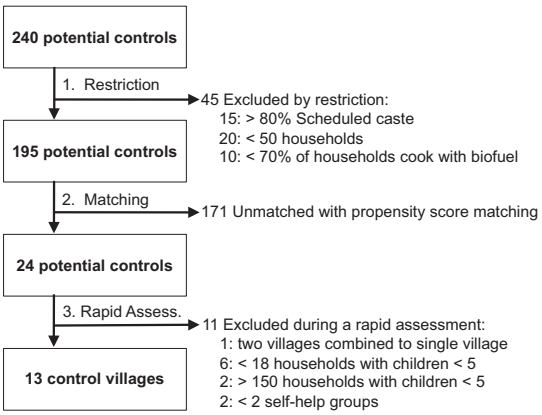


Fig. 2. Control village selection process in the Tamil Nadu study.

analyses using a marginal, *g*-computation estimator with a village-level stratified bootstrap for inference (31).

Matched Cohort Characteristics. The matched cohort design led to a set of matched intervention and control villages that were very similar at baseline, and the approach improved balance greatly for the most imbalanced characteristics (income, biofuel use, sanitation, and water supply) (Table 2). Household characteristics such as durable goods ownership, community participation, and education level were also very similar in our intervention and control groups, postintervention in 2008 (Table S2). Intervention villages were slightly more agricultural than control villages and consequently there were small differences in home ownership, housing materials, and literacy. Our field study sample included 456 control and 444 intervention households [totals exclude 17 control households and 33 intervention households that either moved between listing and enrollment ($n = 44$) or refused]. The 900 households included 1,284 children <5 y old (648 control and 636 intervention). Of these, 612 (94%) control and 608 (96%) intervention children completed the 12-mo follow-up.

Household Sanitation and Water Infrastructure. At baseline in 2003, intervention and control groups were highly similar in toilet and water infrastructure on the basis of census data (Table 2) and retrospective measurement (Fig. 3). Between 2003 and 2008, intervention households were far more likely to build a new private toilet than controls (change in toilet ownership 2003–2008: 48% vs. 15%, $P < 0.0001$, Fig. 3). The intervention increased toilet coverage greatly in the most socially and economically marginalized households (*SI Materials and Methods* and Fig. S1). Of the private toilets in the sample, 89% were pour–flush toilets with a water seal, 5% were ventilated improved pit latrines, and 5% were unimproved concrete slab pit latrines. Toilets were new: 83% were constructed during the 5-y intervention period (since 2003) and 94% were constructed in the last 10 y. Of the 374 households with private toilets, 94% were classified as functional and in use during inspections over the 12-mo period.

Gains in private and public taps were more modest, and increases between 2003 and 2008 did not differ significantly between intervention and control villages (Fig. 3). All households in the study had improved water sources on the basis of the WHO/Unicef Joint Monitoring Program definition (32), and 93% obtained water from public or private taps fed by ground water-supplied overhead tanks. We observed some fecal contamination in household drinking water samples: 27% (120/441) had ≥10 *Escherichia coli* colony-forming units (cfu) per 100 mL. We also found evidence of microbial contamination from general environmental sources in household drinking water: 84% (2,551/3,026) of samples tested positive for hydrogen sulfide (H_2S)-producing bacteria and 91% (2,755/3,015) of samples had ≥100 total coliform cfu per 100 mL (Table S3). Households with private taps spent a median 50 min per day gathering water vs. a median 75 min for households with public taps.

Table 2. Summary of preintervention characteristics before and after village selection

	All villages		Study sample	
Mean	Control	Intervention	Control	Intervention
Demographic				
Total households	170	161	181	161
Persons per household	5	5	5	5
Scheduled caste, %	19	12	15	12
Children ≤5 y old, %	12	12*	12	12
Female literacy, %	52	48	49	48
Socioeconomic				
Employment rate, %	81	78	79	78
Cultivators, %	27	28	31	28
Agricultural laborers, %	24	33	21	33
Marginal workers, %	19	22	21	22
Females work, %	74	69	71	69
Panchayat income (Rp/person)	12,255	7,470***	7,143	7,470
Per-capita cattle ownership	4	4	5	4
Use banking services, %	29	25	25	25
Use biofuel for cooking, %	91	97**	96	97
Own radio, %	43	43	38	43**
Own television, %	21	16	17	16
Own scooter/moped, %	10	10	9	10
Sanitation and water				
Private toilet/latrine, %	15	8**	9	8*
Open defecation, %	85	92**	91	92*
Tap water (private/public), %	75	76	75	76
Hand pump, %	12	14**	18	14
Other water source, %	13	10*	7	10
Persons per hand pump	260	302	240	302
Persons per deep bore well	437	679**	510	679
Water supply level (lpcd)	12	15**	14	15
No. of villages	240	12	13	12

Authors' calculations using India National Census 2001 and Tamil Nadu Water Supply and Drainage 2003 surveys are shown. lpcd, liters per capita per day; Rp, rupees. Scheduled castes include historically disadvantaged, low rank Indian castes, which are currently under government protection. Kolmogorov–Smirnov test for differences in distribution between control and intervention groups: *$P < 0.1$; **$P < 0.05$; ***$P < 0.01$.

Sanitation and Hygiene Behavior. Households in intervention villages were 11 percentage points less likely to report practicing open defecation (77% vs. 88%) than control households (Table S4). Adult open defecation in intervention villages, which had all been declared "open defecation free," ranged between 35% and 83%. Reductions in open defecation were largest among women and smallest among children <5 y old (Table S4). Households that practiced open defecation reported that adult sites were outside the village (98%), but 91% of sites for children <5 y old were within the village. In households with private toilets, 39% reported that adults practice daily open defecation and 52% reported that children <5 y old practice daily open defecation. The most common answers to an open-ended question about the reasons for continuing to practice open defecation despite owning a toilet were no choice (50%), privacy (26%), convenience (25%), and safety (9%). Discrete hygiene spot checks collected by interviewers show overall moderate hygiene conditions, and intervention households fare the same or worse across a large number of indicators (Table S5). Overall, self-reported hand-

washing with soap was rare: Women reported washing their hands after defecation in 24% of 2,657 caregiver interviews (Table S5).

Privacy and Safety for Women and Girls. Private toilet owners were 28 percentage points more likely to report that women and girls feel safe while defecating during the day or night compared with households without private toilets (81% vs. 53%). Overall, the intervention increased the perception of privacy and safety for women and girls during defecation by 13 percentage points compared with controls (72% vs. 59%, Table S4).

Child Health. We identified 259 diarrhea cases from 14,259 child weeks of observation (mean prevalence 1.8%). The mean diarrhea prevalence was slightly higher in intervention villages than in control villages (1.96% vs. 1.67%), and the two groups differed primarily during the summer months (Fig. S2). In unadjusted analyses, we did not observe differences in diarrhea between children in intervention and control villages [longitudinal prevalence difference (LPD) = 0.003, 95% confidence interval (CI) = –0.002, 0.008]. Adjusted estimates, which account for a large set of potentially confounding characteristics (Table S6), also showed no difference between groups (LPD = 0.003, 95% CI = –0.001, 0.008). Despite low diarrhea prevalence, 53% of the children were stunted, 47% were underweight, and 19% were wasted on the basis of weight-for-height. Over 37% were both stunted and underweight (definitions in *SI Materials and Methods*). Mean *Z*-scores were low for both height (mean = –1.96, SD = 1.69) and weight (mean = –1.86, SD = 1.16). We observed no difference in anthropometric *Z*-scores between intervention and control groups [adjusted difference (adj. diff.) in height = 0.01, 95% CI = –0.15, 0.19; and weight = 0.03, 95% CI = –0.11, 0.17; Fig. S3]. Impacts on height are most likely before age 24 mo (33). Restricting the analysis of height-for-age to children who were most likely to benefit from the intervention (<12 mo old at the conclusion of intervention activities, n = 1,093) did not change our findings (adj. diff. = 0.04, 95% CI = –0.28, 0.36).

Discussion

Evaluations of Preexisting Interventions. In this article we have drawn on causal inference theory to develop an evaluation method for nonrandomized, preexisting interventions. Traditionally, such interventions are often evaluated with a before–after comparison in the intervention group alone or with a postintervention, cross-sectional survey in intervention and comparison groups. Before–after comparisons lack a counterfactual comparison and cannot address what would have happened in the absence of intervention. A postintervention, cross-sectional survey neither demonstrates baseline comparability between intervention and comparison communities nor guarantees overlap between groups in important confounding characteristics. In contrast, under appropriate conditions (Table 1), the matched cohort design that we have proposed can demonstrate baseline comparability between intervention and control groups and ensure overlap for observable baseline characteristics so that a valid counterfactual comparison is possible. The attractive features of the design are that it naturally estimates the average effect of an intervention deployed by actually implementing organizations in populations most likely to receive it and yields information about intervention sustainability without years of prospective follow-up. Our motivating example adds to three previous applications of similar methods (to our knowledge) to evaluate preexisting interventions in development settings (19–21), but prior work has not clearly articulated the design's underlying framework, assumptions, and threats to validity. In *SI Materials and Methods* and Fig. S4, we discuss the details of interpreting intervention sustainability in the context of this design. Although we have framed the design in the context of community interventions, in principle it could apply to any unit of intervention (e.g., households or individuals) if appropriate baseline data are available and the study meets the conditions in Table 1.

Our evaluation from Tamil Nadu illustrates many of the strengths and weaknesses of the design. The use of restriction, propensity score matching, and rapid assessment to select control villages (Fig. 2) led to highly similar intervention and control groups on the basis of key exposures and socioeconomic characteristics at baseline (Table 2 and Fig. 3). Although it remains possible that unmeasured confounding has masked the intervention effect, the extremely good overlap in observable confounding characteristics between groups at baseline and follow-up makes this scenario unlikely (Table 2 and Table S2). As a robustness check, we repeated the analyses using caregiver-reported fever in the previous 7 d among children <5 y old as our outcome. Fever is a nonspecific outcome that should not be influenced by the intervention and thus serves as a falsification test. We found no difference between groups in fever (combined prevalence = 11.8%; adj. LPD = 0.008; 95% CI = –0.006, 0.021). Nonetheless, we relied on matched, postintervention differences for all outcomes besides toilet and tap construction; evaluations that use difference-in-differences estimators by comparing the changes in outcomes from baseline to postintervention could be more robust to unmeasured confounding if baseline outcome measures are available (11).

For matching in the design to reduce bias, baseline data must be accurate and complete with respect to key confounders (Table 1, condition 6). Without meeting this condition, matching is unlikely to improve the comparability of intervention and control groups. For sanitation, water, and hygiene interventions, the major confounding variables and intermediate outcomes are often available from national census data, but this condition may not hold for some development research questions. If investigators use secondary data to match groups, as we did in this study, we recommend a brief qualitative and quantitative rapid assessment to validate the data in matched communities before the full field study. This exercise is consistent with integrating ethnographic "thick description" into the selection process (34)—using checks to ensure that intervention units that appear comparable on the basis of computer records are comparable if observed directly.

An additional weakness of this design is its vulnerability to bias from nonrandom subgroups of the population leaving between the intervention and the evaluation (informative censoring). Because such losses are difficult or impossible to measure retrospectively, the evaluation must rely on plausibility arguments. In

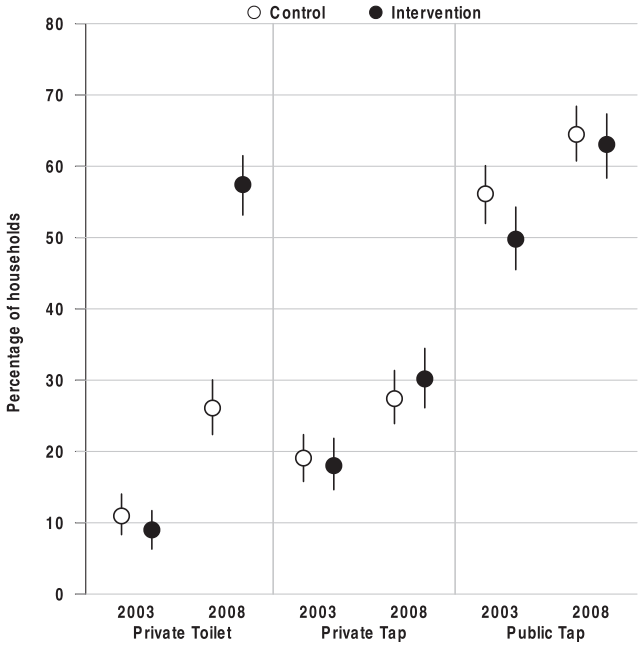


Fig. 3. Population access to private toilets, private water taps, and public water taps in 2003 and 2008. Vertical lines mark bootstrapped 95% confidence intervals. n = 456 control and n = 444 intervention households.

our study sample, just 4.4% of households were lost to follow-up and they were highly similar to those that remained (Table S7). We infer that informative censoring is not a major source of bias in this evaluation.

Combined Interventions, Child Diarrhea, and Growth. Child diarrhea was rare in this population without improved sanitation: 88% of control households practiced open defecation, yet their weekly diarrhea prevalence was just 1.67% over 12 mo. Although we were surprised by this low prevalence, the weekly diarrhea prevalence of all cases reported in the 13 control village health clinics was 1.36% over the same period. (We use the total number of children <5 y old from our original sampling frame as a denominator for the surveillance data. In our control village sample, 80% of diarrhea cases reported visiting the health clinic: $0.8 \times 1.67\% = 1.34\%$, which is very close to the 1.36% prevalence estimated through passive surveillance.)

Our results have a number of implications for government and NGO programs in the sector. They provide evidence that in some populations it is not necessary to combine improvements in water supply, sanitation, and hygiene conditions to achieve very low levels of child diarrhea (35, 36). We infer (although have not tested) that field open defecation is not a primary transmission pathway of diarrhea-causing pathogens for children <5 y old in this population. This study shows that in some rural Indian environments costly sanitation improvements are not guaranteed to have large health benefits, but do improve the perception of privacy and safety for women.

The study also shows that severe growth faltering can persist in populations with rare diarrhea. Poor nutrition is likely a key reason for this (30), but it remains possible that some faltering results from bacterial exposure that is insufficient to cause symptomatic diarrhea, but is sufficient to cause intestinal enteropathy in young children. Enteropathy is hypothesized to cause growth faltering through poor nutrient absorption and low-level immune system stimulation (29). Nutritionists have hypothesized that toilet provision and handwashing with soap could reduce enteropathy and improve growth (37). Our findings indicate that the environmental improvements observed in this study have been insufficient to measurably improve growth (Fig. S3).

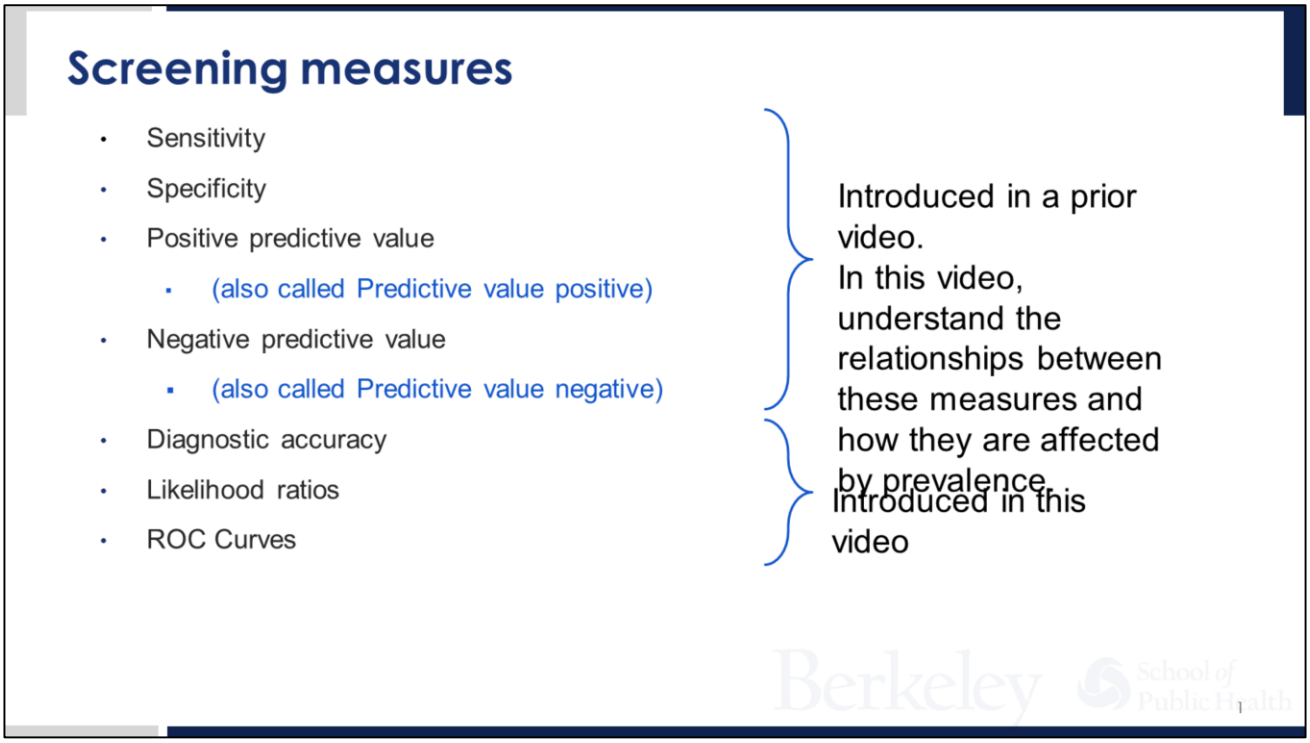
Conclusions. Empirical evaluations of interventions that address the most significant global health and development problems are necessary to ensure that resources are applied most responsibly. It is often difficult or impossible to use randomized studies to measure such impacts. If nonrandomized studies are to be used, they require a more nuanced process of study design and interpretation than randomized studies. In this article we have summarized this process for preexisting interventions, and we expect the methodology could be used to study many types of development programs.

ACKNOWLEDGMENTS. We thank Water.org and Gramalaya for providing us with critical program information and allowing us to evaluate their intervention. We also thank the project field team for their invaluable contribution to collecting the data. This study was funded by a grant from the Open Square Foundation to the Aquaya Institute.

1. Murray CJL, Frenk J (2008) Health metrics and evaluation: Strengthening the science. *Lancet* 371:1191–1199.
2. Murray DM, Varnell SP, Blitstein JL (2004) Design and analysis of group-randomized trials: A review of recent methodological developments. *Am J Public Health* 94:423–432.
3. McCarney R, et al. (2007) The Hawthorne Effect: A randomised, controlled trial. *BMC Med Res Methodol* 7:30.
4. Rothman K, Greenland S (1998) *Modern Epidemiology* (Lippincott–Raven, Philadelphia).
5. Wood L, et al. (2008) Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. *BMJ* 336:601–605.
6. Horton R (2000) Common sense and figures: The rhetoric of validity in medicine (Bradford Hill Memorial Lecture 1999). *Stat Med* 19:3149–3164.
7. Rajan TV, Clive J (2000) NIH research grants: Funding and re-funding. *JAMA* 283:1963.
8. Splawa-Neyman J, Dabrowska DM, Speed TP (1990) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat Sci* 5:465–472.
9. Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 66:688–701.
10. Morgan SL, Winship C (2007) *Counterfactuals and Causal Inference* (Cambridge University Press, Cambridge, UK).
11. Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. *J Econ Lit* 47:5–86.
12. Greenland S, Morgenstern H (1990) Matching and efficiency in cohort studies. *Am J Epidemiol* 131:151–159.
13. Rubin DB (1978) Bayesian inference for causal effects: The role of randomization. *Ann Stat* 6:34–58.
14. Cochran WG (1953) Matching in analytical studies. *Am J Public Health Nations Health* 43:684–691.
15. Sekhon JS (2009) Opiates for the matches: Matching methods for causal inference. *Annu Rev Polit Sci* 12:487–508.
16. Preisser JS, Young ML, Zaccaro DJ, Wolfson M (2003) An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Stat Med* 22:1235–1254.
17. Pattanayak SK, Poulos C, Yang JC, Patil SR, Wendland KJ (2009) Of taps and toilets: Quasi-experimental protocol for evaluating community-demand-driven projects. *J Water Health* 7:434–451.
18. Newman J, et al. (2002) An impact evaluation of education, health, and water supply investments by the Bolivian Social Investment Fund. *World Bank Econ Rev* 16: 241–274.
19. Arnold BF, Arana B, Mäusezahl D, Hubbard A, Colford JM, Jr (2009) Evaluation of a pre-existing, 3-year household water treatment and handwashing intervention in rural Guatemala. *Int J Epidemiol* 38:1651–1661.
20. Cattaneo MD, Galiani S, Gertler PJ, Martinez S, Titiunik R (2009) Housing, health, and happiness. *Am Econ J Econ Policy* 1:75–105.
21. Pradhan M, Rawlings LB (2002) The impact and targeting of social infrastructure investments: Lessons from the Nicaraguan Social Fund. *World Bank Econ Rev* 16: 275–295.
22. Iacus S, King G, Porro G (2009) cem: Software for coarsened exact matching. *J Stat Softw* 30:1–27.
23. Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.
24. Rosenbaum PR, Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat* 39:33–38.
25. Ho DE, Imai K, King G, Stuart EA (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal* 15:199–236.
26. Hernán MA, Hernández-Díaz S, Robins JM (2004) A structural approach to selection bias. *Epidemiology* 15:615–625.
27. Hernán MA, Cole SR (2009) Invited commentary: Causal diagrams and measurement bias. *Am J Epidemiol*, 170:959–962 discussion, 963–964.
28. Schmidt W, et al. (2010) Weight-for-age z-score as a proxy marker for diarrhoea in epidemiological studies. *J Epidemiol Community Health*, 10.1136/jech.2009.099721.
29. Lunn PG (2000) The impact of infection and nutrition on gut function and growth in childhood. *Proc Nutr Soc* 59:147–154.
30. Black RE, et al.; Maternal and Child Undernutrition Study Group (2008) Maternal and child undernutrition: Global and regional exposures and health consequences. *Lancet* 371:243–260.
31. Ahern J, Hubbard A, Galea S (2009) Estimating the effects of potential public health interventions on population disease burden: A step-by-step illustration of causal inference methods. *Am J Epidemiol* 169:1140–1147.
32. UNICEF, WHO (2008) *World Health Organization and United Nations Children's Fund Joint Monitoring Programme for Water Supply and Sanitation (JMP). Progress on Drinking Water and Sanitation: Special Focus on Sanitation* (UNICEF and WHO, New York and Geneva).
33. Victora CG, de Onis M, Hallal PC, Blössner M, Shrimpton R (2010) Worldwide timing of growth faltering: Revisiting implications for interventions. *Pediatrics* 125:e473–e480.
34. Rosenbaum PR, Silber JH (2001) Matching and thick description in an observational study of mortality after surgery. *Biostatistics* 2:217–232.
35. Briscoe J (1984) Intervention studies and the definition of dominant transmission routes. *Am J Epidemiol* 120:449–455.
36. Eisenberg JNS, Scott JC, Porco T (2007) Integrating disease control strategies: Balancing water sanitation and hygiene interventions to reduce diarrheal disease burden. *Am J Public Health* 97:846–852.
37. Humphrey JH (2009) Child undernutrition, tropical enteropathy, toilets, and handwashing. *Lancet* 374:1032–1035.

Lecture: Screening Measures in Depth

Let's spend a bit of time now talking about screening measures in epidemiology.



In earlier videos, we learned the concepts of sensitivity, specificity, positive predictive value which is sometimes called the predictive value positive, the negative predictive value which is sometimes called the predictive value negative. We're going to build on those concepts now to talk about some additional ideas which are diagnostic accuracy, likelihood ratios, and ROC curves. The ROC stands for Receiver Operator Characteristic curves.

Recap: sensitivity and specificity

TABLE 8-4 Schematic representation of the calculation of sensitivity and specificity for a binary variable.

Study's result	Gold standard's result		Total
	Positive	Negative	
Positive	a	b	$a + b$
Negative	c	d	$c + d$
Total	$a + c$	$b + d$	N

Sensitivity = $a/(a + c)$
Specificity = $d/(b + d)$

Sensitivity measures the probability of a **true positive**.

Specificity measures the probability of a **true negative**.

Szklo & Nieto, 3rd Ed. 2

Just as a quick review, let's talk about what sensitivity and specificity represent. This table lays out the results of a test as either positive or negative compared against a gold standard, which has measured whether the patient truly has disease or not. So for example, cell a represents a patient with disease who's test is positive, cell b is a patient who's test is positive but does not have disease, cell c is a patient who has disease but has a negative test, and cell d is a patient who does not have disease and does not have a positive test.

Sensitivity is the concept that measures the probability of a true positive. So that would be calculated from the first column. The positives in the first column would be patients who truly have disease. So if we calculated a over 80 plus c, that's the proportion of the patients with true positive disease whose test is truly positive. That's called sensitivity. And cell d over cell b plus d gives us a calculation-- among the patients who truly do not have disease, how many of them had a negative test. So how accurate is the negative test among the patients without disease, that's specificity.

Relationship between sensitivity and specificity

- When a binary cutoff is used to classify a disease as present/absent based on a continuous measure of disease:
  - Increasing sensitivity decreases specificity.
  - Increasing specificity decreases sensitivity.
- Example: abnormal serum cholesterol is classified as serum cholesterol ≥ 200 mg/dL
- How does changing this cutoff affect sensitivity and specificity?

TABLE 8-5 Comparison of screening values of serum cholesterol under field conditions and values done in a standard laboratory.

Screening values	Standard laboratory values		Total
	Abnormal*	Normal	
Abnormal*	18	19	37
Normal	1	11	12
Total	19	30	49

Sensitivity = 18/19 = 0.95
Specificity = 11/30 = 0.37

Szklo & Nieto, 3rd Ed.

3

There's a relationship between sensitivity and specificity. When a binary cutoff is used to classify a disease as present or absent based on a continuous measure of disease, increasing the sensitivity will decrease the specificity of the test. Increasing the specificity will decrease the sensitivity of the test. So for example, imagine we're talking about an abnormal serum cholesterol, and we're defining it as a serum cholesterol greater than or equal to 200.

Let's think through how changing the cutoff affects sensitivity and specificity. So here's the result comparing the screening values for cholesterol, and we're looking at the standard laboratory values compared to the screening laboratory values. We're treating the standard laboratory values as the truth, and then the screening test as some rapid way to screen for it. And we see the calculation here. The sensitivity of this test is 18 over 19, or 0.95, and the specificity is 11/30, or 0.37.

Relationship between sensitivity and specificity

- Example: change "abnormal" cutoff from serum cholesterol ≥ 200 mg/dL to ≥ 300 mg/dL
- How does changing this cutoff affect sensitivity and specificity?

TABLE 8-5 Comparison of screening values of serum cholesterol under field conditions and values done in a standard laboratory.

Screening values	Standard laboratory values: Abnormal*	Standard laboratory values: Normal	Total	
Abnormal*	10	1	37	11
Normal	9	29	12	38
Total	19	30	49	

Using ≥ 200 mg/dL
Sensitivity: 18/19 = 0.95
Specificity: 11/30 = 0.37

Using ≥ 300 mg/dL
Sensitivity: 10/19 = 0.53
Specificity: 29/30 = 0.97

Szklo & Nieto, 3rd Ed. 4

Let's now change the abnormal cutoff from a serum cholesterol greater than 200 to actually be a serum cholesterol greater than or equal to 300. How does changing this cutoff affect sensitivity and specificity? Well, the new numbers are shown here in bold. So our original sensitivity and specificity was 0.95 and 0.37, but when we change the cutoff to 300, the sensitivity went down dramatically-- 10/19 is 0.3-- and the specificity went up dramatically-- 0.97 or 29/30.

When it makes sense to increase sensitivity at the expense of specificity

- You want to minimize false negatives and you are less concerned about false positives
- Example: Pap smears, which are used to screen for cervical cancer
- Want to detect all possible cases of cervical cancer since it is a serious but treatable disease.
- If the test produces a false positive, the patient will undergo additional testing to confirm they do not have cervical cancer.

There are situations when it makes sense to increase sensitivity at the expense of specificity. For example, if you want to minimize false negatives and you are less concerned about false positives, that would be an example of this situation. Pap smears are one example where a screening test is used to screen for a condition, in this case cervical cancer, and we'd like to detect all possible cases of cervical cancer because it's a serious but treatable disease. So we'd rather accept that we're going to have some false positives using this screening test, but we're OK with that. If the test produces a false positive, the patient will then undergo additional testing to confirm that they do not have cervical cancer.

Sensitivity and specificity depend on the distribution of severity of a disease

- When using one test and the same cutoff point, the amount of potential misclassification in a study is larger if the distribution of values is closer to a true negative.

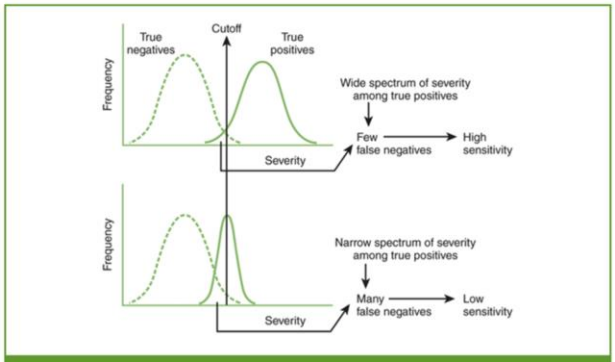


FIGURE 8-5 Two hypothetical situations leading to different estimates of sensitivity, depending on the spectrum of severity of a given condition (e.g.,

Szklo & Nieto, 3rd Ed. 6

Sensitivity and specificity depend on the distribution of the severity of disease. When using one test and the same cutoff point, the amount of potential mis-classifications in a study is larger if the distribution of values is closer to a true negative. So here are two hypothetical situations from your [INAUDIBLE] textbook that lead to different estimates of sensitivity depending on the spectrum of severity of a given condition.

So in the upper example here, we see a wide spectrum of severity among true positives. You see the separation of the curves here with a fixed cutoff that in this example gives us few false negatives and a high sensitivity. But if the distribution of the true positives moves closer to the true negatives, as you see in the second example, now we have a situation where we have many false negatives and a low sensitivity. Notice what's happened here is this narrowing of the spectrum of severity among the true positives has occurred, and it's moved closer to the true negatives.

Sensitivity and specificity estimated with self-reported data

- When disease status is measured through questionnaires, individuals' characteristics may affect their disease classification, introducing bias.
- **Example**:
  - A study of self-reported weight and height found that the sensitivity and specificity of BMI classification varied substantially by age and gender.
- When this is true, external validity of sensitivity and specificity is reduced.

Szklo & Nieto, 3rd Ed. 7

Sensitivity and specificity can be estimated with self-reported data. And when disease status is measured through questionnaires, for example, individuals' characteristics may affect their disease classification, and this can introduce bias. For example, a study of self-reported weight and height found that the sensitivity and specificity of BMI classification varied substantially by age and gender. When this situation is true, the external validity of sensitivity and specificity is reduced.

Recap: PPV and NPV

TABLE 8-4 Schematic representation of the calculation of sensitivity and specificity for a binary variable.

Study's result	Gold standard's result		Total
	Positive	Negative	
Positive	*a*	*b*	*a* + *b*
Negative	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*N*

- **Positive predictive value:** proportion of individuals with a positive test who have preclinical disease
  - PPV = (a / a + b) x 100% = 18 / 37 = 49%
- **Negative predictive value:** proportion of individuals without preclinical disease who test negative
  - NPV = (d / c + d) x 100% = 11 / 12 = 92%

Szklo & Nieto, 3rd Ed. 8

Let's review what positive predictive value and negative predictive value mean. Here again, we have our test result two by two table. The calculation of positive predictive value moves along the rows, so the positive predictive value is along the row of the positive test results. The calculation is a over a plus b, which would give us, in this case, 49%. So what that is saying is that among those who test positive, 49% of them will truly have disease.

Similarly, the negative predictive value calculates along the row with negative testing. The d over c plus d cells represents the proportion of those with negative tests, how many are truly negative for disease. And in this case, 92%.

PPV and NPV are affected by prevalence. Sensitivity and specificity are not.

TABLE 8-4 Schematic representation of the calculation of sensitivity and specificity for a binary variable.

Study's result	Gold standard's result		Total
	Positive	Negative	
Positive	*a*	*b*	*a* + *b*
Negative	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*N*

Prevalence of disease = a + c / N

Sens = a/a+c

- PPV and NPV condition on test results.
- The proportion who test positive depends on how prevalent the disease is.
  - **a** and **c** will be larger if the prevalence is high
  - **b** and **d** will be larger if the prevalence is low
- Prevalence doesn't affect sensitivity and specificity because they condition on true results.

9

The positive predictive value and the negative predictive value condition on the test results. The proportion who test positive depend on how prevalent the disease is. For example, a and c will be larger if the prevalence is high. b and d will be larger if the prevalence is low. Prevalence, however, doesn't affect sensitivity and specificity this way because they condition on true results.

Mathematical relationship between PPV and prevalence

$$PPV = \frac{a}{a+b}$$

1) Multiply by a+b+c+d/(a+b+c+d) and rearrange
2) Multiply by (a+c)/(a+c) and rearrange
3) Multiply by (b+d)/(b+d) and rearrange

$$= \frac{① \ \frac{a}{a+b+c+d}}{\frac{a}{a+b+c+d} + \frac{b}{a+b+c+d}}$$

$$= \frac{② \ \frac{a}{a+c} \times \frac{a+c}{a+b+c+d}}{② \left(\frac{a}{a+c} \times \frac{a+c}{a+b+c+d}\right) + \left(\frac{b}{b+d} \times \frac{b+d}{a+b+c+d}\right) ③}$$

Sensitivity — Prevalence — Sensitivity — Prevalence — 1 - Specificity — 1 - Prevalence

10

PPV = \frac{a}{a+b}
= \frac{\frac{a}{a+b+c+d}}{\frac{a}{a+b+c+d}+\frac{b}{a+b+c+d}}

There are mathematical relationships to understand between the positive predictive value and the prevalence. So by algebraic rearrangements, we can multiply the formula for the positive predictive value by the quantity you see here. And then continuing this line of reasoning, if we multiply by a plus c over a plus c just as an identity, and then multiply by b plus d over b plus d, and rearrange the entire fraction, what we see is that the sensitivity times the prevalence divided by the sensitivity times the prevalence plus 1 minus specificity times 1 minus the prevalence gives us the positive predictive value. So all of this algebraic rearrangement is just to the purpose of showing the relationship between sensitivity, specificity and prevalence.

Mathematical relationship between NPV and prevalence

$$NPV = \frac{d}{c+d}$$

1) Multiply by a+b+c+d/(a+b+c+d) and rearrange
2) ~~Multiply by (b+d)/(b+d)~~ and rearrange
3) Multiply by (a+c)/(a+c) and rearrange

$$= \frac{\frac{d}{a+b+c+d}}{\frac{d}{a+b+c+d} + \frac{c}{a+b+c+d}} \quad (1)$$

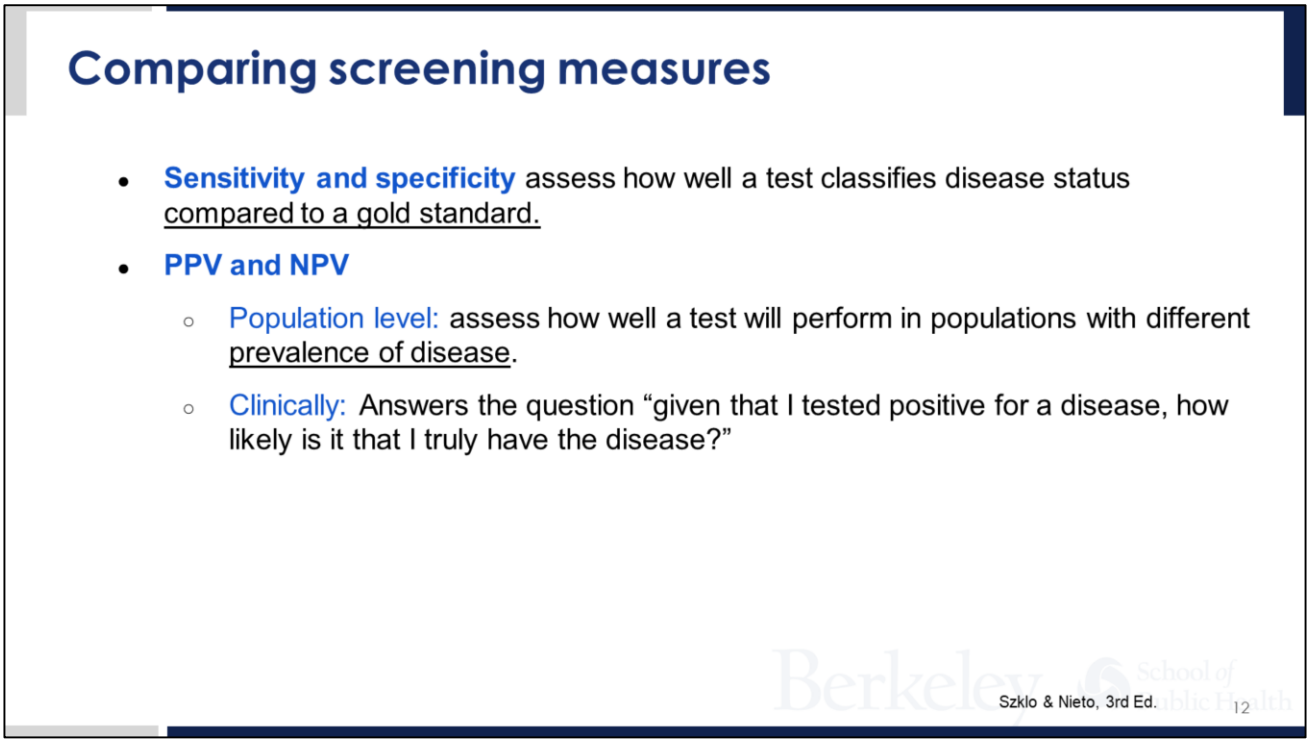
$$= \frac{\frac{d}{b+d} \times \frac{b+d}{a+b+c+d}}{\left(\frac{d}{b+d} \times \frac{b+d}{a+b+c+d}\right) + \left(\frac{c}{a+c} \times \frac{a+c}{a+b+c+d}\right)} \quad (2)\ (3)$$

(2) Specificity; 1 - Prevalence — (3) 1 - Sensitivity; Prevalence

11

NPV = \frac{d}{c+d}
= \frac{\frac{a}{a+b+c+d}}{\frac{a}{a+b+c+d}+\frac{b}{a+b+c+d}}
= \frac{\frac{a}{a+c} \times\frac{a+c}{a+b+c+d} }{\big(\frac{a}{a+c} \times\frac{a+c}{a+b+c+d}\big)+\big(\frac{b}{b+d} \times\frac{b+d}{a+b+c+d}\big)}

We can do similar calculations for the negative predictive value, and you see those done here. I won't step through each of them, but hopefully you can follow the algebra through step 1, step 2, step 3.



Sensitivity and specificity are concepts that assess how well a test classifies disease status compared to some gold standard, whereas positive predictive value and negative predictive value, at the population level, tells us how well a test performs in populations with different prevalence of disease. This is a really critical point. And clinically, these values, the positive predictive value and the negative predictive value, answer the question, given that I have tested positive for a disease, how likely is it that I truly have the disease?

Diagnostic accuracy

TABLE 8-4 Schematic representation of the calculation of sensitivity and specificity for a binary variable.

Study's result	Gold standard's result		Total
	Positive	Negative	
Positive	*a*	*b*	*a* + *b*
Negative	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*N*

- The proportion of results that are correct:
  - a + d / (a + b + c + d)
- Measures overall accuracy combining sensitivity and specificity.

Diagnostic accuracy is a different condition. And if you look at our two by two table again and think about which cells are correct-- quote unquote "correct"-- cell a and cell d are correct, because in cell a, people who truly have the disease test positive for disease, and in cell d, people who truly do not have the disease test negative for the disease. So if you add a plus d divided by the quantity the sum of the cells, that's the overall accuracy. And that combines sensitivity and specificity.

Likelihood ratios

TABLE 8-4 Schematic representation of the calculation of sensitivity and specificity for a binary variable.

Study's result	Gold standard's result		Total
	Positive	Negative	
Positive	a	b	$a+b$
Negative	c	d	$c+d$
Total	$a+c$	$b+d$	N

A likelihood ratio contrasts the proportions of patients with and without a disease for a given positivity criterion.

- Likelihood ratios can be used to connect pre-test and post-test probability of disease.
- **Positive likelihood ratio (LR+)**: is the odds that a positive test result would be seen in a patient with the disease

$$LR^{+} = \frac{sensitivity}{(1-specificity)} = \frac{A/(A+C)}{B/(B+D)}$$

can also be written as:

$$LR^{+} = \frac{p(test+ \mid D+)}{p(test+ \mid D-)} = \frac{p(test+D+)/p(D+)}{p(test+D-)/p(D-)}$$

- **Negative likelihood ratio (LR-)**: the odds that a negative test result would be seen in a patient with the disease

$$LR^{-} = \frac{(1-sensitivity)}{specificity} = \frac{C/(A+C)}{D/(B+D)}$$

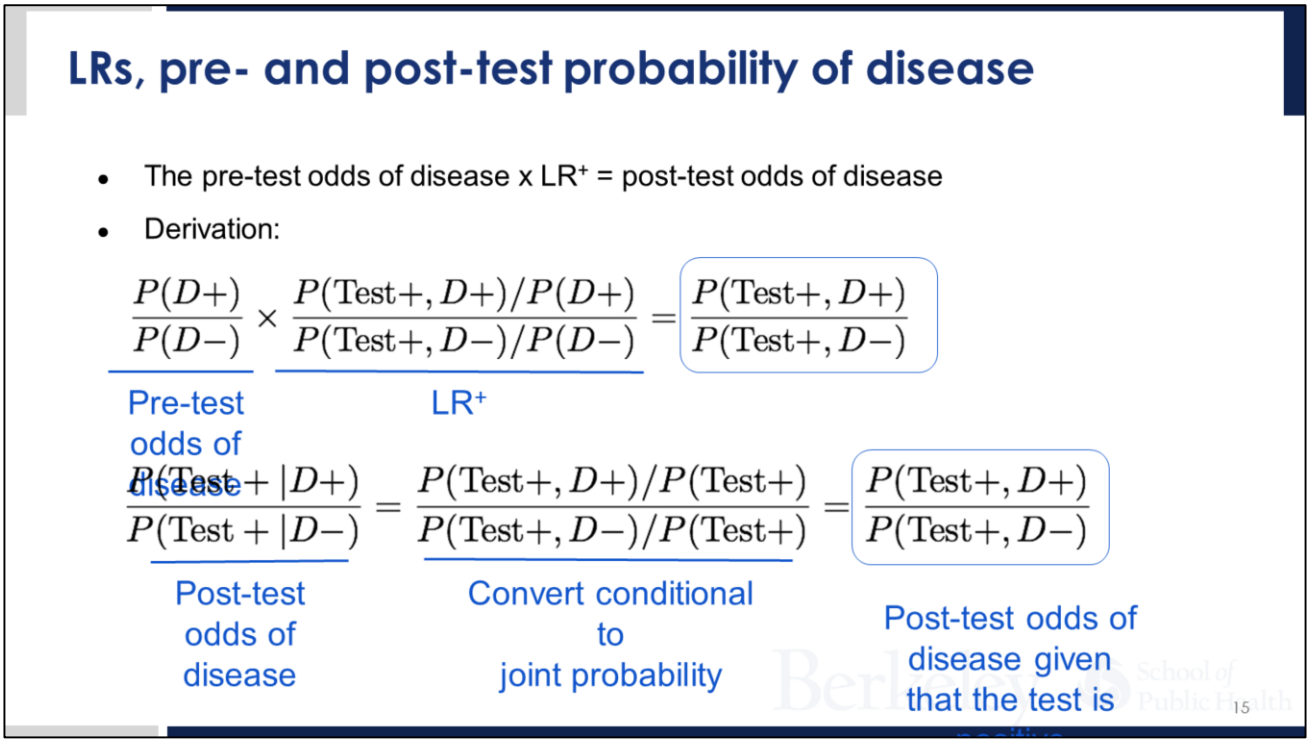
can also be written as:

$$LR^{-} = \frac{p(test- \mid D+)}{p(test- \mid D-)} = \frac{p(test-D+)/p(D+)}{p(test-D-)/p(D-)}$$

Berkeley School of Public Health

14

Yet another concept to use in screening measures is the likelihood ratio, and the likelihood ratio contrasts the proportion of patients with and without a disease for a given positivity criterion. And likelihood ratios can be used to connect our pre-test and post-test probability of disease. So the positive likelihood ratio, or LR positive, is the odds that a positive test result would be seen in a patient with disease. The likelihood ratio positive is defined as the sensitivity over 1 minus the specificity, so we can write that out with letters here. A over a plus c divided by b over b plus d.
Now if you write that out in an alternative form using algebra, you could write that as the probability that the test is positive given that the disease is positive, that's the sensitivity, divided by probability that the test is positive given that the disease is negative, that's 1 minus the specificity. And then you can see how those could be further elaborated on or extended as the probability of the test positive given the disease positive divided by the probability that the disease is positive, divided by the probability that the test is positive given that the disease is negative divided by the probability that the disease is negative.
The negative likelihood ratio is the odds that a negative test result would be seen in a patient with the disease. And you see here how this can be rewritten, but the concept here to capture is that the negative likelihood ratio, or the likelihood ratio negative, is 1 minus the sensitivity over the specificity.



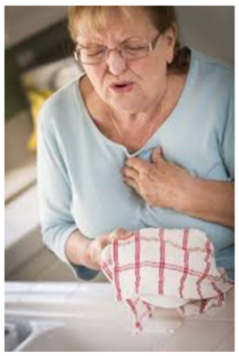
\frac{P(D+)}{P(D-)} \times \frac{P(\text{Test}+, D+) / P(D+)}{P(\text{Test}+, D-) / P(D-)} = \frac{P(\text{Test}+, D+)}{P(\text{Test}+, D-)}

\frac{P(\text{Test}+| D+)}{P(\text{Test}+| D-)} = \frac{P(\text{Test}+,D+)/P(\text{Test}+)}{P(\text{Test}+,D-)/P(\text{Test}+)} = \frac{P(\text{Test}+,D+)}{P(\text{Test}+, D-)}

The likelihood ratios, the pre- and post-test probability of disease are all related to each other, because the pre-test odds of disease times the positive likelihood ratio is the post-test odds of disease. I find the best way to learn this is to work through problems, and we've provided you problems to work through these quantities. You see the derivation algebraically here for this relationship-- that the pre-test odds of disease times the likelihood ratio positive is the probability of the test positive and the disease positive divided by the probability of the test being positive and the disease being negative. This is equal to the post as the odds of disease. This is a very useful relationship, and you see the derivation for it here on this slide.

Example of clinical application

- You are working in an emergency room when a 50 year old woman presents with a complaint of “chest pain”. She is a smoker with a history of heart disease in her father.
- She does not have “classical” symptoms of a heart attack.
- You estimate, based on her story, that there is only a 10% chance that she having a heart attack.
- You decide to apply a new blood test to decide whether to admit her or not. The test has:
  - Sensitivity = 0.86
  - Specificity = 0.95
- What is the post-test probability that she is having a heart attack if the test result is 150 units/liter?



16

Here's an example of a clinical application that I think will make all of these relationships clearer. Assume you're working in an emergency room when a 50-year-old woman presents with a complaint of chest pain. She's a smoker with a history of heart disease in her father. She doesn't have classical symptoms of a heart attack, and so you estimate based just on her story that there's only a 10% chance that she's having a heart attack.

You decide to apply a new blood test to decide whether to admit her or not, and you know that this blood test has a sensitivity of 0.86 six and a specificity of 0.95. Let's calculate the post-test probability that she's having a heart attack if the result of her test is 150 units per liter.

Example of clinical application

- **Pre-test odds of disease:** P(D+) / P(D-)
  - From prior slide, we estimate probability of disease is 0.1
  - Odds of disease = 0.1 / (1-0.1) = 0.11
- **Likelihood ratio + :** sensitivity / 1 - specificity
  - LR^+ = 0.86 / (1 - 0.95) = 17.2
- **Post-test odds of disease:** pre-test odds of disease x LR^+
  - Post-test odds = 0.11 x 17.2 = 1.89
- **Post-test odds of probability:** Probability = odds / (1+odds)
  - Post-test probability = 1.89 / (1+1.89) = 0.65
- We conclude that there is a 65% probability she is having a heart attack. Since this is higher than the pre-test probability, we choose to admit her to the hospital.

Berkeley School of Public Health

17

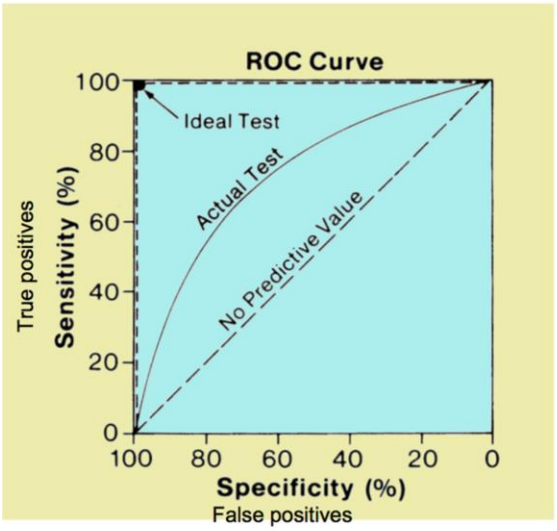
Well, first, let's calculate the pre-test odds of disease. So that's given by the probability of disease divided by the probability of not having disease. That's the odds, the pre-test odds. So from the prior slide, we estimate the probability of disease is 0.1, so the odds of disease would be 0.1 over 1 minus 0.1, or 0.11.

The positive likelihood ratio we defined as the sensitivity divided by 1 minus the specificity. So using the numbers from this example, the positive likelihood ratio would be the sensitivity, 0.86, divided by 1 minus the specificity-- that's 1 minus 0.95. That gives us a quantity of 17.2. So the post-test odds of disease is the pre-test odds of disease times the positive likelihood ratio. So if we do that calculation using the numbers we just derived, we get 0.11 times 17.2, or 1.89. So that's the post-test odds of disease.

Now we have to convert that to a probability. So the way to do that is the probability is equal to the odds over 1 plus the odds. So we have a post-test odds of 1.89. Let's convert that to a probability. That would give us 1.89 over the quantity 1 plus 1.89, or 0.65. So we conclude that there's a 65% probability she is having a heart attack. Since this is higher than the pre-test probability and of concern, we would probably choose to admit her to the hospital.

ROC curve

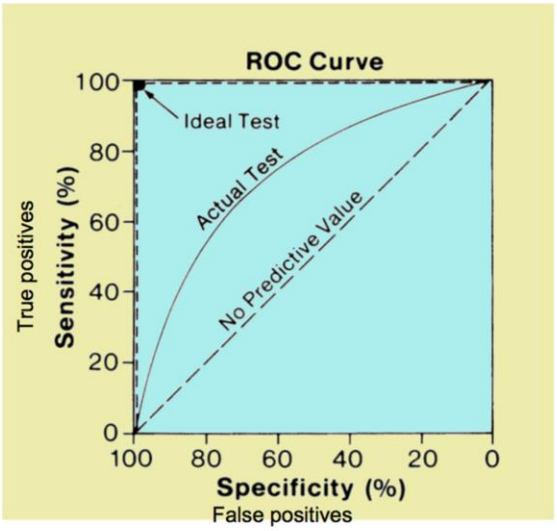
- Allow us to directly compare sensitivity and specificity when assessing how changing a test's cutoff (positivity criterion) affects sensitivity and specificity.
- Plot of the sensitivity against 1-specificity (the false positive rate)
- An ideal test has perfect sensitivity and specificity and has values in the upper left hand corner.



Another tool useful in screening is the ROC curve. The ROC curve allows us to directly compare sensitivity and specificity when we assess how changing a test cutoff or positivity criterion affects sensitivity and specificity. So an ROC curve is a plot of the sensitivity on the y-axis here against 1 minus the specificity on the x-axis. So you see how on the x-axis it goes from 100 down to 0, because it's 1 minus specificity.
So an ideal test would have perfect sensitivity and specificity and have values up in the upper left hand corner. Our actual test might be some intermediate level in between, and if a test had no predictive value, it would be a straight line for true positives versus false positives, or sensitivity versus 1 minus specificity.

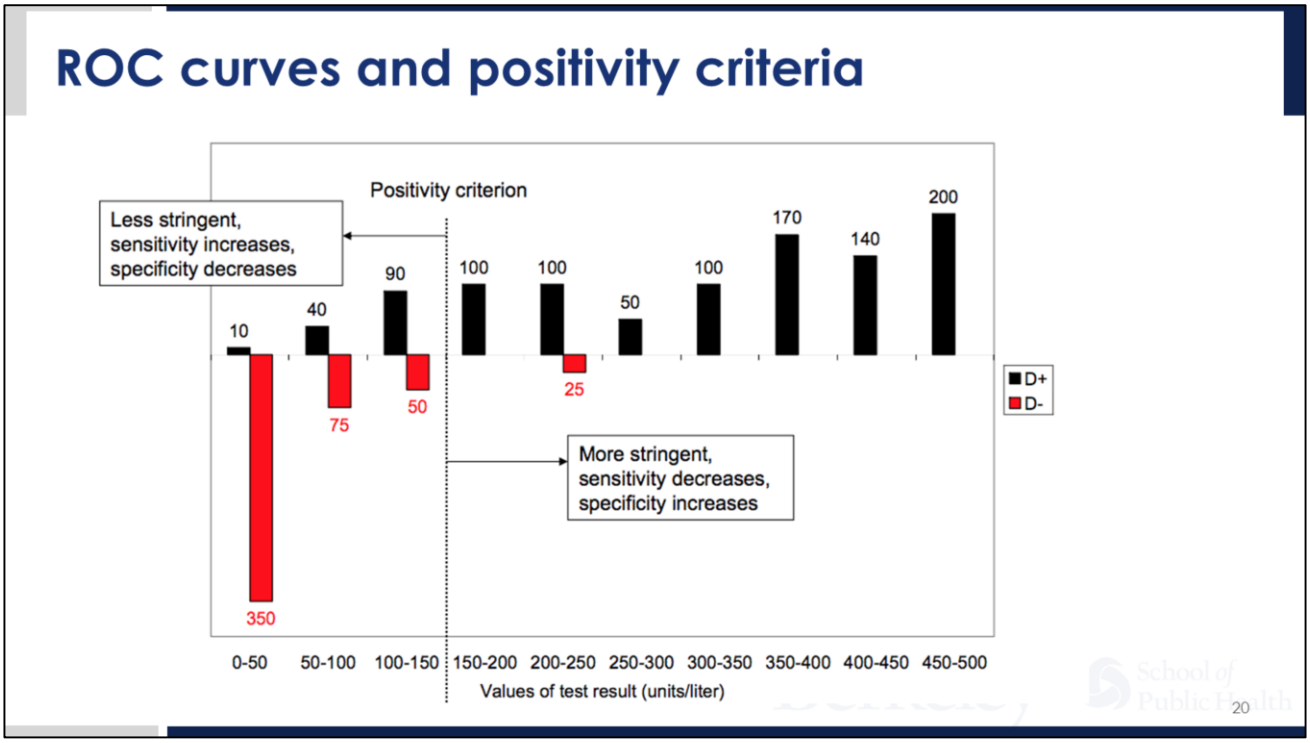
ROC curve

- As the positivity criterion for a test becomes more stringent (criterion has a larger value), the point on the curve corresponding to sensitivity and specificity moves down and to the left (lower sensitivity, higher specificity)
- As the criterion for a test becomes less stringent (criterion has a smaller value), the point on the curve corresponding to sensitivity and specificity moves up and to the right (higher sensitivity and lower specificity)



19

As the positivity criterion for a test becomes more stringent-- that is, the criterion has a larger value-- the point on the curve corresponding to sensitivity and specificity moves down and to the left. That's a lower sensitivity and a higher specificity. In contrast, as the criterion for a test becomes less stringent-- that is, the criterion has a smaller value-- the point on the curve corresponding to sensitivity and specificity moves up and to the right. That is, it has a higher sensitivity and lower specificity.



So let's look at the relationship with these ROC curves and positivity criterion. So on the x-axis of this figure here are the values of our test result, the cholesterol result. And we've set a cholesterol positivity criterion in this example as 150. So less than 150 is negative, greater than 150 is positive. As the positivity criterion becomes lower, less stringent, the sensitivity increases, we capture more cases-- those are the black bars-- and the specificity decreases. Conversely, if the positivity criterion increases, we increase our specificity but decrease our sensitivity.

Comparing two tests

- ROC curves are useful devices for comparing two or more screening tests
- Statistical procedures exist to allow determination of whether two ROC curves differ significantly from each other
- Usual method involves a determination of the area under the curve for each ROC curve and a modification of the Wilcoxon rank-sum procedure to compare them

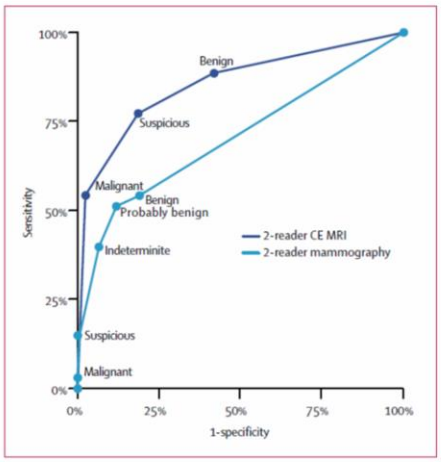
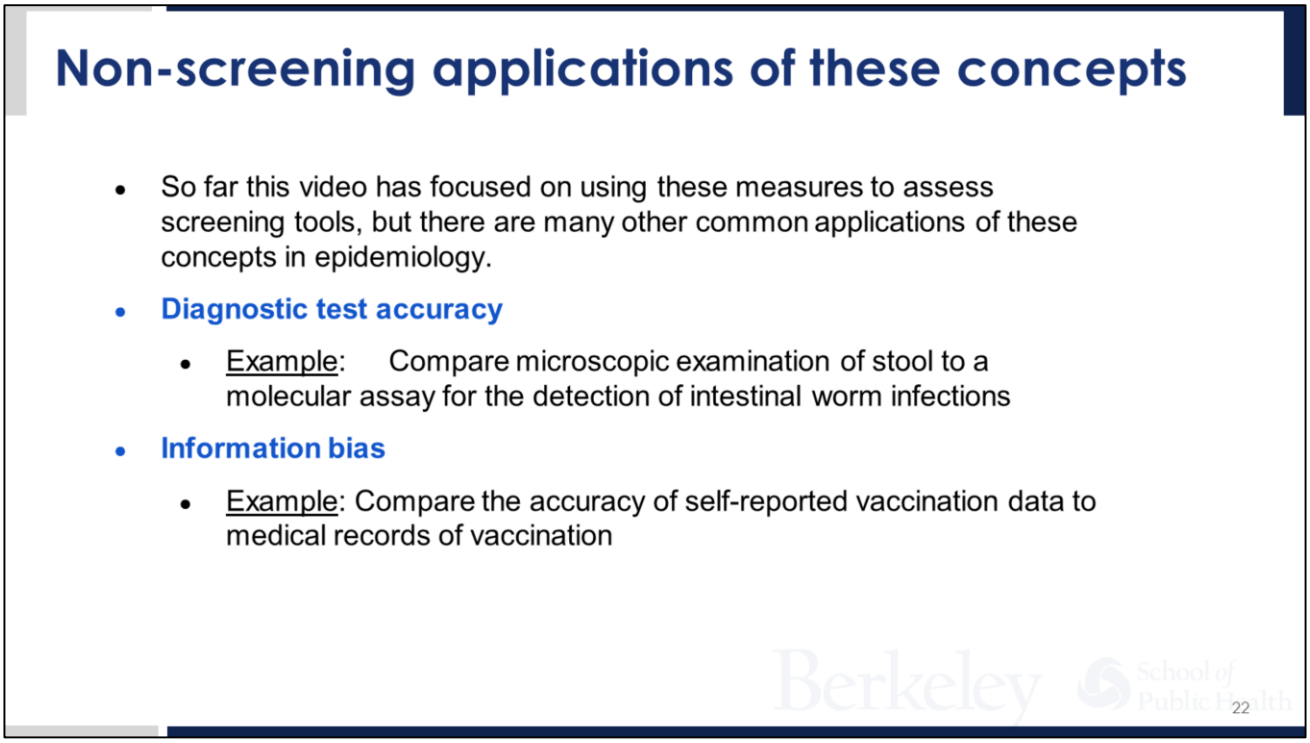


Figure 2: Receiver operator characteristic curves for two-reader CE MRI and mammography

So how would we use ROC curves to compare two tests? Well, they're useful devices for comparing two or more screening tests, and statistical procedures exist to allow determination of whether two ROC curves differ significantly from each other. The usual method involves a determination of the area under the curve for each ROC curve and a modification of a test called the Wilcoxon rank-sum procedure to compare them.

So in this example here, we're seeing a comparison of two tests. We're looking for breast cancer with mammography, one with two-reader mammography and one with a procedure called two-reader CE-MRI, so an MRI scan. And what we see here is the MRI scan appears to be better. See how it's drawn up and to the left, so at a higher sensitivity and a lower specificity in its results. So this allows us to compare, assuming that the area under the curve of the test on the left is greater than the area under the curve of the test on the right and is statistically significant-- it's hard to tell just by looking at it if it's statistically significantly different. But here it certainly appears that one test is better than the other.



So far in this video, we've focused on using these measures to assess screening tools, but there are many other common applications of these concepts in epi. One example is diagnostic test accuracy. For example, we might want to compare the microscopic examination of stool to a molecular assay for the detection of intestinal worm infections, or we might want to evaluate information by comparing the accuracy of self-reported vaccination to medical records of vaccination, treating each of these approaches as a test in a sense.

Summary of key points

- Sensitivity and specificity depend on the cutoff value for disease presence/absence when disease status is classified based on a continuous measure.
- Sensitivity and specificity depend on the distribution of severity of a disease
- PPV and NPV depend on the prevalence of disease in the population.
- Likelihood ratios can be used to connect pre-test and post-test probability of disease.
- ROC curves can be used to compare two tests or assess how sensitivity and specificity change when the positivity criterion changes.

So in summary, the sensitivity and specificity of a test depend on the cutoff value for disease presence or absence when disease status is classified based on some continuous measure-- we set up a positivity criterion. Sensitivity and specificity depend on the distribution of severity of disease. The positive predictive value and the negative predictive value depend on the prevalence of disease in the population. Likelihood ratios can be used to connect pre-test and post-test probability of disease. We saw that with the formula. And ROC curves can be used to compare two tests or to assess how sensitivity and specificity change when the positivity criterion changes.

Epidemiology series

Uses and abuses of screening tests

David A Grimes, Kenneth F Schulz

Screening tests are ubiquitous in contemporary practice, yet the principles of screening are widely misunderstood. Screening is the testing of apparently well people to find those at increased risk of having a disease or disorder. Although an earlier diagnosis generally has intuitive appeal, earlier might not always be better, or worth the cost. Four terms describe the validity of a screening test: sensitivity, specificity, and predictive value of positive and negative results. For tests with continuous variables—eg, blood glucose—sensitivity and specificity are inversely related; where the cutoff for abnormal is placed should indicate the clinical effect of wrong results. The prevalence of disease in a population affects screening test performance: in low-prevalence settings, even very good tests have poor predictive value positives. Hence, knowledge of the approximate prevalence of disease is a prerequisite to interpreting screening test results. Tests are often done in sequence, as is true for syphilis and HIV-1 infection. Lead-time and length biases distort the apparent value of screening programmes; randomised controlled trials are the only way to avoid these biases. Screening can improve health; strong indirect evidence links cervical cytology programmes to declines in cervical cancer mortality. However, inappropriate application or interpretation of screening tests can rob people of their perceived health, initiate harmful diagnostic testing, and squander health-care resources.

Screening is a double-edged sword, sometimes wielded clumsily by the well-intended. Although ubiquitous in contemporary medical practice, screening remains widely misunderstood and misused. Screening is defined as tests done among apparently well people to identify those at an increased risk of a disease or disorder. Those identified are sometimes then offered a subsequent diagnostic test or procedure, or, in some instances, a treatment or preventive medication.[1] Looking for additional illnesses in those with medical problems is termed case finding;[2,3] screening is limited to those apparently well.

Screening can improve health. For example, strong indirect evidence lends support to cytology screening for cervical cancer. Insufficient use of this screening method accounts for a large proportion of invasive cervical cancers in industrialised nations.[4] Other beneficial examples include screening for hypertension in adults; screening for hepatitis B virus antigen, HIV-1, and syphilis in pregnant women; routine urine culture in pregnant women at 12–16 weeks' gestation; and measurement of phenylalanine in newborns.[5] However, inappropriate screening harms healthy individuals and squanders precious resources. The nearly universal antenatal screening for gestational diabetes (a diagnosis in search of a disease)[6] in the USA[7] exemplifies the widespread confusion about the nature and aim of screening. Here, we review the purposes of screening, the selection of tests, measurement of validity, the effect of prevalence on test outcome, and several biases that can distort interpretation of tests.

Ethical implications

What are the potential harms of screening?

Screening differs from the traditional clinical use of tests in several important ways. Ordinarily, patients consult with clinicians about complaints or problems; this prompts testing to confirm or exclude a diagnosis.[8] Because the patient is in pain and requests our help, the risk and expense of tests are usually deemed acceptable by the patient. By contrast, screening engages apparently healthy individuals who are not seeking medical help (and who might prefer to be left alone). Alternatively, consumer-generated demand for screening, such as for osteoporosis and ovarian cancer, might lead to expensive programmes of no clear value.[5,9] Hence, the cost, injury, and stigmatisation related to screening are especially important (though often ignored in our zeal for earlier diagnosis); the medical and ethical standards of screening should be, correspondingly, higher than with diagnostic tests.[10] Bluntly put: every adverse outcome of screening is iatrogenic and entirely preventable.

Screening has a darker side that is often overlooked.[2] It can be inconvenient (the O'Sullivan screen for gestational diabetes), unpleasant (sigmoidoscopy or colonoscopy), and expensive (mammography). For example, a recent Markov model revealed that new screening tests for cervical cancer that are more sensitive than the Papanicolaou test (and thus touted as being better) will drive up the average cost of detecting an individual with cancer.[11] Paradoxically, these higher costs could make screening unattainable by poor women who are at highest risk.[4] The net effect might be more instances of cancer.

A second wave of injury can arise after the initial screening insult: false-positive results and true-positive results leading to dangerous interventions.[2] Although the stigma associated with correct labeling of people as ill might be acceptable, those incorrectly labeled as sick suffer as well. For example, labeling productive steelworkers as being hypertensive led to increased absenteeism and adoption of a sick role, independent of treatment.[12,13] More recently, women labeled as having gestational diabetes reported deterioration in their health and that of their infants over the 5 years after diagnosis.[14] By what right do clinicians rob people of their perceived health, and for what gain?[2]

Screening can also lead to harmful treatment. Treatment of hyperlipidaemia with clofibrate several decades ago provides a sobering example. Treatment of the cholesterol count (a risk factor, rather than an illness itself) inadvertently led to a 17% increase in mortality among middle-aged men given the drug.[2] This screening

Lancet 2002; **359:** 881–84

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

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

misadventure cost the lives of more than 5000 men in the USA alone.[2] Because of these mishaps, reviews of screening practices have recommended that clinicians be more selective.[5,15]

Criteria for screening

If a test is available, should it be used?

The availability of a screening test does not imply that it should be used. Indeed, before screening is done, the strategy must meet several stringent criteria. One checklist separates criteria in three parts: the disease, the policy, and the test.[1] The disease should be medically important and clearly defined, and its prevalence reasonably well known. The natural history should be known, and an effective intervention must exist. Concerning policy, the screening programme must be cost effective, facilities for diagnosis and treatment must be readily available, and the course of action after a positive result must be generally agreed on and acceptable to those screened. Finally, the test must do its job. It should be safe, have a reasonable cut-off level defined, and be both valid and reliable. The latter two terms, often used interchangeably, are distinct. Validity is the ability of a test to measure what it sets out to measure, usually differentiating between those with and without the disease. By contrast, reliability indicates repeatability. For example, a bathroom scale that consistently measures 2 kg heavier than a hospital scale (the gold standard) provides an invalid but highly reliable result.

Although an early diagnosis generally has intuitive appeal, earlier might not always be better. For example, what benefit would accrue (and at what cost) from early diagnosis of Alzheimer's disease, which to date has no effective treatment? Sackett and colleagues[2] have proposed a pragmatic checklist to help decide when (or if) seeking a diagnosis earlier than usual is worth the expense and bother. Does early diagnosis really benefit those screened, for example, in survival or quality of life? Can the clinician manage the additional time required to confirm the diagnosis and deal with those diagnosed before symptoms developed? Will those diagnosed earlier comply with the proposed treatment? Has the effectiveness of the screening strategy been established objectively?[5,15] Finally, are the cost, accuracy, and acceptability of the test clinically acceptable?

Assessment of test effectiveness

Is the test valid?

For over half a century,[16] four indices of test validity have been widely used: sensitivity, specificity, and predictive values of positive and negative. Although clinically useful (and far improved over clinical hunches), these terms are predicated on an assumption that is often clinically unrealistic—ie, that all people can be dichotomised as ill or well. (Indeed, one definition of an epidemiologist is a person who sees the entire world in a 2×2 table.) Often, those tested simply do not fit neatly into these designations: they might be possibly ill, early ill, probably well, or some other variant. Likelihood ratios, which incorporate varying (not just dichotomous) degrees of test results, can be used to refine clinicians' judgments about the probability of disease in a particular person.

For simplicity, however, assume a population has been tested and assigned to the four mutually exclusive cells in figure 1. Sensitivity, sometimes termed the detection rate,[10] is the ability of a test to find those with the disease. All those with disease are in the left column. Hence, the sensitivity is simply those correctly identified by the test (a) divided by all those sick (a+c). Specificity denotes the

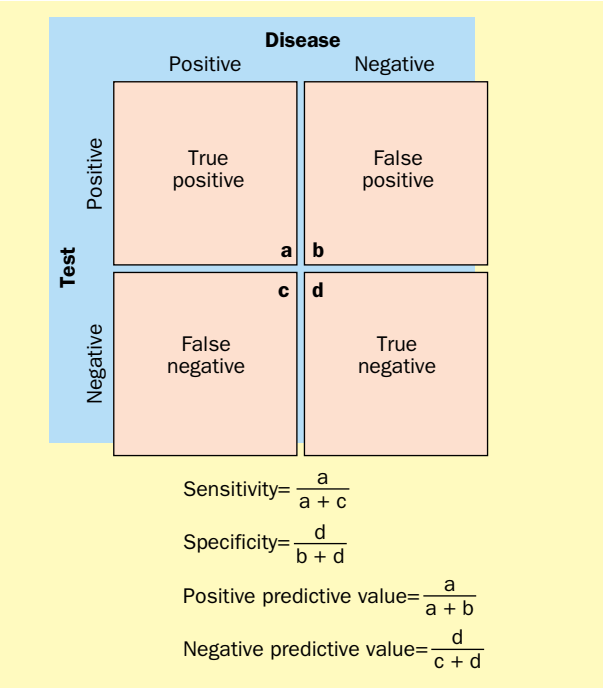


Figure 1: **Template for calculation of test validity**

ability of a test to identify those without the condition. Calculation of this proportion is trickier, however. By analogy to sensitivity, many assume (incorrectly) that the formula here is b/(b+d). However, the numerator for specificity is cell d (the true negatives), which is divided by all those healthy (b+d).

Although sensitivity and specificity are of interest to public-health policymakers, they are of little use to the clinician. Stated alternatively, sensitivity and specificity (population measures) look backward (at results gathered over time).[8] Clinicians have to interpret test results to those tested. Thus, what clinicians need to know are the predictive values of the test (individual measures, which look forward). To consider predictive values, one needs to shift the orientation in figure 1 by 90 degrees: predictive values work horizontally (rows), not vertically (columns). In the top row are all those with a positive test, but only those in cell a are sick. Thus, the predictive value positive is a/(a+b). The "odds of being affected given a positive result (OAPR)" is the ratio of true positives to false positives, or a to b.[10] For example, in figure 1, the OAPR is 75/5, or 17/1. This corresponds to a positive predictive value of 89%. Advocates of use of the OAPR note that these odds better describe test effectiveness than do probabilities (predictive values). In the bottom row of figure 1 are those with negative tests, but only those in cell d are free of disease. Hence, the predictive value negative is d/(c+d).

Learning (and promptly forgetting) these formulas was an annual ritual for many of us in our clinical training. If readers understand the definitions above and can recall the 2×2 table shell, then they can quickly figure out these formulas when needed. As a mnemonic, disease goes at the top of the table shell, since it is our top priority. By default, test goes on the left border.

Through the years, researchers have tried to simplify these four indices of test validity by condensing them into a single term.[8] However, none adequately depicts the important trade-offs between sensitivity and specificity that generally arise. An example is diagnostic accuracy, which is the proportion of correct results.[3] It is the sum of

the correctly identified ill and well divided by all those tested, or (a+d)/(a+b+c+d). Cells b and c are noise in the system. Another early attempt, Youden's J, is simply the predictive value positive plus the predictive value negative minus one.[17] The range of values extends from zero (for a coin toss with no predictive value) to 1·0, where predictive values of both positive and negative tests are perfect.

Trade-offs between sensitivity and specificity

Where should the cut-off for abnormal be?

The ideal test would perfectly discriminate between those with and without the disorder. The distributions of test results for the two groups would not overlap. More commonly in human biology, test values for those with and without a disease overlap, sometimes widely.[18] Where one puts the cut-off defining normal versus abnormal determines the sensitivity and specificity. For any continuous outcome measurement—for example, blood pressure, intraocular pressure, or blood glucose—the sensitivity and specificity of a test will be inversely related. Figure 2 shows that placing the cut-off for abnormal blood glucose at point X produces perfect sensitivity; this low cut-off identifies all those with diabetes. However, the trade-off is poor specificity: those in the part of the healthy distribution in pink and purple are incorrectly identified as having abnormal values. Placing the cut-off higher at point Z yields the opposite result: all those healthy are correctly identified (perfect specificity), but the cost here is missing a proportion of ill individuals (portion of the diabetic distribution in purple and blue). Placing the cut-off at point Y is a compromise, mislabeling some healthy people and some people with diabetes.

Where the cut-off should be depends on the implications of the test, and receiver-operator characteristic curves are useful in making this decision.[19] For example, screening for phenylketonuria in newborns places a premium on sensitivity rather than on specificity; the cost of missing a case is high, and effective treatment exists. The downside is a large number of false-positive tests, which cause anguish and further testing. By contrast, screening for breast cancer should favour specificity over sensitivity, since further assessment of those tested positive entails costly and invasive biopsies.[20]

Prevalence and predictive values

Can test results be trusted?

A badly understood feature of screening is the potent effect of disease prevalence on predictive values.

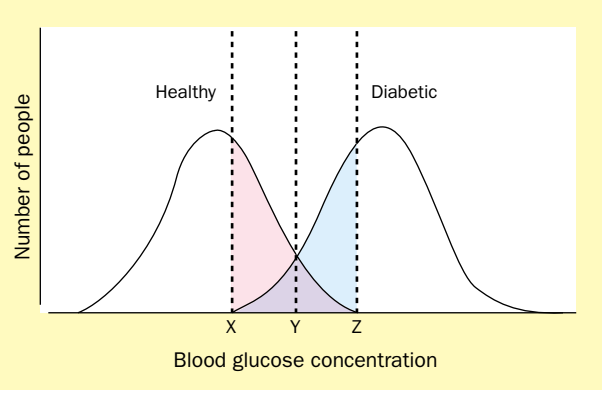


Figure 2: **Hypothetical distribution of blood glucose concentrations in people with and without diabetes**
Setting cut-off for abnormal at X yields perfect sensitivity at the expense of specificity. Setting cut-off at Z results in perfect specificity at the cost of lower sensitivity. Cut-off Y is a compromise.

Clinicians must know the approximate prevalence of the condition of interest in the population being tested; if not, reasonable interpretation is impossible. Consider a new PCR test for chlamydia, with a sensitivity of 0·98 and specificity of 0·97 (a superb test). As shown in the left panel of figure 3, a doctor uses the test in a municipal sexually transmitted disease clinic, where the prevalence of *Chlamydia trachomatis* is 30%. In this high-prevalence setting, the predictive value of a positive test is high, 93%—ie, 93% of those with a positive test actually have the infection.

Impressed with the new test, the doctor now takes it to her private practice in the suburbs, which has a clientele that is mostly older than age 35 years (figure 3, right panel). Here, the prevalence of chlamydial infection is only 3%. Now the same excellent test has a predictive positive value of only 0·50. When the results of the test are positive, what should the doctor tell the patient, and what, in turn, should the patient tell her husband? Here, flipping a coin has the same predictive positive value (and is considerably cheaper and simpler than searching for bits of DNA). This message is important, yet not widely understood: when used in low-prevalence settings, even excellent tests have poor predictive positive value. The reverse is true for negative predictive values, which are nearly perfect in figure 3. Although failing to diagnose sexually transmitted diseases can have important health implications, incorrectly labeling people as infected can wreck marriages and damage lives.

Tests in combination

Should a follow-up test be done?

Clinicians rarely use tests in isolation. Few tests have high sensitivity and specificity, so a common approach is to do tests in sequence. In the instance of syphilis, a sensitive (but not specific) reagin test is the initial screen. Those who test positive then get a second, more specific test, a diagnostic treponemal test. Only those who test positive on both receive the diagnosis. This strategy generally increases the specificity compared with a single test and limits the use of the more expensive treponemal test.[20] Testing for HIV-1 is an analogous two-step procedure.

Alternatively, tests can be done in tandem (parallel or simultaneous testing).[3,21] For example, two different tests might both have poor sensitivity, but one might be better at picking up early disease, whereas the other is better at

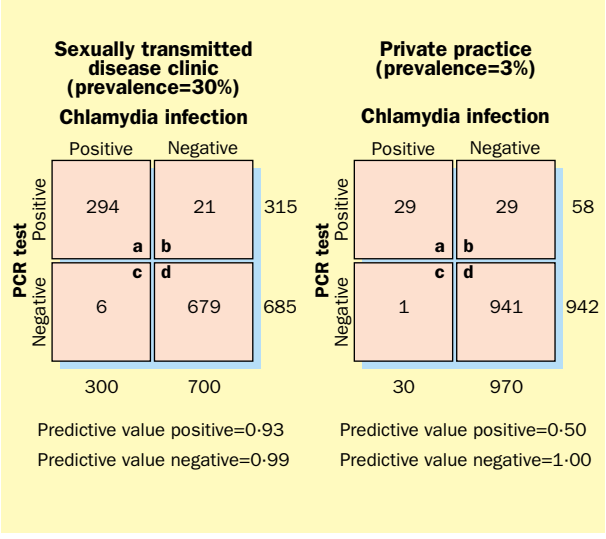


Figure 3: **Predictive values of a PCR test for *Chlamydia trachomatis* in high-prevalence and low-prevalence settings**

identifying late disease. A positive result from either test would then lead to diagnostic assessment. This approach results in higher sensitivity than would arise with either test used alone.

Benefit or bias?

Does a screening programme really improve health?

Even worthless screening tests seem to have benefit.[2] This cruel irony underlies many inappropriate screening programmes used today. Two common pitfalls lead to the conclusion that screening improves health; one is an artifact and the other a reflection of biology.

Lead-time bias

Lead-time bias refers to a spurious increase in longevity associated with screening. For example, assume that mammography screening leads to cancer detection 2 years earlier than would have ordinarily occurred, yet the screening does not prolong life. On average, women with breast cancer detected through screening live 2 years longer than those with cancers diagnosed through traditional means. This gain in longevity is apparent and not real: this hypothetical screening allows women to live 2 years longer with the knowledge that they have cancer, but does not prolong survival, an example of zero-time shift.[2]

Length bias

Length bias is more subtle than lead-time bias: the longevity association is real, but indirect. Assume that community-based mammography screening is done at 10-year intervals. Women whose breast cancers were detected through screening live 5 years longer on average from cancer initiation to death than those whose cancers were detected through usual means. That screening is associated with longer survival implies clear benefit. However, in this hypothetical example, this benefit indicates the inherent variability in cancer growth rates and not a benefit of screening. Women with indolent, slow-growing cancers are more likely to live long enough to be identified in decennial screening. Conversely, those with rapidly progressing tumours are less likely to survive until screening.

The only way to avoid these pervasive biases is to do randomised controlled trials and then to assess age-specific mortality rates for those screened versus those not screened.[10] Moreover, the trials must be done well. The quality of published trials of mammography screening has raised serious questions about the utility of this massive and hugely expensive enterprise.[22–24]

Conclusion

Screening can promote or impair health, depending on its application. Unlike a diagnostic test, a screening test is done in apparently healthy people, which raises unique ethical concerns. Sensitivity and specificity tend to be inversely related, and choice of the cut-off point for abnormal should indicate the implications of incorrect results. Even very good tests have poor predictive value positive when applied to low-prevalence populations. Lead-time and length bias exaggerate the apparent benefit of screening programmes, underscoring the need for rigorous assessment in randomised controlled trials before use of screening programmes.

Acknowledgments

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

References

1 Cuckle HS, Wald NJ. Principles of screening. In: Antenatal and neonatal screening. Oxford: Oxford University Press, **1984:** 1–22.
2 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine, 2nd edn. Boston: Little, Brown and Company, 1991.
3 Lang TA, Secic M. How to report statistics in medicine. Philadelphia: American College of Physicians, 1997.
4 Sawaya GF, Grimes DA. New technologies in cervical cytology screening: a word of caution. *Obstet Gynecol* 1999; **94:** 307–10.
5 US Preventive Services Task Force. Guide to clinical preventive services, 2nd edn. Baltimore: Williams and Wilkins, 1996.
6 Enkin M, Keirse MJNC, Neilson J, et al (eds). A guide to effective care in pregnancy and childbirth, 3rd edn. Oxford: Oxford University Press, 2000.
7 Gabbe S, Hill L, Schmidt L, Schulkin J. Management of diabetes by obstetrician–gynecologists. *Obstet Gynecol* 1998; **91:** 643–47.
8 Feinstein AR. Clinical biostatistics XXXI: on the sensitivity, specificity, and discrimination of diagnostic tests. *Clin Pharmacol Ther* 1975; **17:** 104–16.
9 NIH Consensus Development Panel on Ovarian Cancer. Ovarian cancer: screening, treatment, and follow–up. *JAMA* 1995; **273:** 491–97.
10 Wald N, Cuckle H. Reporting the assessment of screening and diagnostic tests. *Br J Obstet Gynaecol* 1989; **96:** 389–96.
11 Myers ER, McCrory DC, Subramanian S, et al. Setting the target for a better cervical screening test: characteristics of a cost-effective test for cervical neoplasia screening. *Obstet Gynecol* 2000; **96:** 645–52.
12 Haynes RB, Sackett DL, Taylor DW, Gibson ES, Johnson AL. Increased absenteeism from work after detection and labeling of hypertensive patients. *N Engl J Med* 1978; **299:** 741–44.
13 Taylor DW, Haynes RB, Sackett DL, Gibson ES. Longterm follow-up of absenteeism among working men following the detection and treatment of their hypertension. *Clin Invest Med* 1981; **4:** 173–77.
14 Feig DS, Chen E, Naylor CD. Self-perceived health status of women three to five years after the diagnosis of gestational diabetes: a survey of cases and matched controls. *Am J Obstet Gynecol* 1998; **178:** 386–93.
15 Canadian Task Force on the Periodic Health Examination. The Canadian guide to clinical preventive care. Ottowa: Minister of Supply and Services Canada, 1994.
16 Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Pub Health Rep* 1947; **62:** 1432–49.
17 Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3:** 32–35.
18 Griffith CS, Grimes DA. The validity of the postcoital test. *Am J Obstet Gynecol* 1990; **162:** 615–20.
19 Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987; **6:** 411–23.
20 Hennekens CH, Buring JE. Epidemiology in medicine. Boston: Little, Brown and Company, 1987.
21 Riegelman RK, Hirsch RP. Studying a study and testing a test, 2nd edn. Boston: Little, Brown and Company, 1989.
22 Gotzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? *Lancet* 2000; **355:** 129–34.
23 Olsen O, Gotzsche PC. Cochrane review on screening for breast cancer with mammography. *Lancet* 2001; **358:** 1340–42.
24 Horton R. Screening mammography: an overview revisited. *Lancet* 2001; **358:** 1284–85.

Epidemiology Case Studies – Risk Factors for Menstrual Toxic Shock Syndrome - Multistate Case-Control Study Part 2

MICHELLE RUIZ: This is Michelle Ruiz, we are now on our seventh episode of Epidemiology case studies. So, now we'll hear the second part of Dr. Jack Colford and Dr. Art Reignold. In this part, they continue to discuss the case-control study performed to asess risk factors for developing toxic shock syndrome (TSS) during menstruation, their conversation now focuses on addressing bias.

JACK COLFORD: So welcome again. We're talking with Professor Art Reingold about his study of the association between tampon use and toxic shock syndrome in the *JAMA* article that you've all read. And I'd like to begin, Art, today talking about some of the potential biases in a study like this and how the case-control study you used might address things like, for instance, selection bias, information bias.

ART REINGOLD: So I would say of the three main areas where there'd be concern about inference-- confounding information bias and selection bias-- we were fairly comfortable that uncontrolled confounding was not a major problem. We certainly collect either matched on, as we already indicated, age and geographic area of residence. We collected information about other potential confounders. There weren't big differences, but they were controlled for in the analysis.

So I think confounding was not so much an issue, but there are certainly concerns in this study or in similar prior studies about information bias and selection bias. And so one question at the heart of any of these studies of menstrual toxic shock is whether the cases are representative of all the cases occurring during menstruation, or whether there might be a greater diagnosis in reporting of cases in tampon users than in non-tampon users.

And once the association had been reported and was known to women, was known to health care providers, it's certainly possible that cases that were diagnosed and reported and included in the study were more likely to be tampon users than non-tampon users. To the extent that that was true in our study, that would have biased the results toward an association with tampon use in general. And there's not much we can do to evaluate that, although we did, in fact, using a lot of resources, look through medical records of people with overlapping clinical diagnoses to see if any of them were undiagnosed cases of toxic shock syndrome.

JACK COLFORD: OK.

ART REINGOLD: A huge amount of work that didn't produce very much. But I would point out that within this study, the primary question wasn't whether tampon use increased the risk, but whether use of one particular brand--

JACK COLFORD: Or type.

ART REINGOLD: --or style or absorbency increased the risk, compared to another brand or style or absorbency. So for selection bias to have influenced that, we would have needed patients or their providers to be more suspicious of and report a case in users of one brand--

JACK COLFORD: Particular type, yeah.

ART REINGOLD: --versus another.

JACK COLFORD: Sure, sure.

ART REINGOLD: Higher absorbency versus lower absorbency. And I, personally, don't think that was very likely. But I suppose one could argue about that. So certainly, questions about whether there could have been some selection bias. So the odds ratio for tampon use versus no tampon use could certainly, possibly have been influenced by that.

In the era of information bias, in all of these case-control studies of menstrual toxic shock syndrome, there have been concerns about whether women accurately report on their tampon use or their other exposures. And we actually had the women look for their package of tampons that had been used. About 60% of them had the package.

JACK COLFORD: Yup.

ART REINGOLD: They could read the label. They could describe the package and tell us the color.

JACK COLFORD: Sure.

ART REINGOLD: We were pretty sure that they were giving us accurate information. I would also assert that the average woman can reliably report whether she used tampons or not during her last month period. You know, it's not a subtle thing to report on.

So when you start asking questions about sexual activity and other things where there might be some sensitivity, I suppose you could question whether women would report honestly about something like that. But we have reason to think that women were being pretty candid about the things we were asking them about.

JACK COLFORD: Great. So you estimated odds ratios in this study. Can you tell us kind of the statistical methods you used to estimate those?

ART REINGOLD: Well, it was pretty straightforward, case-control studies 101 in terms of looking at basically what proportion of the cases were exposed, what portion of the controls were exposed, and initially using univariate analyses and then, ultimately, logistic regression, in this case because they were matched to conditional logistic regression to control for possible confounders and generate an odds ratio that took into account the matching.

JACK COLFORD: The matching, great.

ART REINGOLD: So there was nothing very fancy about the analysis.

JACK COLFORD: Conditional on the two factors you matched on.

ART REINGOLD: Exactly.

JACK COLFORD: So perfect. OK. So what associations did you find between the various tampon use and types of tampons and toxic shock syndrome?

ART REINGOLD: So beyond the association of tampon use--

JACK COLFORD: In general.

ART REINGOLD: --versus no tampon use, which, as I've said, there certainly could be some selection bias in that, fundamentally we confirmed our hypothesis that some brands and styles were higher risk compared to other brands and styles. And so they are one particular brand, Tampax Regular, which was thought to be, if anything, the lowest risk product. When that was the standard of

comparison, higher absorbency tampons were associated with a higher odds of disease compared to Tampax Regular use.

And we also actually produced an odds ratio for increasing absorbency because tampons can be measured in vitro or in vivo for how many grams of liquid they absorb and produce an odds ratio per gram of absorbency. And showed that for every increase in gram of absorbency of a tampon product, there was a 34% increase in the odds of toxic shock syndrome.

JACK COLFORD: My recollection of that figure in the article was it was very linear, that relation.

ART REINGOLD: Yeah, that was pretty much a straight, linear relationship. So again, I have a hard time believing that was due to bias in terms of what people were reporting, and certainly suggest that absorbency was the critical factor in determining a risk of toxic shock syndrome.

Now, I would say that two things remain unresolved by that-- one is the role of chemical composition. Because not all tampons are made of the same chemicals-- cotton versus rayon being a major distinction. We didn't really end up having a large enough number of cases to look at that factor. The good news is, by the time we were doing that study, the rate of this disease declined and we didn't have enough cases.

But the other real question is, what is it in a tampon, or what is it about absorbency that changes your risk of this disease? And there are lots of theories. So one is, how much oxygen is introduced when you introduce a tampon. Because oxygen is being introduced into a normally oxygen-less or anaerobic environment.

So maybe the amount of oxygen introduced is important. Maybe these products bind certain cations, like calcium and magnesium that, in vitro, is important in terms of toxin production. So our studies really couldn't begin to tease apart those questions about what is it about absorbency and is there an added effect of chemical composition.

JACK COLFORD: How generalizable do you think your study in a particular population that you used is to other populations, either in the US or in other countries or?

ART REINGOLD: Well, I don't have any reason to think it isn't generalizable. Now, the risk of this disease has always been substantially higher in younger women than in older women. And that's consistently been found to be the case.

Is it possible that these results-- that there's an effect modification by age and the tampon use

is not as much of a risk factor when you're an older woman than a younger woman? That's possible. One reason might be that older women are more likely to have antibodies to the toxin and basically be immune to the disease--

JACK COLFORD: Sure.

ART REINGOLD: --as opposed to younger women, who are not. So it's possible that the results don't extrapolate well to all age groups.

JACK COLFORD: Mm-hmm.

ART REINGOLD: In terms of racial groups or geographic groups, I honestly don't see any reason to think they wouldn't be generalizable.

JACK COLFORD: Be any different. Sure, sure. How would you say your study-- this particular case-control study-- improved upon or built upon prior studies of the toxic shock syndrome in risk factors?

ART REINGOLD: Well, I think, first of all, it quantified the relationship with absorbency in a way that earlier studies had not.

JACK COLFORD: Hadn't done. Yup.

ART REINGOLD: Secondly, as I said, because in response to the earlier studies, manufacturers had substantially reduced the absorbency of products they were selling and changed the chemical composition. The real question was whether the risk continued, given that the products had changed. So I think it contributed to our understanding that, yes, the risk was still there, in addition to giving us a pretty good quantitative estimation of what that risk was.

I think it also provided some reassurance that some of the other variables we looked at, like contraception and the like, were not important. Or over-the-counter medication use didn't contribute to the risk of this disease.

JACK COLFORD: Do you have any specific stories or tales of manufacturers' responses to your study or?

ART REINGOLD: Well, to our particular study, I think the industry was basically fairly quiescent. They didn't raise a fuss. They didn't start pointing to a lot of flaws. I don't think they made a lot in the way of further changes to their products. I think they pretty much accepted it at face value and went on with their jobs.

JACK COLFORD: What happened to the national incidence of the syndrome as time went on?

ART REINGOLD: Well, there's a lot of controversy about that, even to this day. And several studies, including one that I did with colleagues at Kaiser Permanente, would suggest that the rate of menstrual toxic shock syndrome in some areas did go down during this time period. In other areas, it didn't go down very much. And so while the number of reported cases seem to have gone down after the early 1980s, these other studies that pretty much remove either diagnostic bias or reporting bias completely sort of give a mixed picture about that.

So that's somewhat unclear, how much the rate declined at that point. I would say at this point in time, in 2018, it remains an exceedingly rare disease. So the estimates are at this point well under 1 in 100,000 menstruating women per year, and perhaps more like 1 in 500,000.

JACK COLFORD: OK.

ART REINGOLD: And so is it a common problem? No. And at this point, it's very hard to study because it's so rare.

JACK COLFORD: So rare. Yeah. Yeah.

ART REINGOLD: So there are not ongoing studies that I know of.

JACK COLFORD: So if you could go back and do anything different about the design and conduct of this particular study, what would you have done?

ART REINGOLD: Well, that's a good question. You know, I think we certainly did our best around control selection and trying to make sure the cases were representative of all the cases that were out there. I can't think of any hypotheses that I would have tried to study that would seem plausible that we would have looked at.

So I guess, maybe if we'd carried it on longer and enrolled enough more cases, we'd have had a better understanding about the role of chemical composition. But the decision was, we would only enroll patients for 18 months. And when that time was over, we stopped.

JACK COLFORD: You were done. Great.

ART REINGOLD: [CHUCKLES]

JACK COLFORD: Well, thanks a lot, Art, for talking to us about your case-control study of tampons and toxic

shock syndrome.

ART REINGOLD: Great, happy to do it.

JACK COLFORD: Thank you.

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

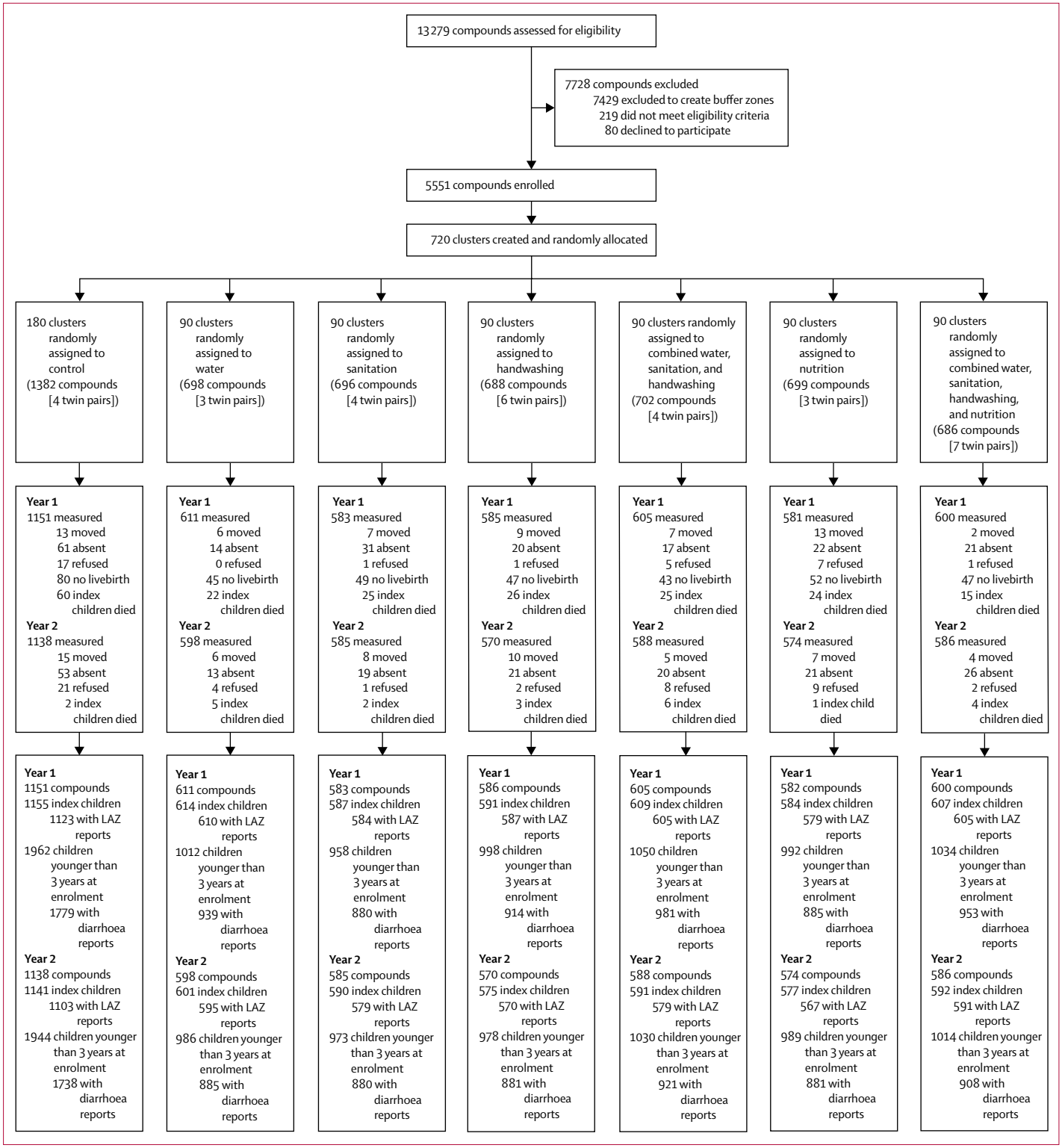


Figure 1: **Trial profile and analysis populations for primary outcomes**
LAZ=length-for-age *Z* scores.

	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.

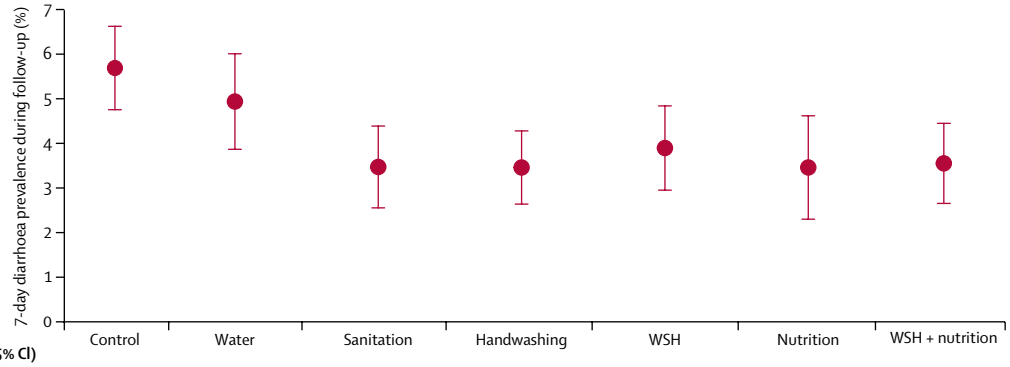
The trial is registered at ClinicalTrials.gov, number NCT01590095. The International Centre for Diarrhoeal Disease Research, Bangladesh convened a data and safety monitoring board and oversaw the study.

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



Prevalence ratio (95% CI)	Control	Water	Sanitation	Handwashing	WSH	Nutrition	WSH + nutrition
Intervention *vs* control	ref	0·89 (0·70–1·13)	0·61 (0·46–0·81)	0·60 (0·45–0·80)	0·69 (0·53–0·90)	0·64 (0·49–0·85)	0·62 (0·47–0·81)
WSH *vs* water		ref			0·77 (0·57–1·04)		
WSH *vs* sanitation			ref		1·13 (0·80–1·58)		
WSH *vs* handwashing				ref	1·07 (0·77–1·51)		

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

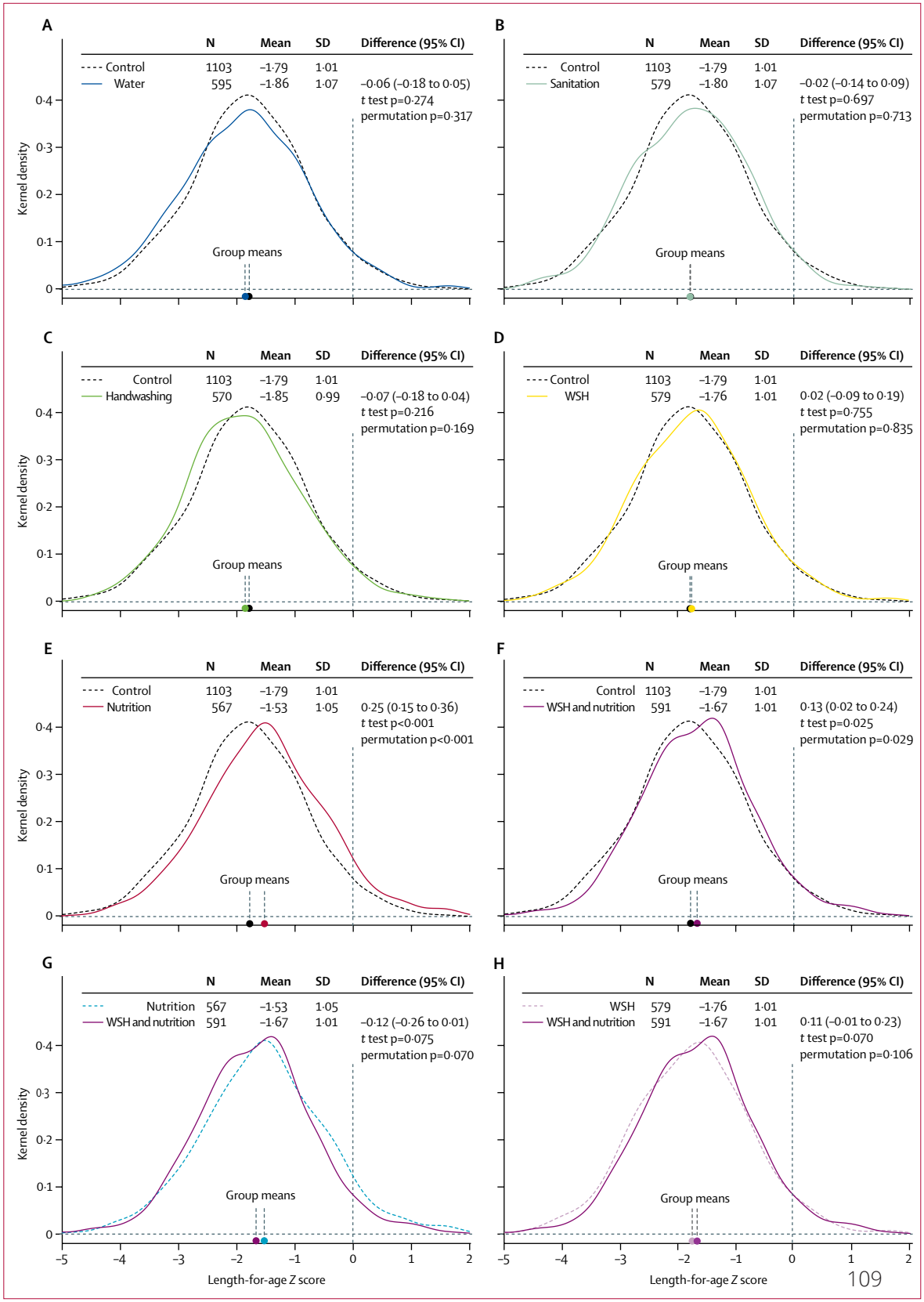


Figure 3: **Intervention effects on length-for-age Z scores in 4584 children after 2 years of intervention**

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

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

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

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

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

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

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

Contributors

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

Declaration of interests

We declare no competing interests.

Acknowledgments

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Methods

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

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

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

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

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

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

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

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

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

Results

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

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

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

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

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

* Values given are mean ± SD (range).

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

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

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

Table 2. Tampon use during the index menstrual period.

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

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

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

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

* Vs. no tampon use.

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

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

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

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

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

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

* Single brand and style use only.

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

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

* Deodorant and nondeodorant, combined.

Discussion

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

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

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

* Values indicate number (percentage) of women.

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

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

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

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

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

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

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

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

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

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

References

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

Discussion

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

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

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

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

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

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

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

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