PHW250B Week 6 Reader

Topic 1: Types of Randomized Trials, Equivalence Trials, Cluster-Randomized Trials

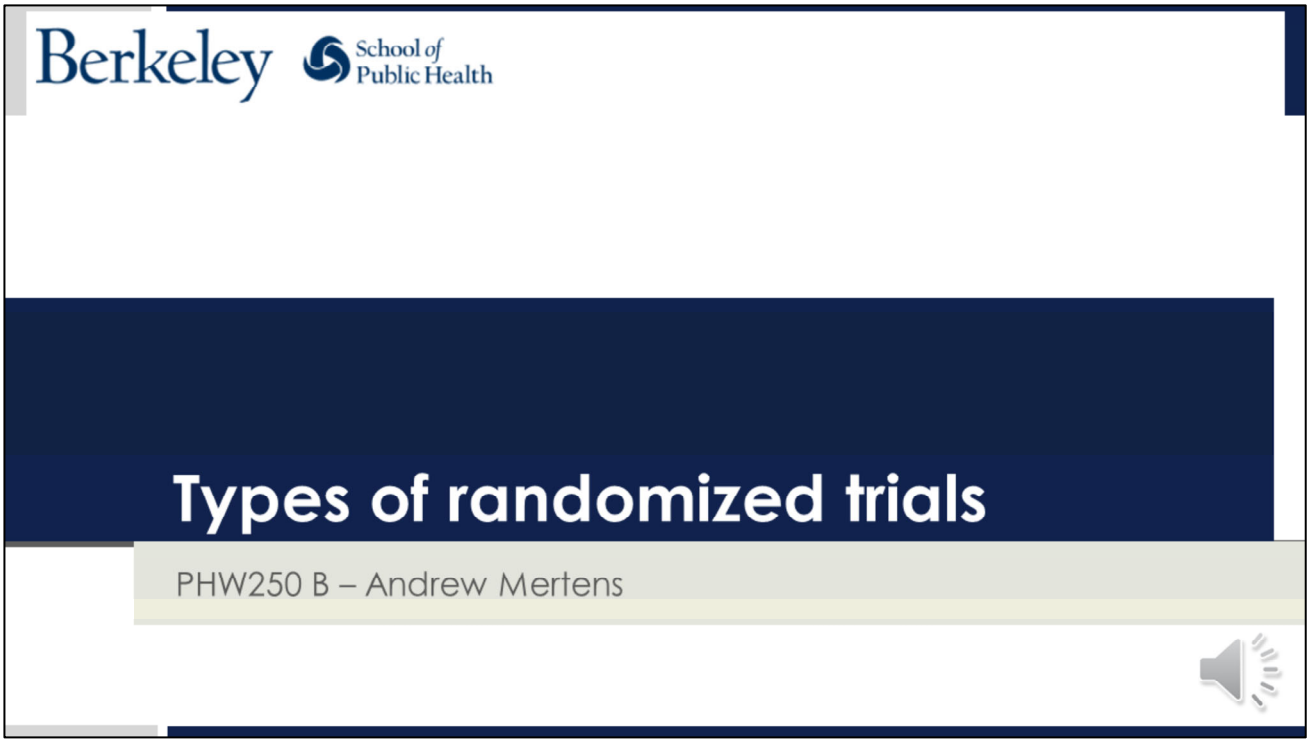
Topic 2: Evaluating and Reporting Randomized Trials

Topic 3: Case Studies

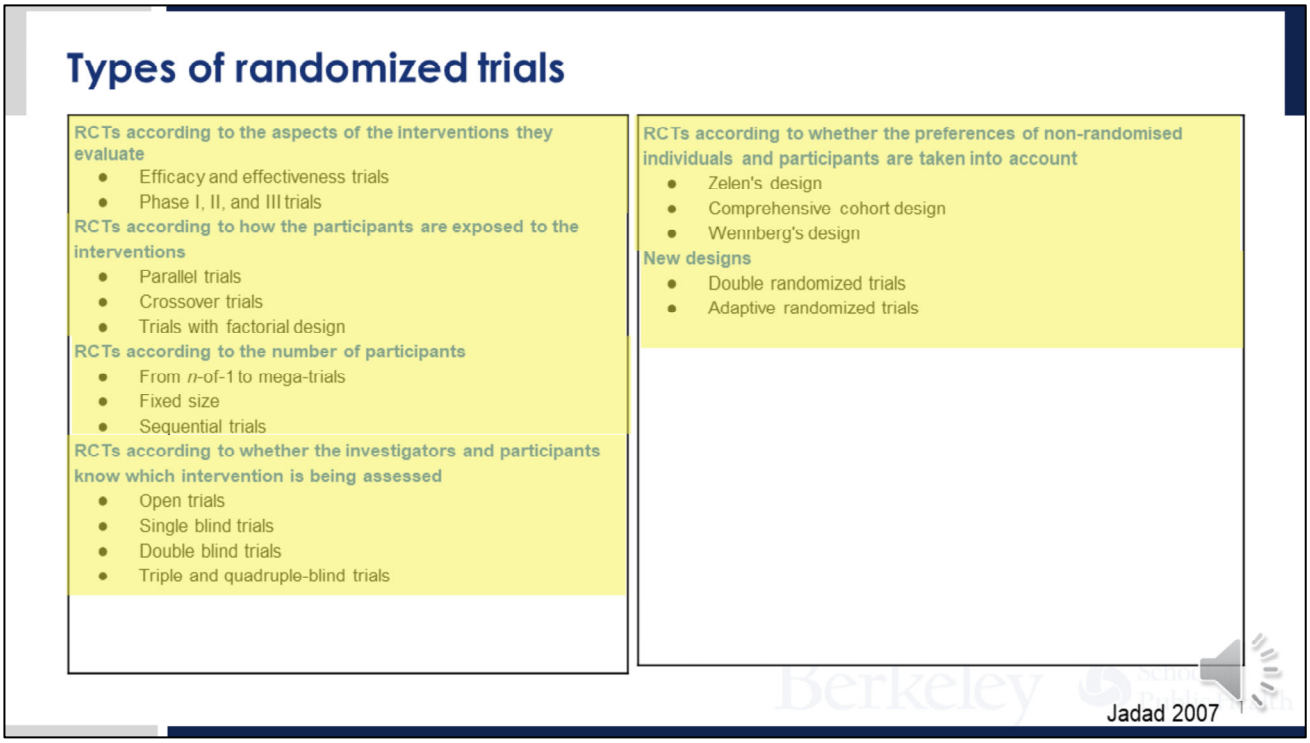
Podcast

Journal Club

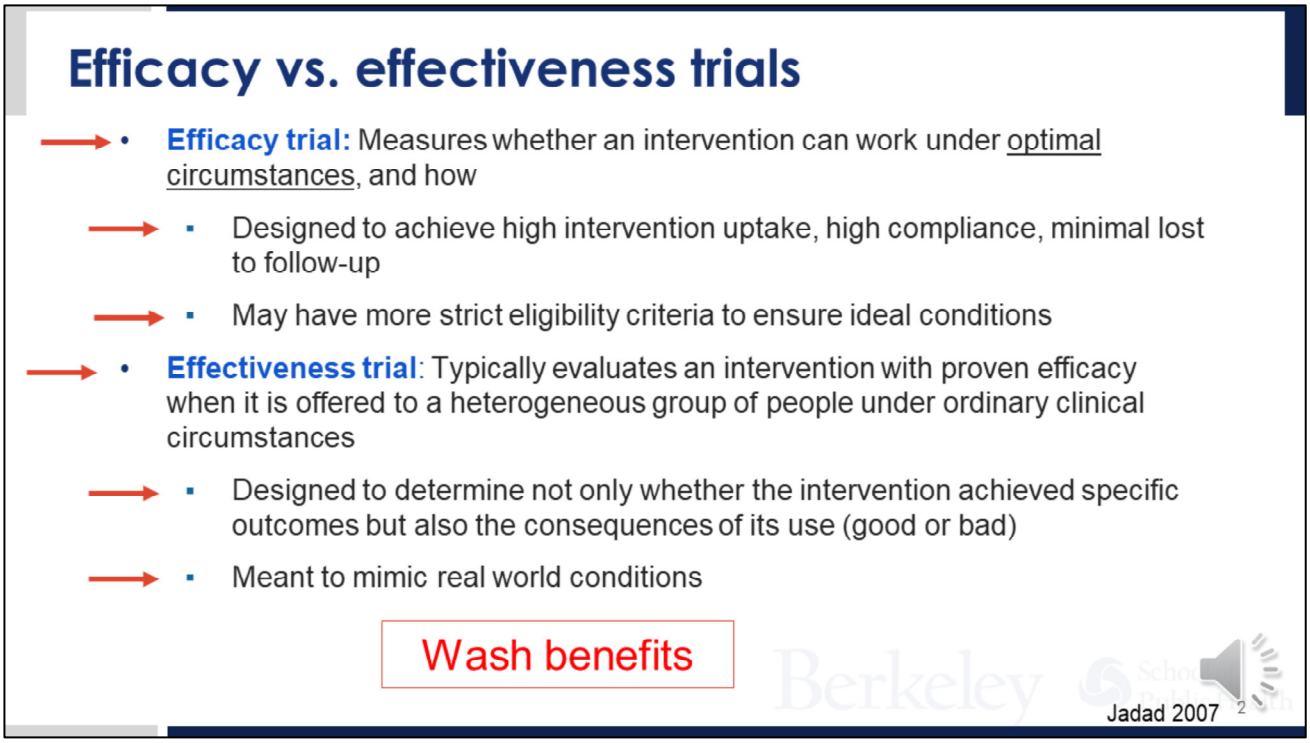
Lecture: Types of Randomized Trials



In this video, we'll go over a wide range of different types of randomized trials, and then in subsequent videos this week, we'll go into more depth around a key set of randomized trials that are quite common in epidemiology.



This slide shows us how we can categorize the wide range of randomized trials that are in use today. *Some trial designs depend on aspects of interventions that they evaluate, *some RCTs focus on how participants are exposed to intervention, *some are designed based on the number of participants. *There's a set of different kinds of RCTs that are designed based on whether investigators and participants know which intervention is being assessed. *There are RCTs that are designed around preference for non-randomized and randomized treatment, *and then there's a couple of new designs that we'll cover the end of this video. And these different types are not necessarily mutually exclusive, so you could have, for example, a Phase III, sequential, double-blind, adaptive randomized trial.



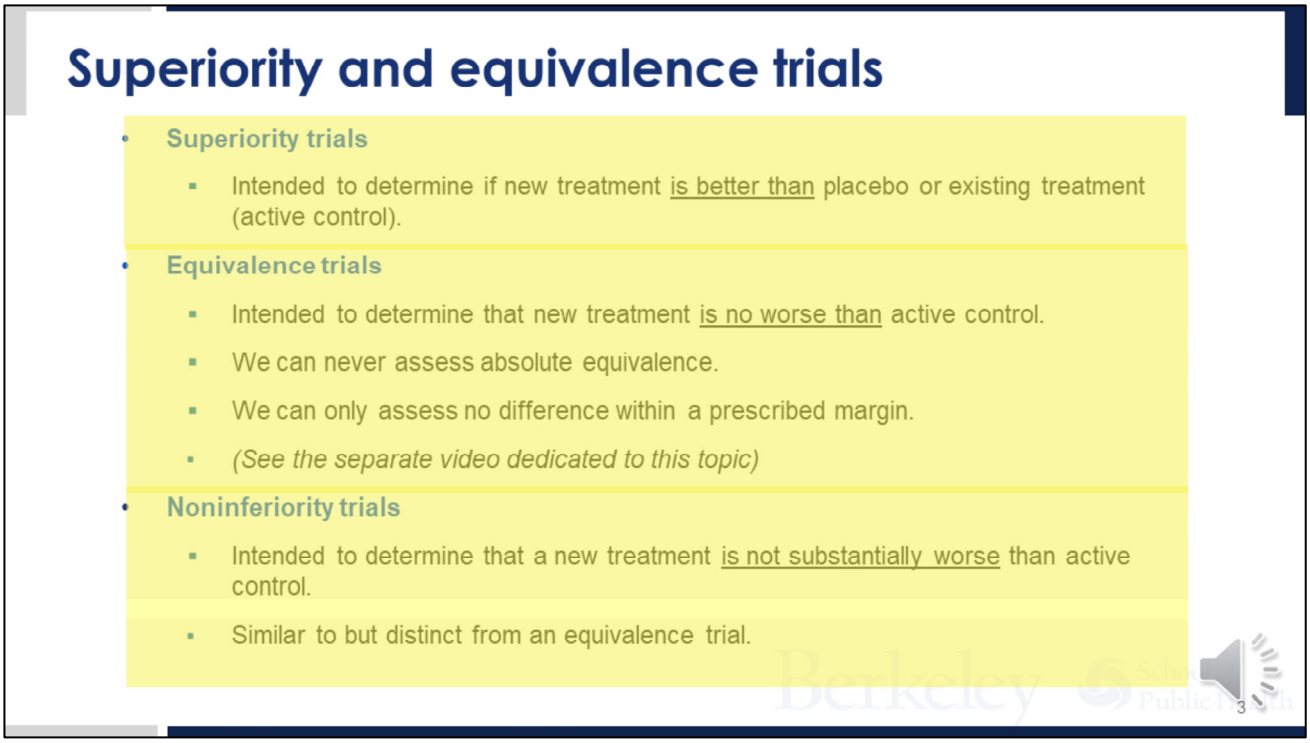
Let's start with efficacy vs effectiveness trials. *An efficacy trial measures whether an intervention can work under optimal circumstance and how. *It's designed to achieve high intervention uptake, high compliance, minimal lost to follow-up. *In order to achieve a really high intervention uptake and compliance, it's usually necessary to restrict eligibility to people who are able to participate in the ideal circumstances for the trial.

*An effectiveness trial, on the other hand, typically evaluates an intervention that already has been proven to be efficacious in an efficacy trial. And then in the effectiveness trial, the intervention is offered to a more heterogeneous group of people under ordinary clinical or other real-world circumstances.

*Effectiveness trials are designed to determine not only whether the intervention achieves specific outcomes, but also the consequences of its use, whether they're positive or negative, good or bad, and *they're meant to mimic real world conditions.

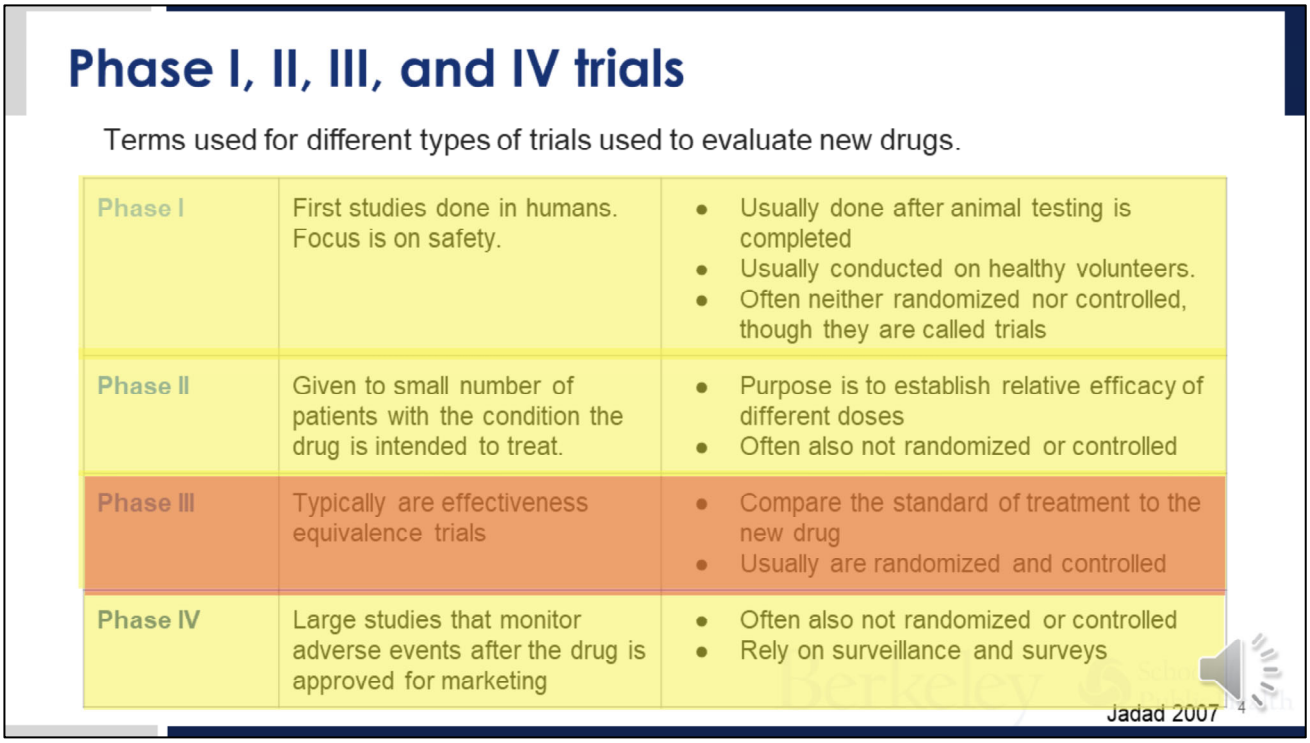
You're going to watch a video later this week about **an efficacy trial called wash benefits that Jack and I worked on in Bangladesh and Kenya. These trials were designed to implement an improved water sanitation hand-washing and nutrition intervention in ideal circumstances in order to achieve high uptake because no previous trial had ever asked this research question.

After an efficacy trial such as wash benefits is completed, it's very common for other investigators, or perhaps the same investigation team, to then conduct an effectiveness trial in a much broader, potentially nationally-representative population, in order to see if a wider range of people, who are perhaps going to be less compliant , or may have a harder time achieving high intervention uptake, can still benefit from the intervention.



A set of designs that's common in the field of drug trials is superiority and equivalence designs. *A superiority trial is intended to determine if a new treatment is better than an existing treatment or a placebo, which we'll refer to as an active control. *On the other hand, an equivalence trial intends to determine that the new treatment is not worse than the active control. We can't ever assess absolute equivalence, but we can assess that there is no difference between the new treatment and the active control within a predetermined margin. And we have a whole separate video dedicated to this topic.

*Another term you'll see is non-inferiority trials, and these are intended to determine that a new treatment is not substantially worse than the active control. It's similar to but distinct from an equivalence trial. These kinds of trials are really common for new drugs and for treatments in clinical settings, where there's a preexisting drug or a standard of care that is being compared to a new alternative.



Another way we can classify trials that evaluate new potential drugs and therapies are by their phase. So phase one through four. *The first phase is typically done in humans, and the focus is on safety. And I want to make an important distinction here. When we use the term phase one through four trial, technically many of these are not randomized trials in the way that we'll use the term in the remainder of this course.

This is one of these tricky semantics things that we just have to be really careful about.

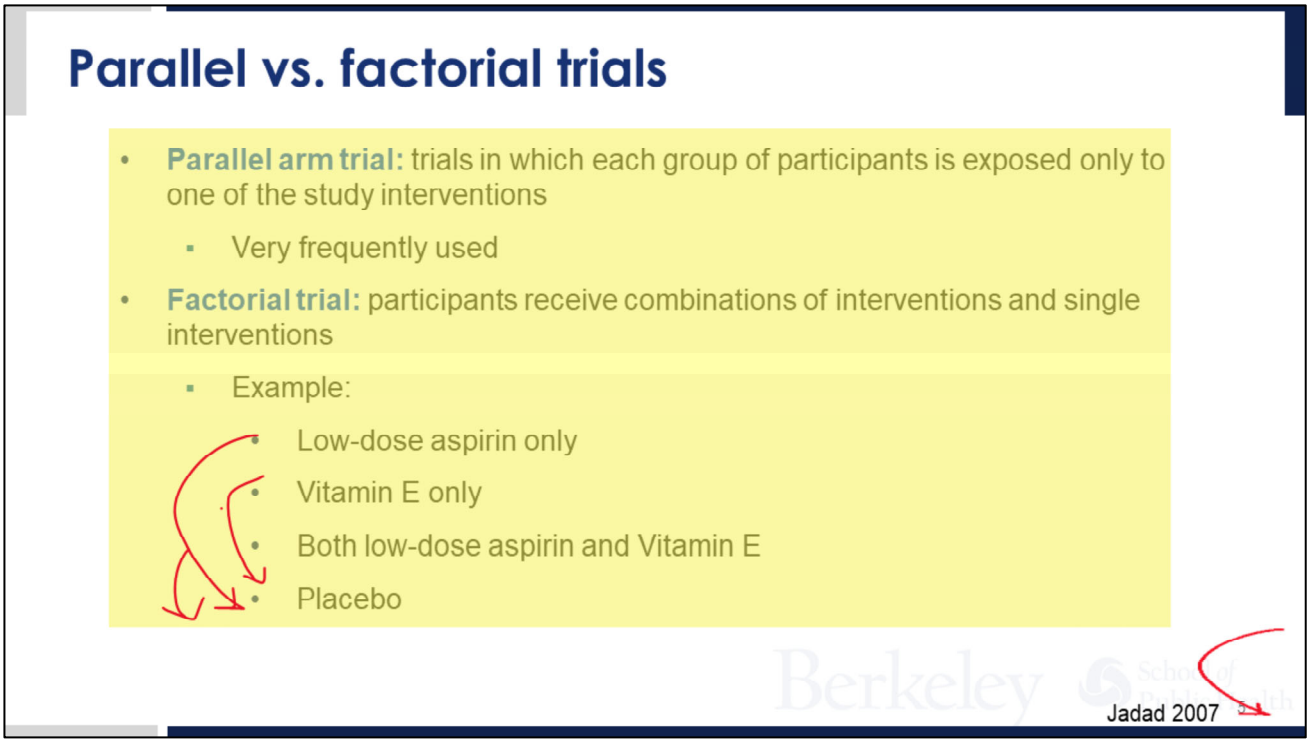
Phase one trials are really not trials. They're often done right after animal testing is completed on a small number of healthy volunteers. They're not randomized, and the trial is not controlled, though they are called trials. This is essentially taking a small group of people who are willing to try a new therapy that's been minimally tested and seeing how that therapy affects their health.

*After phase one trial is complete, we move onto a phase two trial, where the new therapy is given to a small number of patients with the condition the drug is intended to treat. This is different from phase one. In phase one, the trials are done on healthy volunteers, and in phase two, the trial is done on people with a particular disease that is being treated by the new regimen. The purpose of a phase two trial is to establish the relative efficacy of different doses of the new therapy or medication, and again, it's often not randomized or controlled.

*Once a phase two trial is complete, we move on to phase three trials, and these are typically equivalent to an effectiveness equivalence trial. They're real trials where participants are randomly assigned to the standard of treatment and then also to the new drug. In this case, that standard of treatment is serving as the comparison group, if it's an equivalence trial or perhaps a superiority trial.

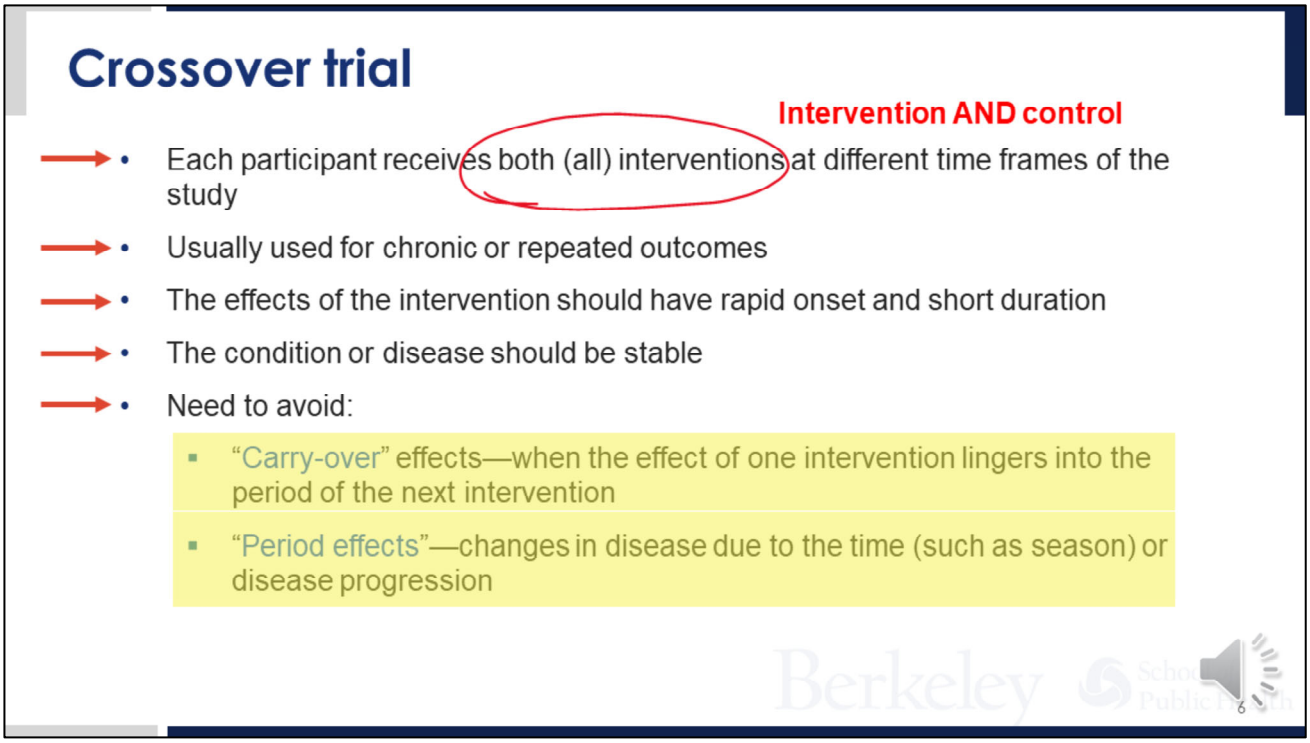
Once a phase three trial is done, the FDA and the United States can approve the drug for marketing, and then doctors can begin to prescribe it. *Phase four trials are even larger studies that monitor potential adverse events in very large populations, since often these adverse events are rare and would be difficult to detect in a phase three trial. These are studies that are done after marketing is approved, and these are also not truly trials. They're often not randomized or controlled. Typically they just involve surveillance from a clinical standpoint and surveys of different patients who've taken the medication.

So out of all of these so-called trials *it's really just the phase three trials that meet our standard definition of what a trial is in epidemiology



Next are parallel versus factorial arm trials. *Parallel arm trials are very frequently used. These are trials in which each group of participants is exposed to just one type of intervention or to control. New drug versus placebo, for example. *On the other hand, a factorial trial randomizes participants to combinations of interventions as well as single intervention.

For example, participants might be randomized to low dose aspirin only, vitamin E only, or both low dose aspirin and vitamin E, or to a placebo.* This trial would have a large enough sample that each of these different arms could theoretically be compared to each other, although usually what is done is that, for example, *the low dose aspirin only arm would be compared to placebo. *Vitamin E would be compared to placebo. *And then the both low dose aspirin and vitamin E would be compared to placebo. If the sample size was large enough though, any of these arms could be compared to each other. These are very powerful designs as long as the sample size is large enough to evaluate the effect of combining different interventions.



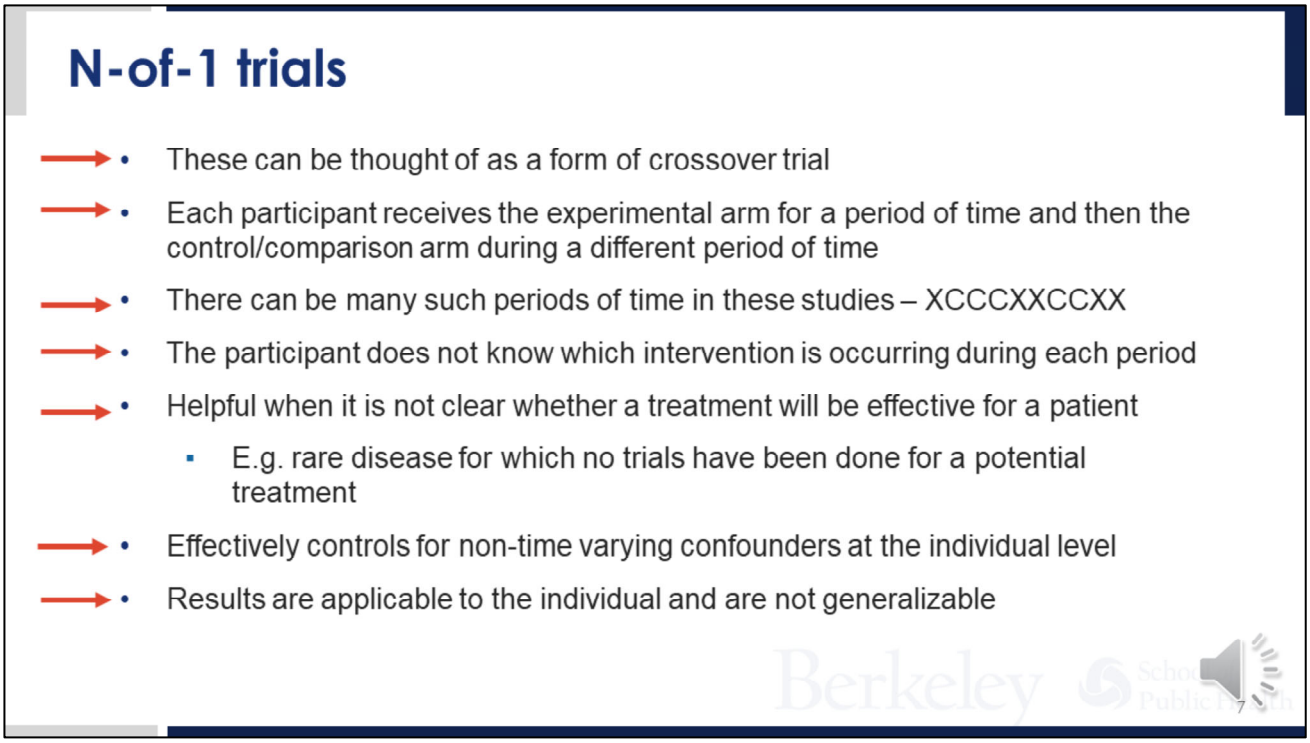
In a crossover trial, *each participant receives *both or all of the different kinds of interventions at different times during the study. *That means that they serve not only as an intervention participant but also potentially as a control participant over the course of one study. *These trials are appropriate for diseases that are chronic in nature or where there's multiple episodes of a particular health condition.

*But the effect of the intervention needs to be quite rapid and short in duration in order for this trial to work. *The condition or disease should also be stable. So for many chronic diseases, when they're early in their development stage, there may be lots of changes in physiology for the patient. This isn't a good time to conduct a crossover trial because the progression of the disease may have a mixed effect of the intervention, making it difficult to disentangle the independent contributions of each. *It's important to avoid two types of effects in crossover trials, *the first being carryover effects, when the effect of one intervention lingers into the period of the next.

For example, if the intervention period is too short and the intervention effect is lingering, when the patient switches to what is considered a controlled period, you would see that the measure of effect would become biased towards the null, because the treatment would be active during the control period, making the treatment and control periods more similar to each other.

*Another thing to avoid is period effects. This is changes in disease due to the time

of year, or some other time factor, that affects disease progressions. Seasonality is common for certain types of infectious diseases, for example. If you're concerned about period effects, one option would be to ensure that there is a similar balance of intervention and control periods across different seasons in order to capture the effects in a standard way as the seasons change.



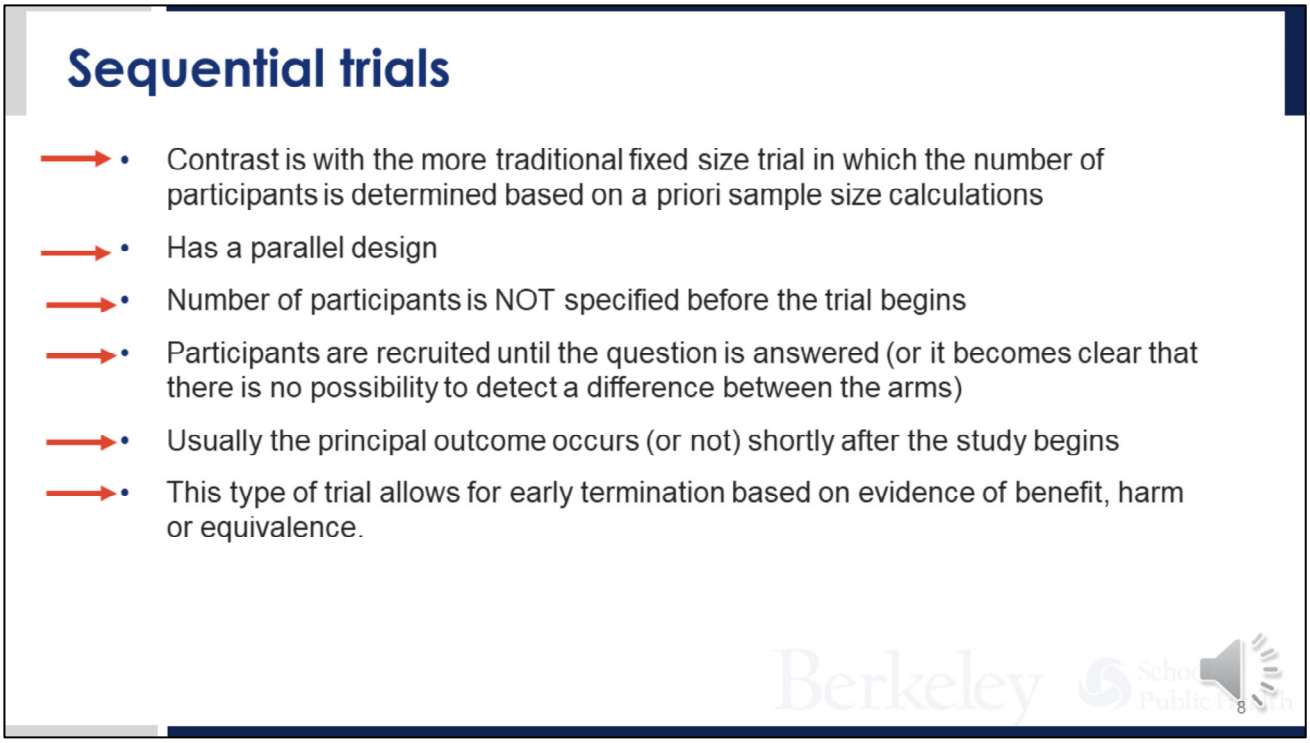
*A special type of crossover trial is called an n-of-1 trial, *where there is just one participant receiving the experimental arm for a period of time and then receiving the control or comparison arm for a different period of time. *There can be many different periods of time in an n-of-1 trial. Here x stands for intervention and c stands for control.

You could have a wide variety of different patterns where you start with intervention, followed by three control periods, followed by two intervention, and so on, depending on what the specific hypothesis of interest. We don't see these kinds of patterns in typical crossover trials just because of the logistical complexity of doing that, but in an n-of-1 study, it's a lot easier to experiment with these different periods of time in different arrangements.

*Ideally, the participant does not know which intervention is occurring in each period. So they're blinded to their treatment assignment, but that can be difficult for certain kinds of interventions. *This type of trial is very helpful when a clinician or a physician wants to know whether a specific treatment will be effective for a specific patient.

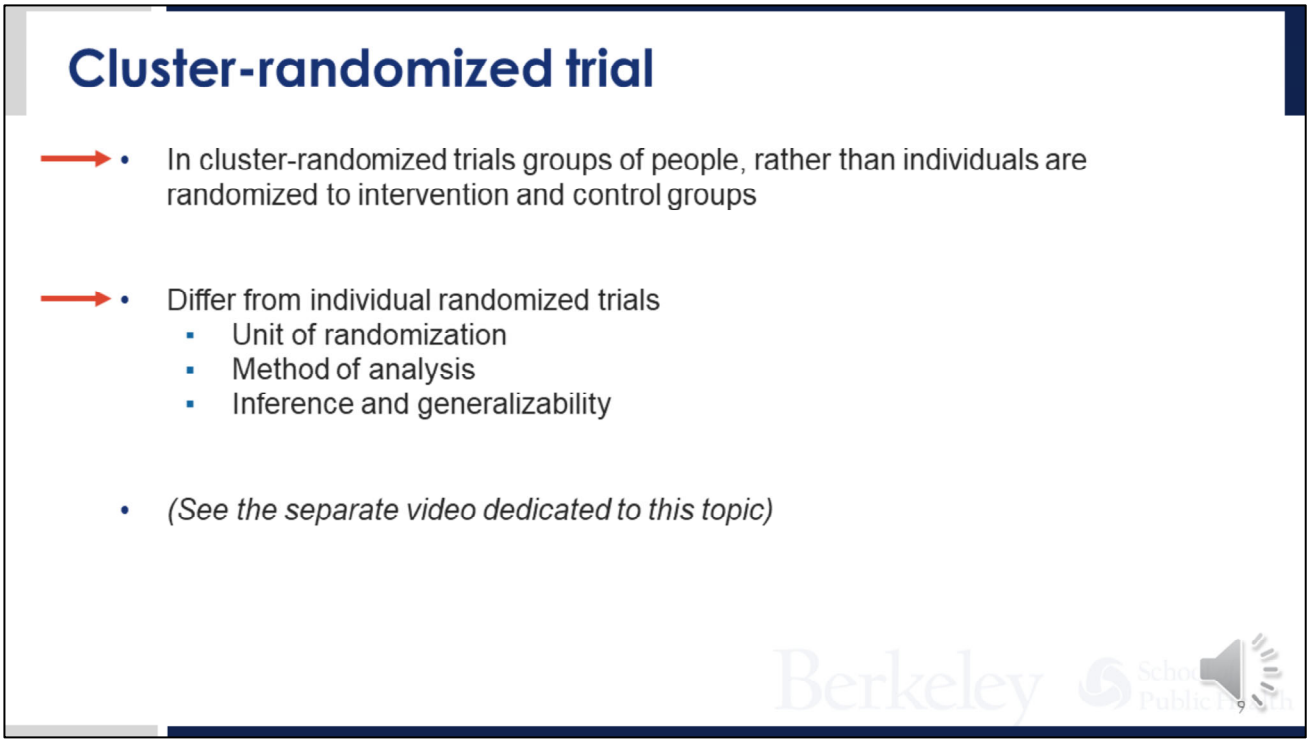
*This might be true for a rare disease, for which there have been no other trials to date. There is limited evidence about whether the new drug or new regimen will be able to treat the rare disease effectively. *One nice feature of an n-of-1 trial is that it's able to really effectively control for non-time varying confounders at the

individual level because it's conducted on a single person. So any fixed confounders will be the same throughout the course of the trial, regardless of which intervention or control time period the person is in. *And because an n-of-1 trial is conducted in an individual. the generalizability of an n-of-1 trial is to the patient, not to a population.



Another type of trial that's classified by the number of participants is a sequential trial. *This is in contrast to the traditional fixed-size trial. *A sequential trial will have a parallel design. *In a sequential trial, the number of people is not determined before the trial begins. *The trial starts, and interim analyzes are done, and additional participants are recruited as needed. This means that the question of interest, and potentially certain kinds of statistical results, need to be pre- specified before the study begins. Once there's enough data to answer that question of interest, or if it becomes clear that there is no possibility that there will be a difference detected between arms, the study can then terminate further recruitment.

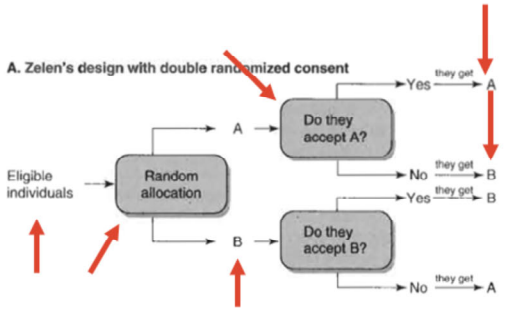
*This kind of trial is most effective when the primary outcome of interest occurs pretty quickly after the study begins. *A nice advantage of this is that if there's concerns about potential harms of a trial, the trial can be terminated very early. And on the other hand, if the benefits of the regimen that's being tested are very strong and can be seen rapidly, the trial can also be terminated quickly.



In a cluster-randomized trial, groups of people are randomized to intervention or control, rather than individuals. These kinds of trials differ from individual randomize trials in a number of ways. As I just mentioned, the unit of randomisation differs, the methods of analysis differ, and also the inference and generalizability of these trials differ, and we have a separate video fully dedicated to this topic.

Trials that account for non-randomized individuals' preferences

- **Zelen's design:** eligible individuals are randomized before they consent
- Evaluates the effect of offering the intervention to eligible individuals
- Limitations
  - Cannot be blinded
  - May have low statistical power if one arm is much more popular than the other



Jadad 2007

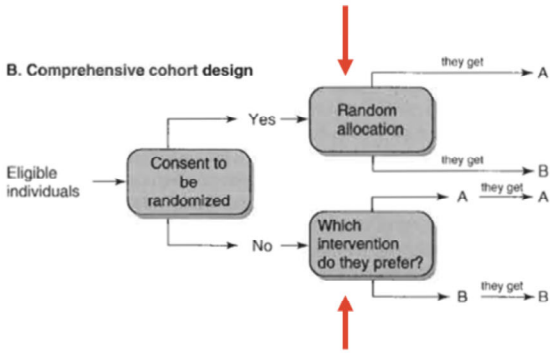
10

Next, I'll go through trial designs that account for non-randomized individuals treatment preferences. These particular trials are less common in a public health setting and are more commonly used in commercial settings for marketing research and other social types of interventions, but I'll go over them briefly here. *The first one is called Zelen's design, and in this design, eligible individuals are randomized before they consent to a particular treatment. You take a subset of **eligible individuals, *randomly allocate them to a or b in this particular figure, and *then if they accept treatment a in the group randomized to a, *they will get a. If they don't accept a, *they may get b or perhaps nothing depending on the design of the trial.

*The same process goes for the people allocated to arm b. *This design is evaluating the effect of offering the intervention to eligible individuals. A typical randomized design would be evaluating the effect of the intervention rather than offering the intervention, so that's the distinction here. *Two of the main limitations are first, that trial can't be blinded because the person has consented to a particular intervention. The second is that because one intervention may be much more popular than the other, you run the risk that the trial will have lower statistical power if you get a really large imbalance between the percentage accepting a versus b

Trials that account for non-randomized individuals' preferences

- **Comprehensive cohort design:**
  - Individuals who consent to randomization are randomly allocated to intervention groups.
  - Individuals who want to participate but do not consent to randomization choose the intervention they prefer.
- Convenient when participants have strong preferences and might not consent to randomization.
- Limitations
  - The non-randomized arm is a prospective cohort, so analyses must account for potential confounding



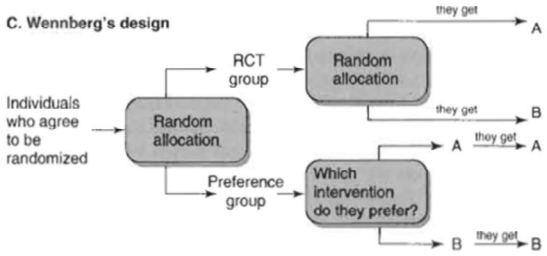
Jadad 2007

In a comprehensive cohort design, *individuals who consent to randomisation are randomly allocated to intervention groups, and *then individuals who want to participate but do not consent to randomization, choose the intervention that they prefer. So if we look at this figure here on the right, *on the top where it says random allocation, these people who have consented to be randomized are participating in a randomized trial. *But on the bottom where it says “which intervention do they prefer”, these individuals are essentially participating in a prospective cohort study. It's like two different studies with two different designs nested in one. *And this is a convenient design when participants have strong preferences and might not consent to randomization.

*In the second part of the study shown at the bottom of the figure, the prospective cohort, the individuals who choose a or b may be systematically different from each other, and so the analysis of this data must account for a potential confounding.

Trials that account for non-randomized individuals' preferences

- **Wennberg's design:** individuals who agree to be randomized are randomly allocated to participate in:
  - An RCT in which they are randomly assigned to intervention groups
  - A prospective cohort in which they choose which intervention they prefer
- Limitations
  - The non-randomized arm is a prospective cohort, so analyses must account for potential confounding



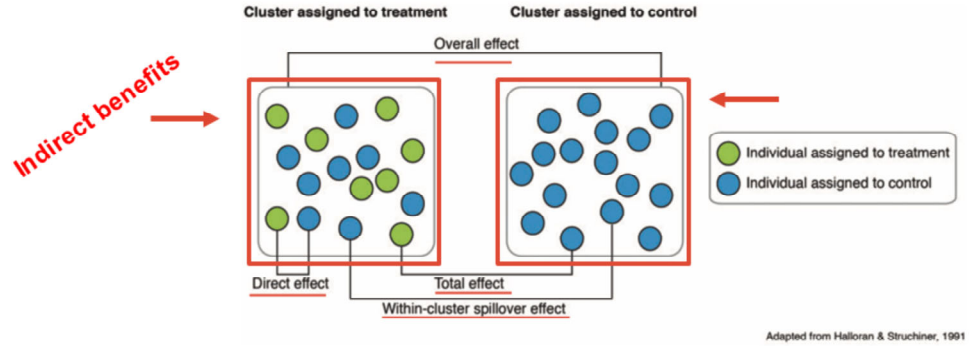
Jadad 2007

12

The last of these designs is called *Wennberg's design. In this design, individuals who agree to be randomized are randomly allocated to participate either in *an RCT, in which they are then randomly assigned to intervention group a or b, *or a prospective cohort, in which they choose which intervention they prefer. *As we mentioned in the last slide, the non-randomized arm, the bottom of the figure, is a prospective cohort study, essentially. The analysis of the data from the prospective cohort must account for potential confounding. And note that the main difference between the Wennberg's design and the comprehensive cohort design is Wennberg's design only enrolls those who consent to randomization to reduce bias if those who consent to randomization are different in important ways from those who don't consent.

Double-randomized trial

- Also called "two-stage randomized trials"
- First randomize clusters to intervention or control
- Then randomize individuals in intervention clusters to intervention or control
- Appropriate for interventions that may indirectly affect non-recipients (e.g., through reduced disease transmission)



Benjamin-Chung et al. Int J Epi, 2018, 332–347

13

Now I'll briefly touch on two new randomized designs that estimate different parameters of interest from those in traditional randomized trials. *The first is called a double-randomized trial, also can sometimes be called a two-stage randomized trial. *In this design, you start by randomizing clusters to intervention or control, and then within the intervention clusters, *you then randomize individuals to intervention or control. *So in this figure, on the left side, we see a cluster assigned to treatment, *and on the right side is a cluster assigned to control. *Within the cluster assigned to treatment, individuals, represented by these circles, are randomly assigned to treatment in green or control in blue. *In the control cluster everyone is going to receive the control. *This type of design is appropriate for interventions that may indirectly affect non-recipients through, for example, reductions in disease transmission. For the most part, this design has so far been used to evaluate infectious disease interventions, such as vaccines. And this is because vaccines produce something called herd effects, which are also called indirect effects or spillover effects, as noted in the figure. Essentially, the idea is that people who don't receive the intervention, but live in close proximity to people who do receive the intervention, may experience *indirect benefits simply from that proximity.

For example, if you vaccinate people in the treatment cluster, because of the vaccination you may see reductions in disease transmission, and the people who are assigned to control within the treatment cluster may thus have a lower risk of getting the vaccine-prevented disease, compared to the individuals in the control

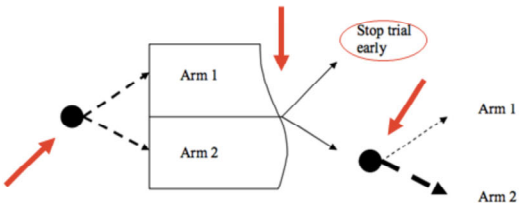
cluster, where no one was vaccinated.

There's a bunch of different kinds of effects that can be estimated using this design. *We can estimate a direct effect comparing people who were individually randomized to treatment or control in the treatment cluster. *We can estimate a total effect, comparing all of the people assigned to treatment in the treatment cluster to those in the control cluster, and that would be consistent with the typical parameter in a cluster-randomized trial. *We can also calculate a within-cluster spillover effect comparing the control individuals in the treatment cluster to those in the control cluster. *And then an overall effect would just average across people in the treatment cluster and compare that to the average in the control cluster.*

The reason that these different kinds of effects are of interest is that we may find that we don't have to offer an intervention to every single person to see large impacts on disease. If an intervention has indirect effects or spillover effects on untreated individuals who are near to treated individuals, it implies that we could just treat a subset of the population, which would greatly reduce cost and be more logistically feasible than treating everyone in the population. This kind of design is increasingly of interest in the infectious disease world, and this idea of herd effects will be familiar to anyone following covid-19 vaccine promotion.

Adaptive randomized trials

- Trial in which characteristics of the study itself, such as the proportion assigned to active intervention versus control, change during the trial in response to data being collected.
- Subjects are randomly assigned to Arm 1 or Arm 2.
- Perform an interim analysis:
  - If there is sufficient evidence that one intervention is superior or that it is virtually impossible that the interventions will differ from each other at the end of the trial, the trial is ended early.
  - If there is evidence of an effect in Arm 2, adopt a "play-the-winner rule" that adjusts the probabilities of assignment to favor Arm 2.
- Can help minimize risk for study participants and provide more rapid results.



Brown et al. Annu. Rev. Public Health 2009. 30:1–25

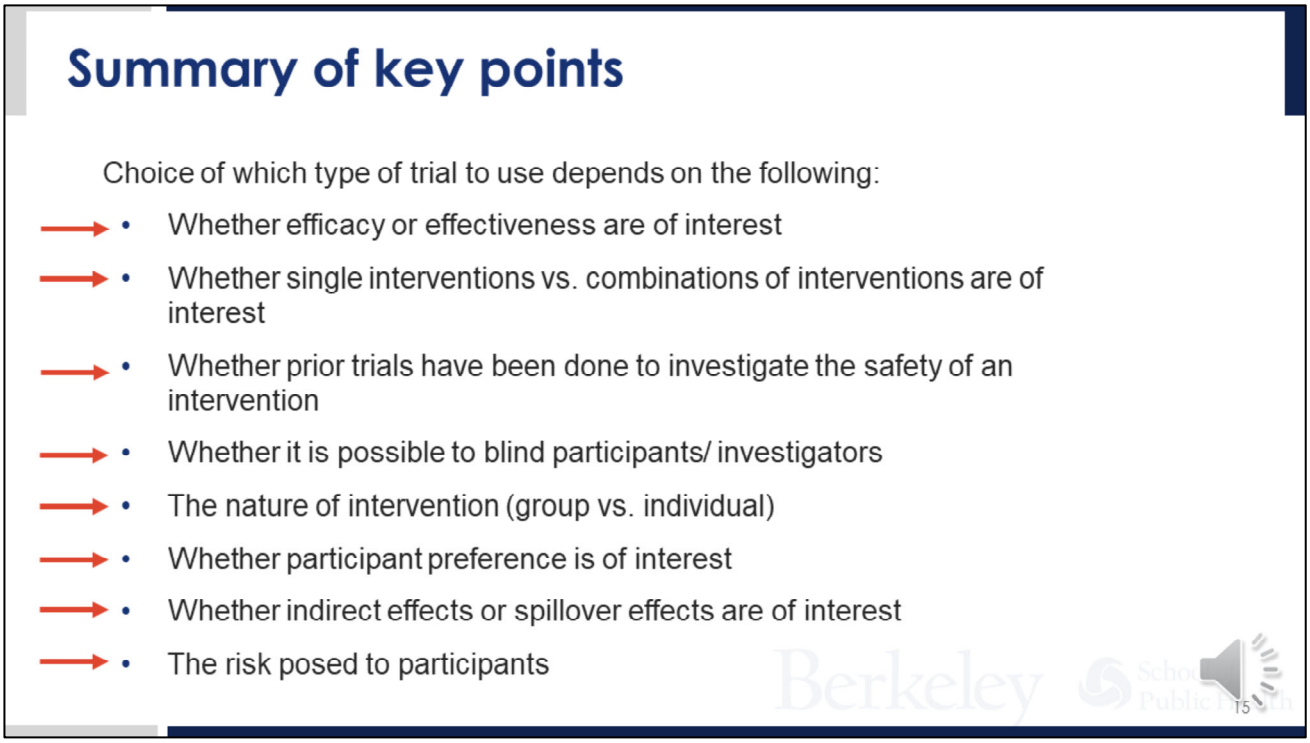
14

The last kind of design is an adaptive randomized trial. *This is a trial in which the characteristics of the study, such as the proportion assigned to active intervention versus control, can change during the trial in response to data that's being collected. *In this figure below, the black circle represents randomisation. People are randomized to arm one or two, and partway through the study an interim analysis is performed.

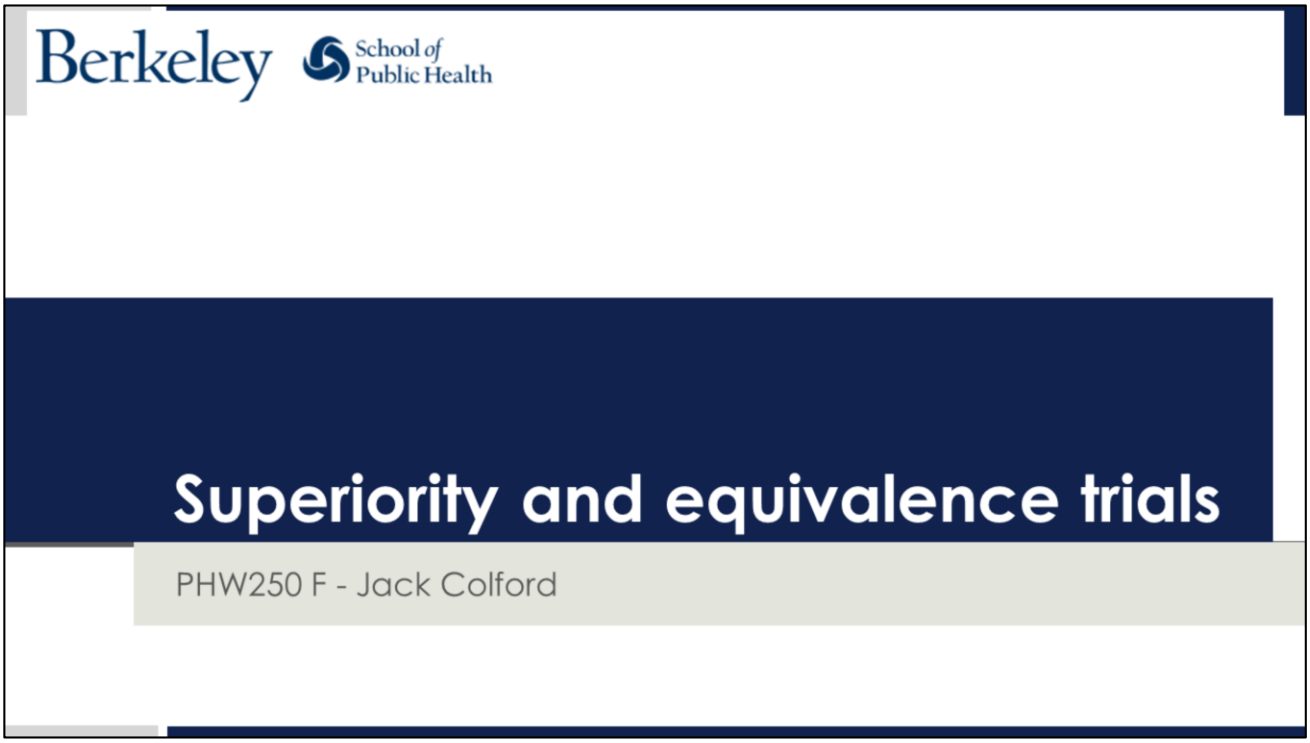
*At the time of the interim analysis, if there's sufficient evidence that one intervention is far superior or that it's virtually impossible to distinguish between the effects of the interventions, and that we don't expect that to change by the end of the trial, *the investigators can decide to end the trial early. However, if there is evidence of an effect in arm two, for example, *investigators can adopt something called a play the winner rule that adjusts the probability of treatment assignment in favor of arm two.

Participants would be randomized again and more people would be randomized to arm two than arm one if it's suspected that arm two is more beneficial.

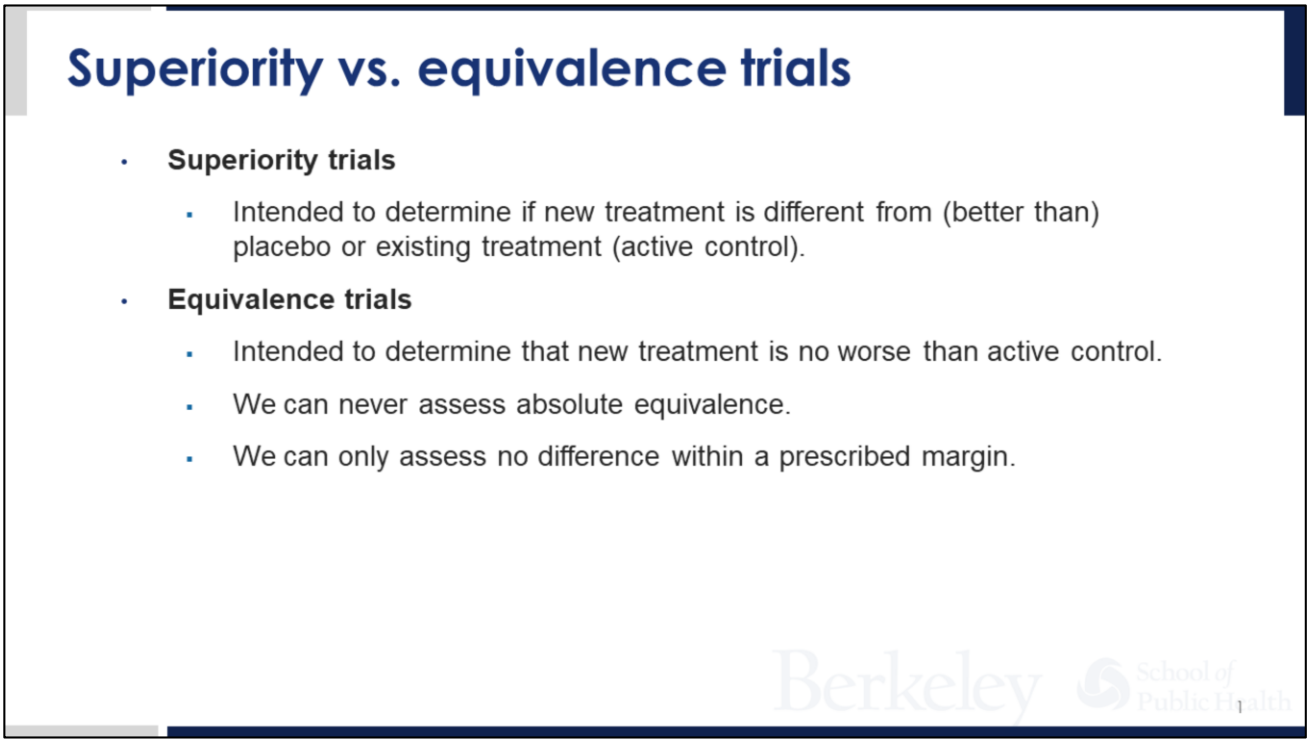
There are lots of variants of this, and I've just shown you one type of adaptive randomized trial, but this is something that's increasingly of interest epidemiologists, and one reason is that it can help minimize risk for study participants and provide much more rapid results than a trial that's not adaptive



To summarize, when thinking about what kind of trial design to use, it's helpful to consider whether efficacy or effectiveness are of interest, whether single or combined interventions are of interest, whether prior trials have been done to investigate the safety of an intervention, whether it's possible to blind participants or investigators, whether the intervention is at heart a group-level intervention or an individual-level intervention, whether you're interested in the preferences of participants, whether indirect effects or spillover effects are of interest, and also the level of risk posed to participants. All of these factors affect our choice of trial design.

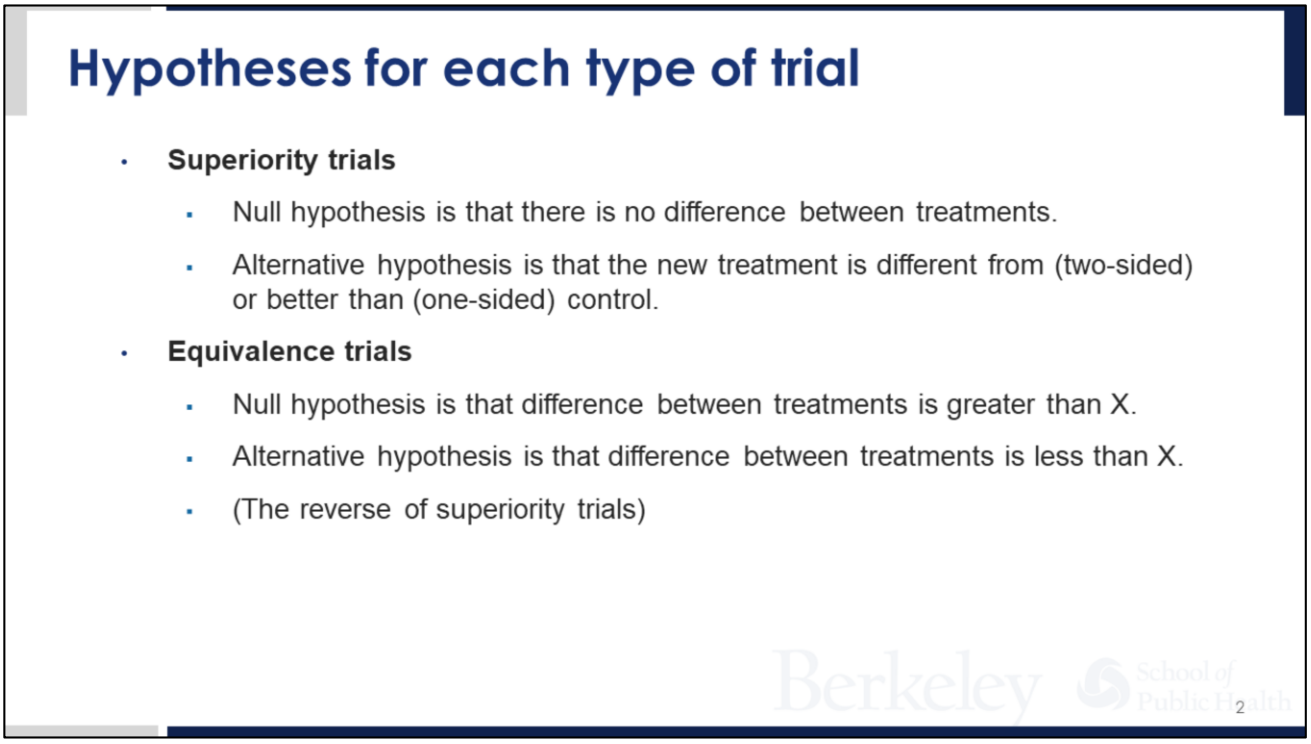


There are many ways to describe trials. One useful distinction is to ask whether a trial is a superiority or an equivalence trial.



Superiority trials are designed and intended to determine if a new treatment is different from, whether it's (better than) placebo or an existing treatment (or active control). In an equivalent trial, we're doing something different. We're trying to determine whether the new treatment is not worse than the active control.

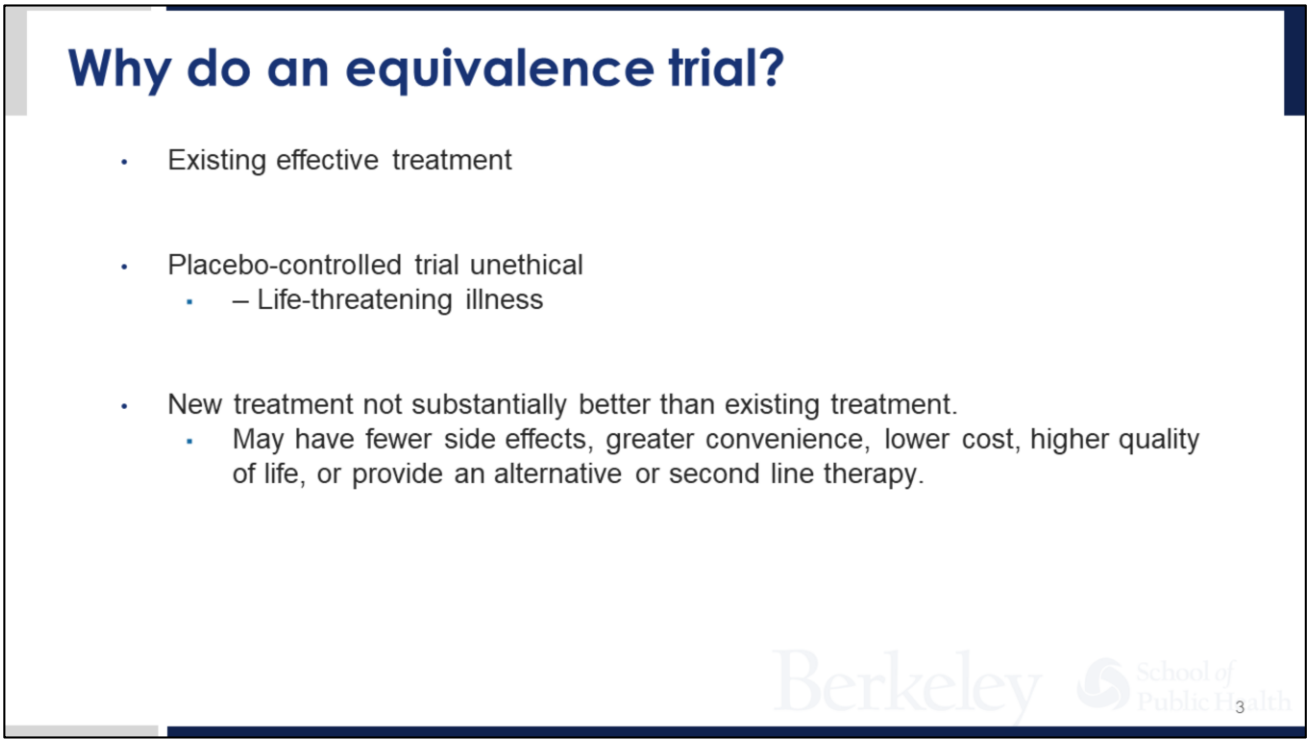
We can never assess absolute equivalence. This will be clear to you as we show some pictures in a few moments. And we can only assess no difference within some prescribed margin. Before the equivalent trial begins, the investigator needs to state what the margin is, within which, equivalence will be determined.



There are different hypotheses for each type of trial. In a superiority trial, the null hypothesis is that there is no difference between treatments. The alternative hypothesis, is that the new treatment is different from (in a two-sided way) or better than a (one-sided) control.

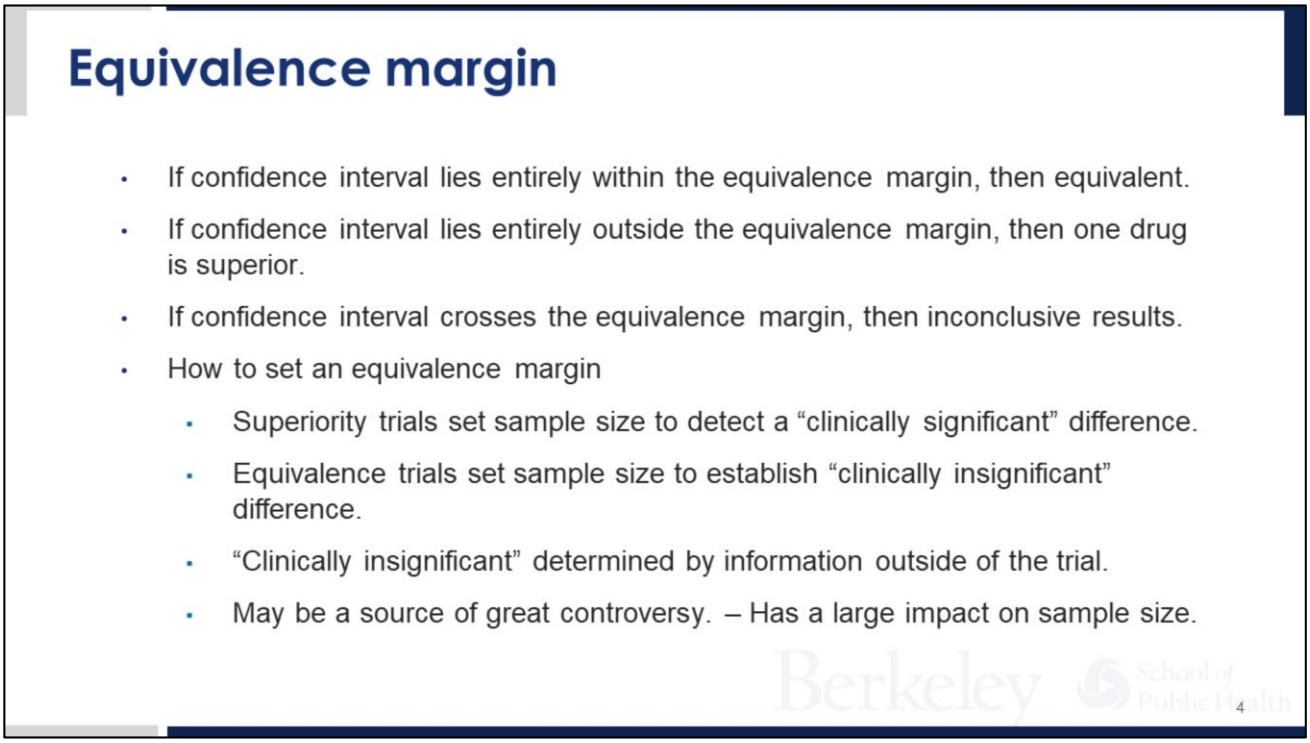
We're trying to show that the new treatment is either different from or better than the control. If trying to show that it's different from, that called a two-sided test, or if we're just trying to show that better than, that's a one-sided test, or a one-sided control.

In an equivalent trial, the null hypothesis, however, is that the difference between treatments is greater than some level x that we predetermine. The alternative hypothesis is that the difference between the treatments is less than that predetermined margin of x. This is the reverse of a superiority trial. Look at how the null and alternatives are reversed between the superiority and equivalent trials.



Why do we do equivalence trials? We often do them to evaluate an existing effective treatment. It might be done in a situation where a placebo-controlled trial would be unethical to do. That is, there might be a life-threatening illness involved, so it won't be possible to give one group nothing. Both groups need to get some sort of treatment to be ethical.

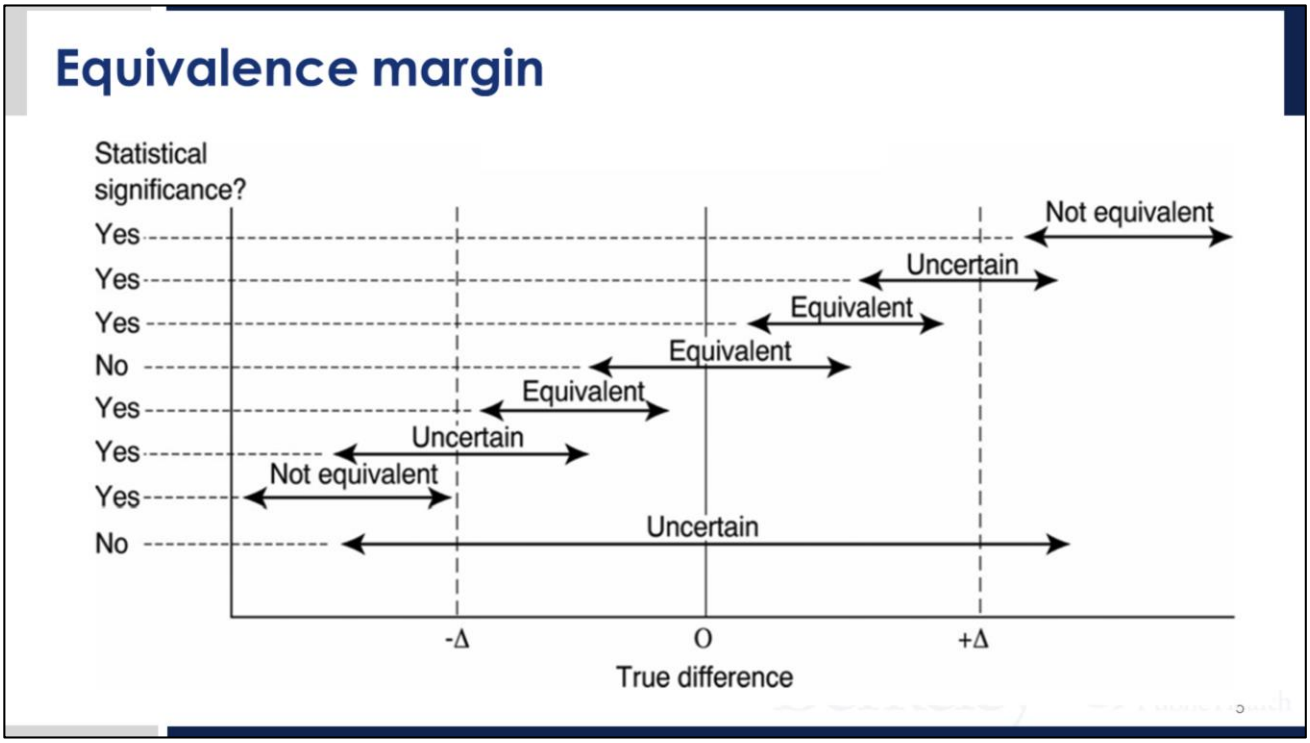
Or we may have a situation where the new treatment is not substantially better than the existing treatment but the new treatment might be alleged to have fewer side effects, greater convenience, lower cost, higher quality of life, or provide an alternative or second line therapy. And that's why we might do an equivalence trial.



I mentioned earlier this idea of setting up a margin. This is called an equivalence margin. If the confidence interval-- describing the difference between the two arms of the trial-- lies entirely within this equivalence margin, then the two treatments are declared equivalent. And I'll show this in a picture in a moment. If the confidence interval lies entirely outside the equivalence margin, then one drug is superior. If the confidence interval crosses the equivalence margin, then the results are inconclusive.

How do we set these equivalence margins? Well, superiority trials set sample size to detect a clinically significant difference. And that's a different situation than the equivalence trial. Equivalence trials set sample size to establish clinically insignificant differences between the arms. We're trying to demonstrate that the two arms of the trial are equivalent to each other.

But this concept of clinically insignificant is important. And it's determined by information outside of the trial. And it can be something that investigators argue about, and causes a source of great controversy, because it has a large impact on what the sample size is. I think all of this will make much more sense as we look at some pictures of equivalence versus superiorly trials.



In this example, we're showing the difference between two arms of a trial. Imagine we had arm A and arm B-- new treatment versus old treatment. If there was no difference in the effectiveness of the two arms of the trial, it would fall at the level of zero. If there was no difference right in the middle there, between the two arms of the trial we'd record a value of zero.

If the new drug is better, then we might be moving to the right. And if the new drug is worse, we might be moving to the left, in terms of the difference here. The bottom, on the x-axis, we're showing a hypothetical equivalence margin-- plus-delta and minus-delta. The investigator determines how big delta should be. So maybe delta is a difference of 10%. We would have plus 10% way down to minus 10% difference between the two arms of the trial.

Let's look at all the different results that could occur when we compare drug A to drug B. The very first line, let's say that drug A versus drug B is greater than 10% difference. The confidence interval falls completely above that plus-delta or plus 10% margin.

That's the very first row-- the top of the picture. There the two drugs are not equivalent, because the confidence interval is completely outside the margin of equivalence. The margin of equivalence goes from minus-delta to plus-delta. The confidence interval of our result-- comparing A to B-- is completely outside that. So the drugs are not equivalent.

Let's look at the second row. In the second row, the statistical significance between the drugs, because the confidence interval does not overlap zero, which would imply no statistical significance, no difference. However, because the confidence interval falls both below and above the plus-delta margin-- which, in this example I just pretend is 10%-- because it falls both above and below, we say that we are uncertain. We can't determine whether the drugs are equivalent or not.

Third row-- now we've compared drug A and B. This time, confidence interval does not cross zero, so the drugs are statistically different. However, the confidence interval falls completely within the margin between minus-delta and plus-delta. We declare the drugs to be equivalent, because the difference between the two drugs falls between minus-delta and plus-delta-- minus 10% and plus 10%.

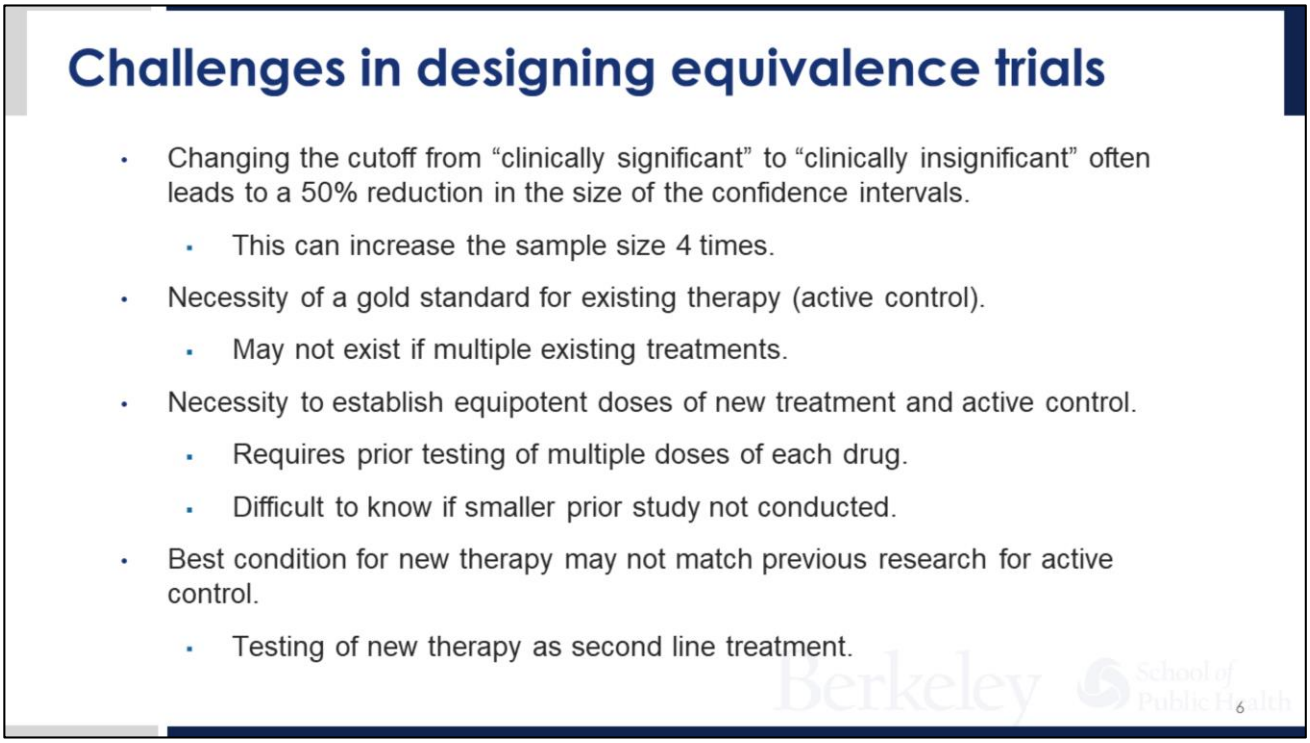
The fourth example now. This time, the confidence interval crosses a difference of zero. In this situation, the drugs are not statistically different from each other, because it crosses the null hypothesis. However, the drugs are still equivalent, because the entire confidence interval falls between the negative delta of minus 10%, and the positive delta of plus 10%.

Next example, row five--these two drugs compared now, are statistically different. The confidence interval does not cross zero. However, they are still equivalent, because of the confidence interval falls completely within the margin from minus-delta to plus-delta.

Next row, row 6-- here, we are uncertain. We know that drugs are statistically different from each other, however, the confidence center falls less than the minus-delta margin. That's outside the boundary between minus-delta and plus-delta. We are uncertain of whether there is equivalence of the two drugs. They might or they might not be.

In the next row-- that's row 7-- we still have statistical significance. The confidence interval does not cross zero. But in this example, the drugs are not equivalent, because the entire confidence interval is outside the margin between minus-delta and plus-delta.

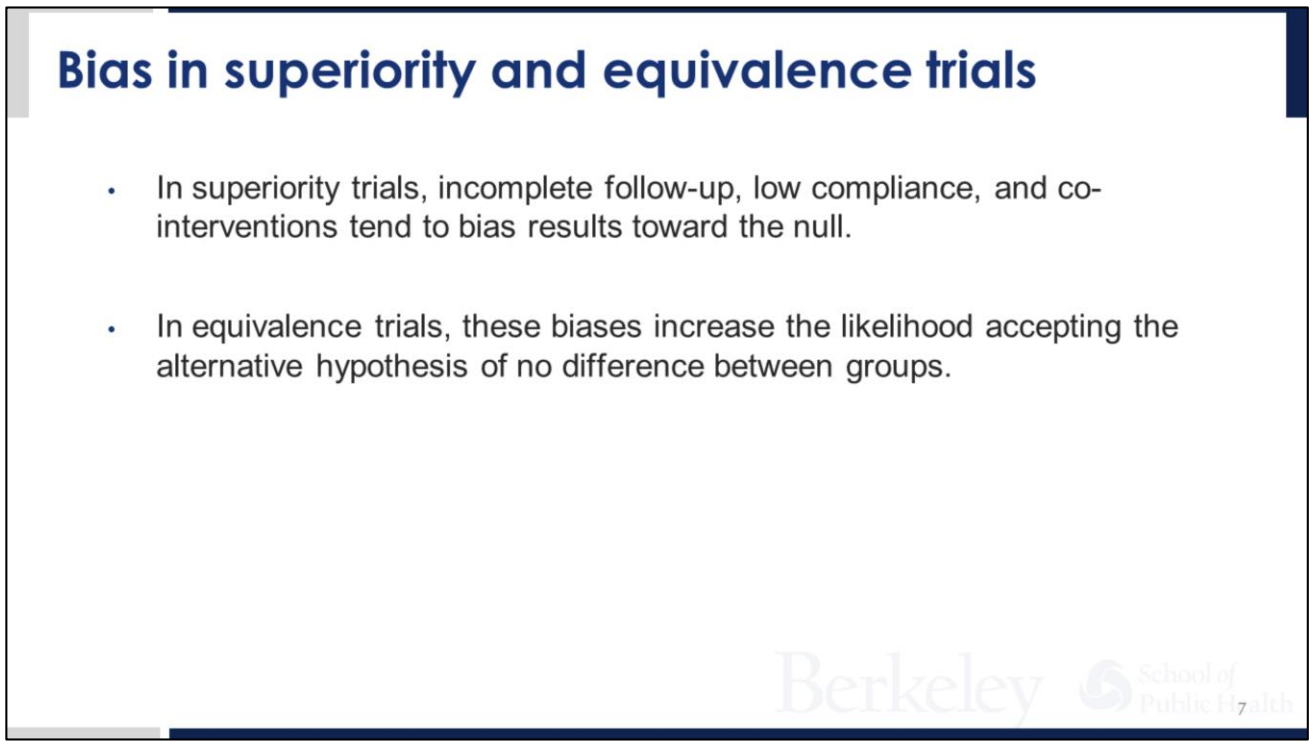
In the final example, we have a very wide confidence interval, and we're completely uncertain. Because, well, we know there's no statistical significance between the drug, because the confidence interval crosses zero. However, the confidence interval also extends outside the margin from minus-delta to plus-delta. In fact, it extends both above and below the margin. Therefore, we are uncertain about the equivalence of the drugs.



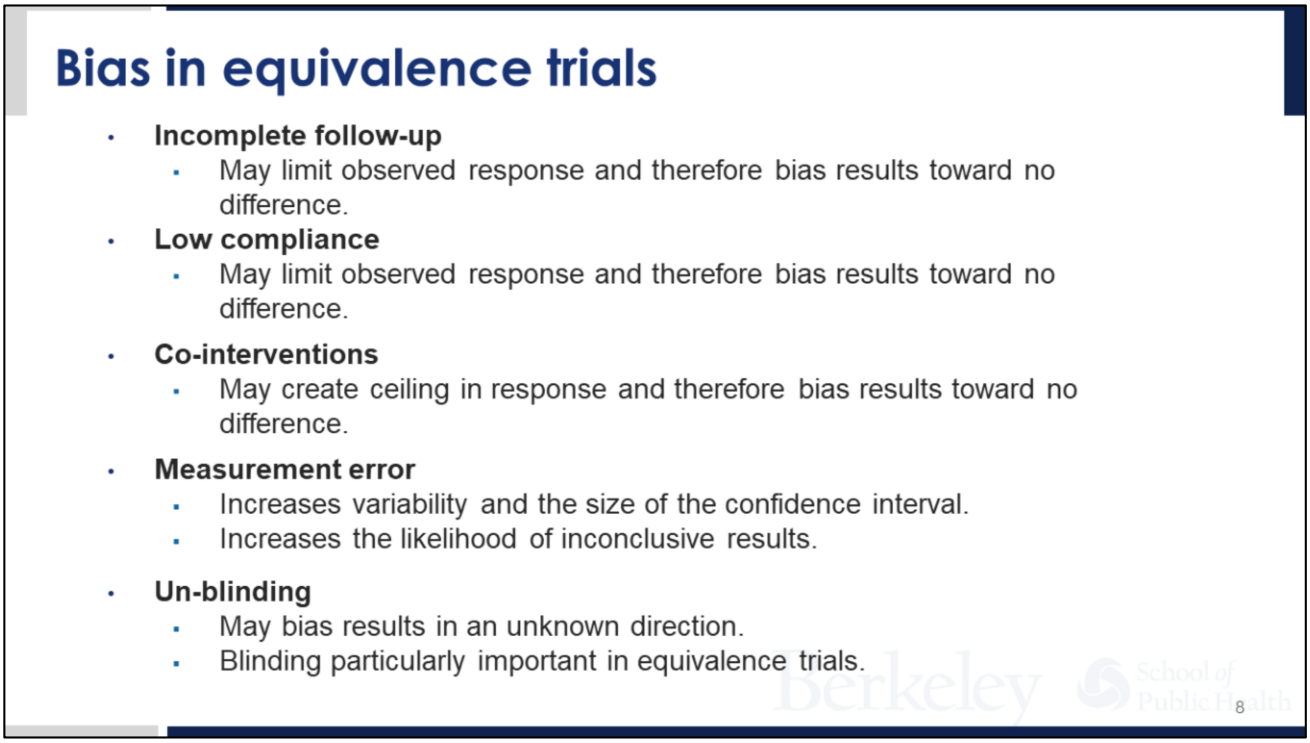
There are challenges in designing equivalence trials. Changing the cutoff from clinically significant to clinically insignificant often leads to a 50% reduction in the size of the confidence interval. This can increase the sample size four times. This is something we haven't learned yet in the course. But this has a tremendous impact on the size of a trial required. Showing that drugs are equivalent requires a much larger sample size.

There is a necessity of using, as a gold standard, the existing therapy-- the active control that we're comparing against. This may not exists if there are multiple possible existing treatments against which to compare the new treatment. There is a necessity to establish equipotent doses of the new treatment in active control. This requires prior testing of multiple doses of each drug, and it's difficult to know if there were no smaller prior studies done to help us know this information.

And the best condition for new therapy may not match previous research for active controls. Here, we have the testing of new therapy as a second-line treatment or possibility.

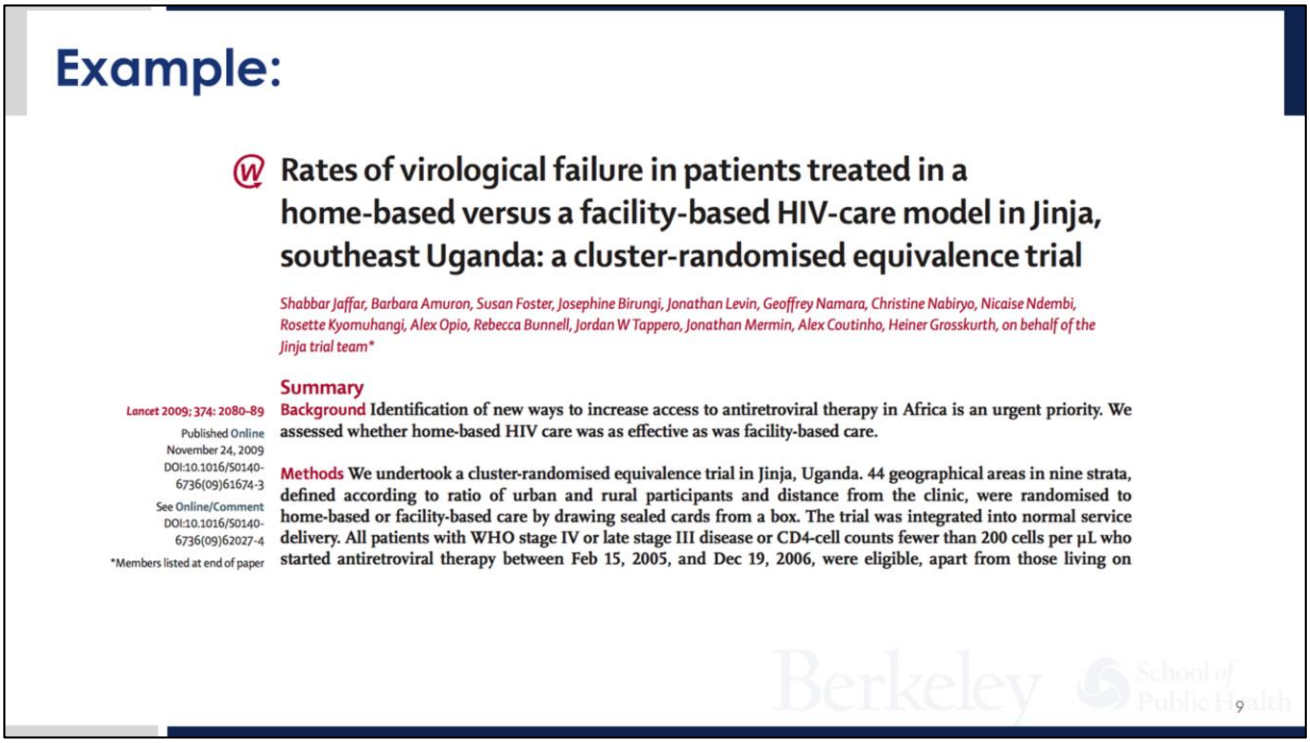


There can be bias in superiority and equivalence trials. In superiority trials, incomplete follow-up, low compliance, and co-interventions tend to bias results toward the null. In equivalence trials, and these biases increase the likelihood of accepting the alternative hypothesis of no difference between groups.



Sources of bias in equivalence trials include the possibility of incomplete follow-up, which may limit observed response, and therefore the bias results towards no difference. Or low compliance, which may limit the observer response, and therefore, again, the bias results tend towards no difference.

There might be co-interventions. These might create a ceiling in the response, and therefore the bias results in no difference. Or there might be measurement error, which increases the variability in the size of the confidence interval, and increases the likelihood of inconclusive results. Finally, there might be unblinding. And unblinding may result in bias in an unknown direction. Blinding's particularly important in equivalent trials because of this.



Let's try to make all this operational with looking at a specific study by Jaffar and colleagues, looking at the rates of virological failure in patients treated in a home-based versus a facility-based HIV care model in Uganda. This was a cluster randomized equivalence trial. And what's being done here, is an attempt to compare home-based therapy to facility-based therapy for HIV care.

Background: location of HIV care in Africa

- To reduce the burden of HIV, a large proportion of people infected with HIV must receive care (including antiretroviral therapy).
- There is a severe shortage of clinical staff to deliver treatment, and it is difficult for patients to accessing clinics
- The trial compared home-based versus a facility-based HIV-care model in Uganda
- Outcome: virological failure (a measure of the body's ability to suppress reproduction of the HIV virus)

Image source: https://redpepper.co.ug/man-dies-of-suspected-marburg-virus-in-mityana-hospital/

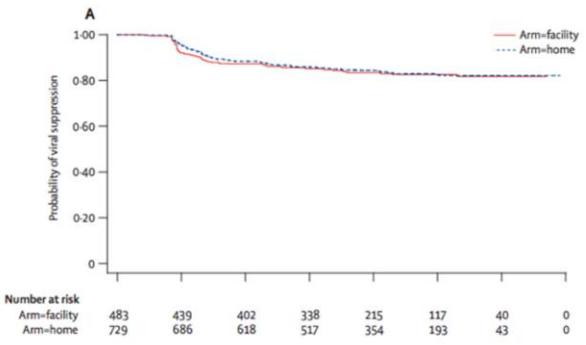
Berkeley School of Public Health

10

To reduce the burden of HIV, a large proportion of people infected with HIV must receive care, including anti-retroviral therapy. There is a severe shortage of clinical staff to deliver treatment, and it's difficult for patients to assess clinics. This trial compared home-based versus facility-based HIV care in Uganda. The two therapies here are home-based therapy versus facility-based therapy. And the outcome being measured was virological failure, which is a measure of the body's ability to suppress reproduction of the HIV virus.

Trial design: location of HIV care in Africa

- Assumed that in one group the rate of virological failure would be about 20%
- The other group would be considered equivalent if the rate of virological failure did not exceed 20% by more than 9% (ie, it was ≤ 29%)
- Results:
  - **Home care:** viral suppression rate per 100 person-years = 8.19 (95% CI 6.84, 9.82)
  - **Facility care:** 8.67 (6.96, 10.79)
  - Rate ratio = 1.04 (0.78, 1.40)
    - equivalence shown

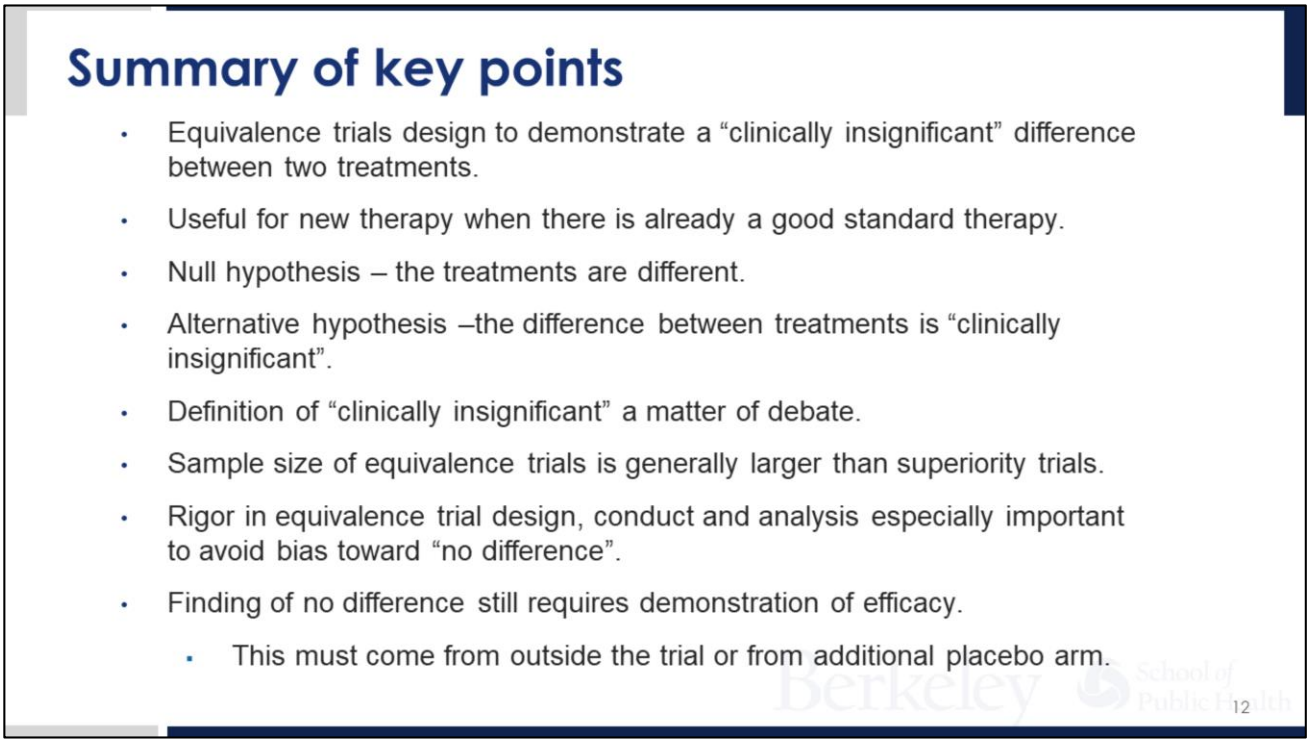


Kaplain-Meier Curve of Viral Suppression

11

It was assumed that in one group, the rate of virological failure would be about 20%. Prior to this study, the investigators defined themselves that the other group would be considered equivalent if the rate of virological failure did not exceed 20% by more than 9%. In other words, that it was between plus 9% or it could have been minus 9%. So that would be from 20 up to 29, or from 20 down to 11.

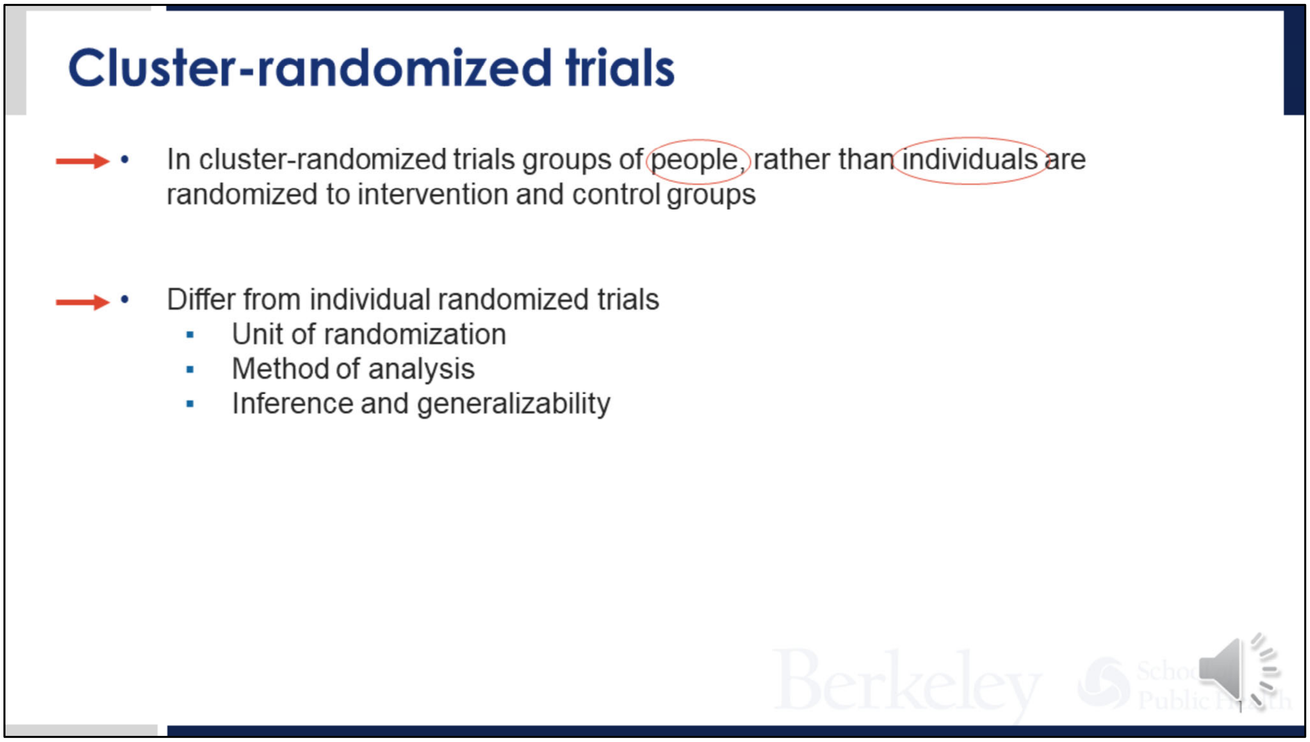
Let's look at the results of the study. In the home-care group, the viral suppression rate per 100 person-years was 8.19. And you see the confidence interval there-- 6.84 up to 9.82. In facility care, the rate was 8.67. The confidence interval there was 6.96 up to 10.79, giving us a rate ratio of 1.04. This demonstrated equivalence, because the two groups fell within the margin of difference-- a 9% difference above and below-- that was defined before the study as the equivalent margin.



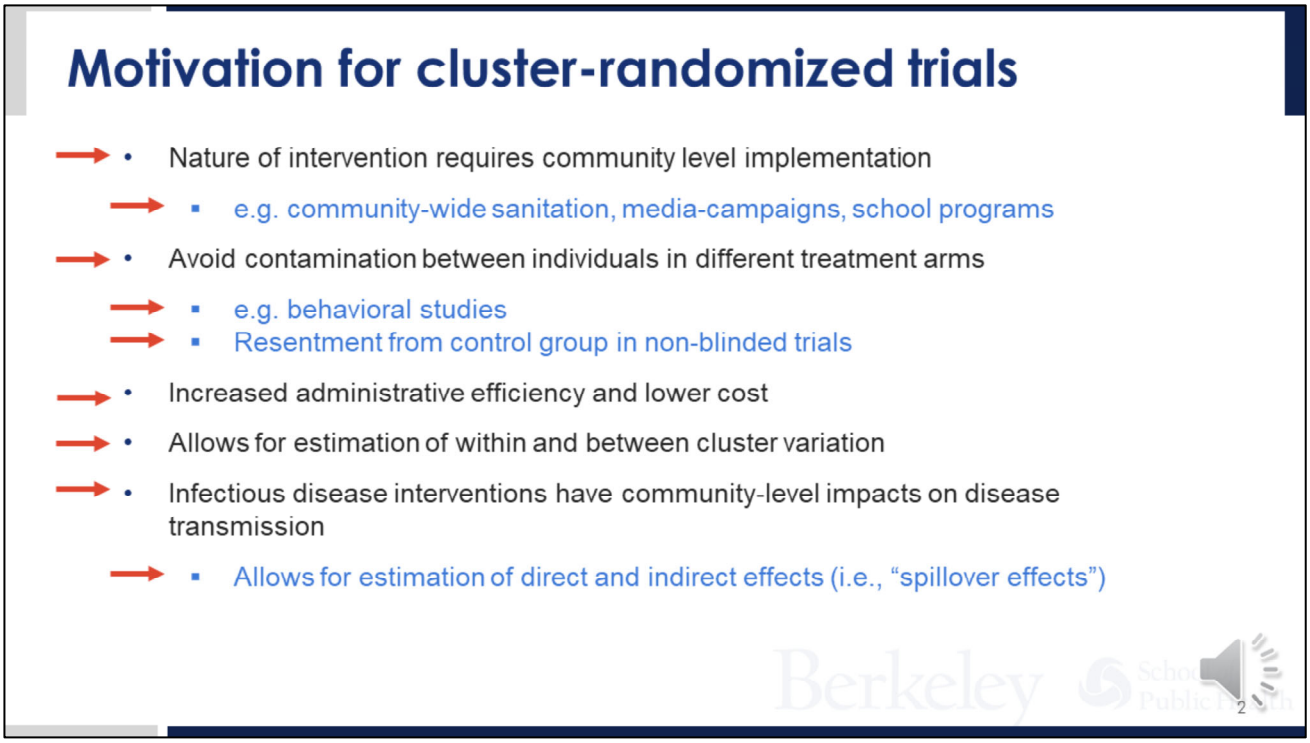
To summarize the key points, equivalence trials are designed to demonstrate a clinically insignificant difference between two treatments. They're useful for new therapy when there's already a good standard therapy. The null hypothesis is that the treatments are different. The alternative hypothesis is that difference between the treatments is clinically insignificant. And this definition of clinically insignificant is a matter of debate. So investigators have to define it, but people can then still argue about it.

The sample size of an equivalence trial is generally larger than that of a superiority trial. Rigor is required in equivalence trial design, conduct, and analysis-- especially because it's important to avoid bias toward no difference. And I discussed how many different ways can push a trial towards no difference. And therefore, push it towards looking like equivalence in a biased way. A finding of no difference still requires a demonstration of efficacy. This has to come from outside the trial, or from some additional placebo arm.

This video focuses on cluster-randomized trials.



*In cluster randomized trials, groups of *people, rather than *individuals, are randomly assigned to the intervention or the control group. *This kind of trial differs from individual randomized trials. The unit randomization is different, the method of analysis is different, and the inference that we make, as well as the generalizability of the trial, can also be different.



The motivation for cluster-randomized trials is that *certain kinds of interventions require community level implementation. These types of interventions are best evaluated using a cluster-randomized trial instead of an individually-randomized trial. *For example, a community wide sanitation intervention may include a social media aspect or a social campaign, such as billboards to promote good sanitation. It may include household as well as shared latrine improvements and hand-washing stations, and it might include other types of behavior change interventions that are delivered through schools, churches, and other community organization. For this type of intervention, it wouldn't be appropriate to use an individually-randomized design, because the intervention is not delivered at the individual level. So it's not only expected to have effects on the individual, but it's also expected to have community-wide effects. This is also true, generally speaking, for media campaigns and programs delivered through schools.

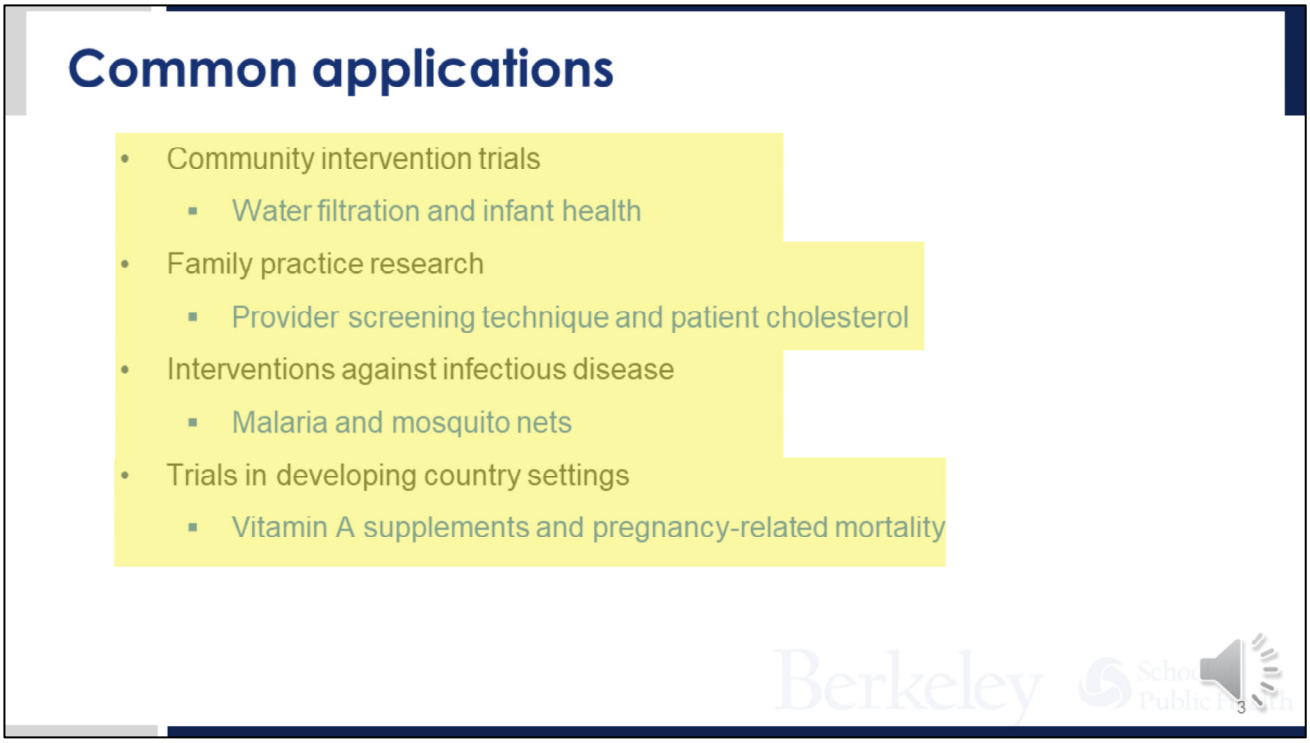
*Another motivation for using cluster-randomized trials is when we're concerned about contamination between individuals in different treatment arms. *This could be true in a behavioral study or in a study of infectious diseases.

For any kind of outcome that could be diffused or transmitted between different people in the population, we have to be very careful that people assigned to different treatments don't interact with each other in a way that causes their treatment assignment to change. For example, for an exercise promotion intervention, if an individual who is assigned to the exercise promotion program

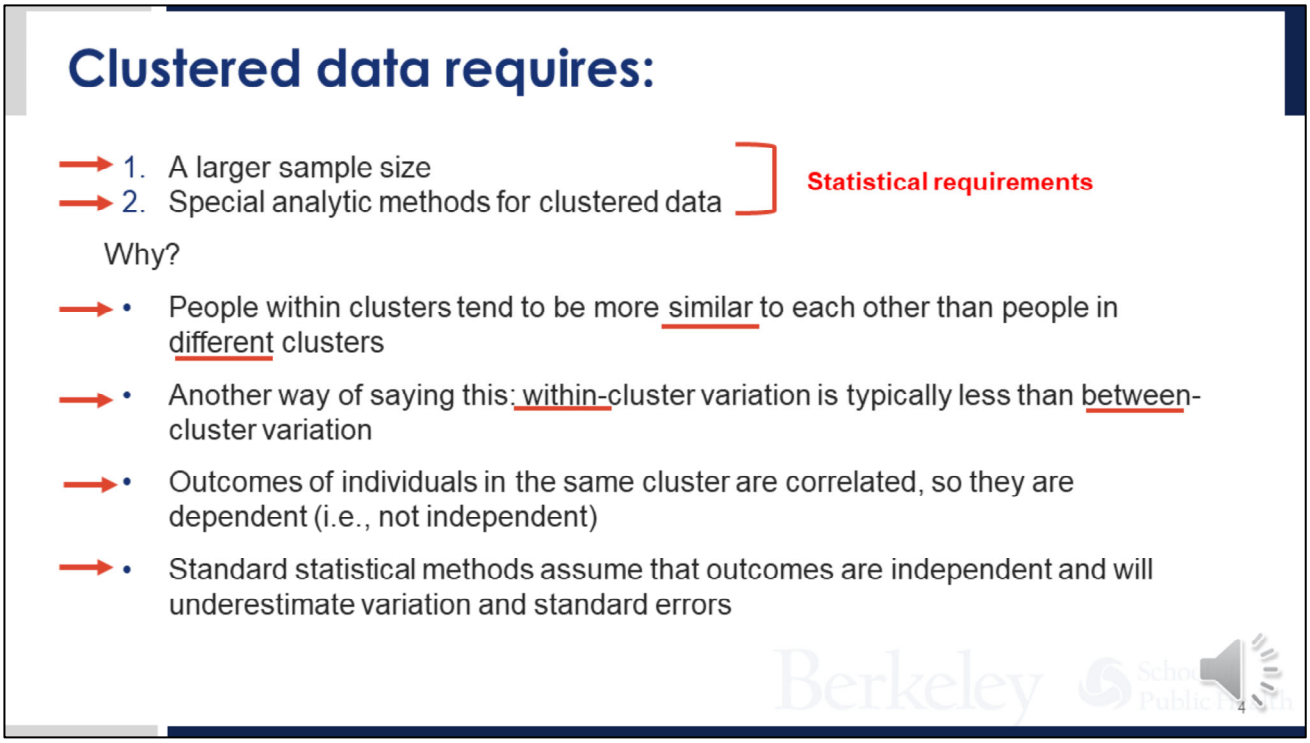
came into contact with an individual assigned to control, they may share details with the control person about the intervention that causes the control person to change their physical activity. We would call this contamination. Cluster-randomized trials can help minimize this kind of contamination by randomizing groups of people with a large distance between the groups, in order to avoid contact between different groups, but to allow for contact within the same group.

*Cluster-randomized trials are also a good choice when there's concerns that in non-blinded trials people in the control group may resent people in the intervention group. Because of the distance that typically is set between intervention and control clusters, this kind of resentment is less of an issue in cluster-randomized trials than in individually randomized trials.

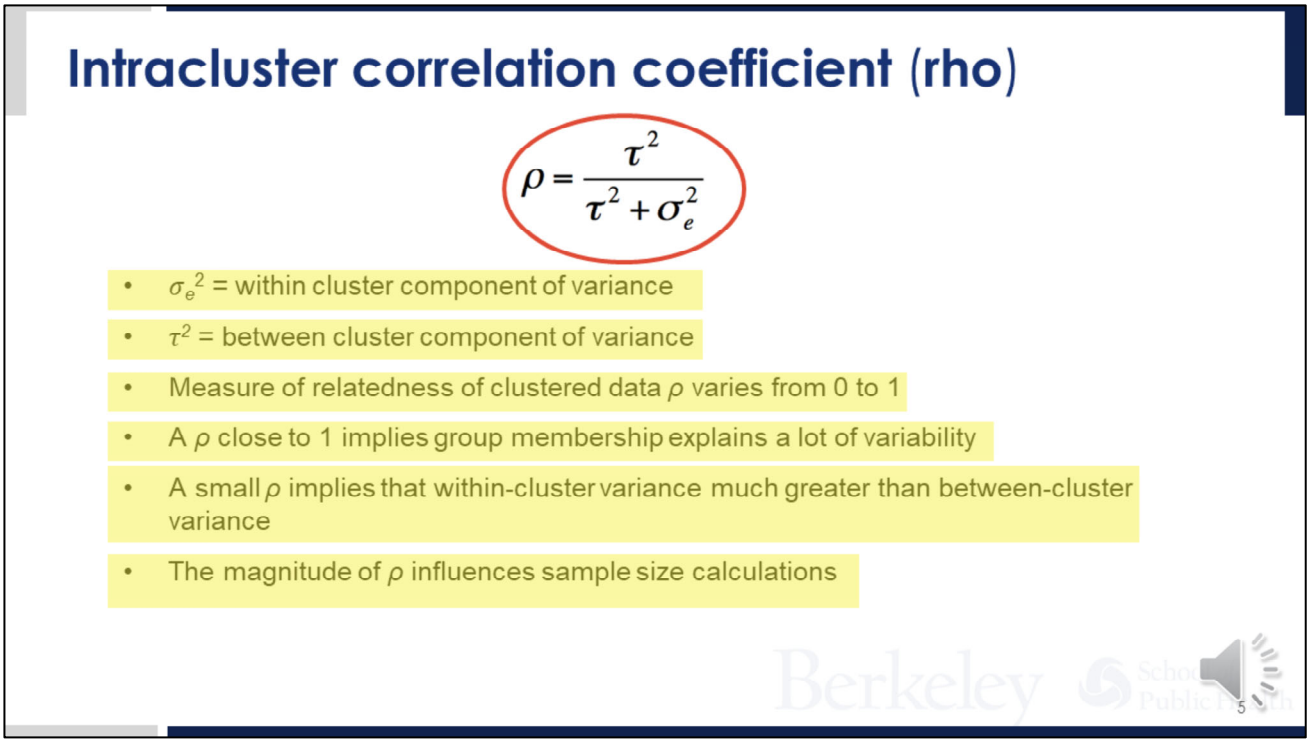
* In some cases, randomizing clusters instead of individuals can increase administrative efficiency and reduce cost because logistics can be planned around groups of people, whereas in an individually-randomized cluster, it can take a quite a bit of effort to track down individual people who are scattered across population. *Also, sometimes we're interested in explicitly estimating the within and between cluster variation. And if that is of interest, we need to use a cluster-randomized design instead of an individually-randomized design. *And finally, as I mentioned for infectious disease interventions, they almost always had community-level impacts on disease transmission. And it's also often of interest to calculate direct and indirect effects or spillover effects. Cluster-randomized trials allow us to do this in many cases.



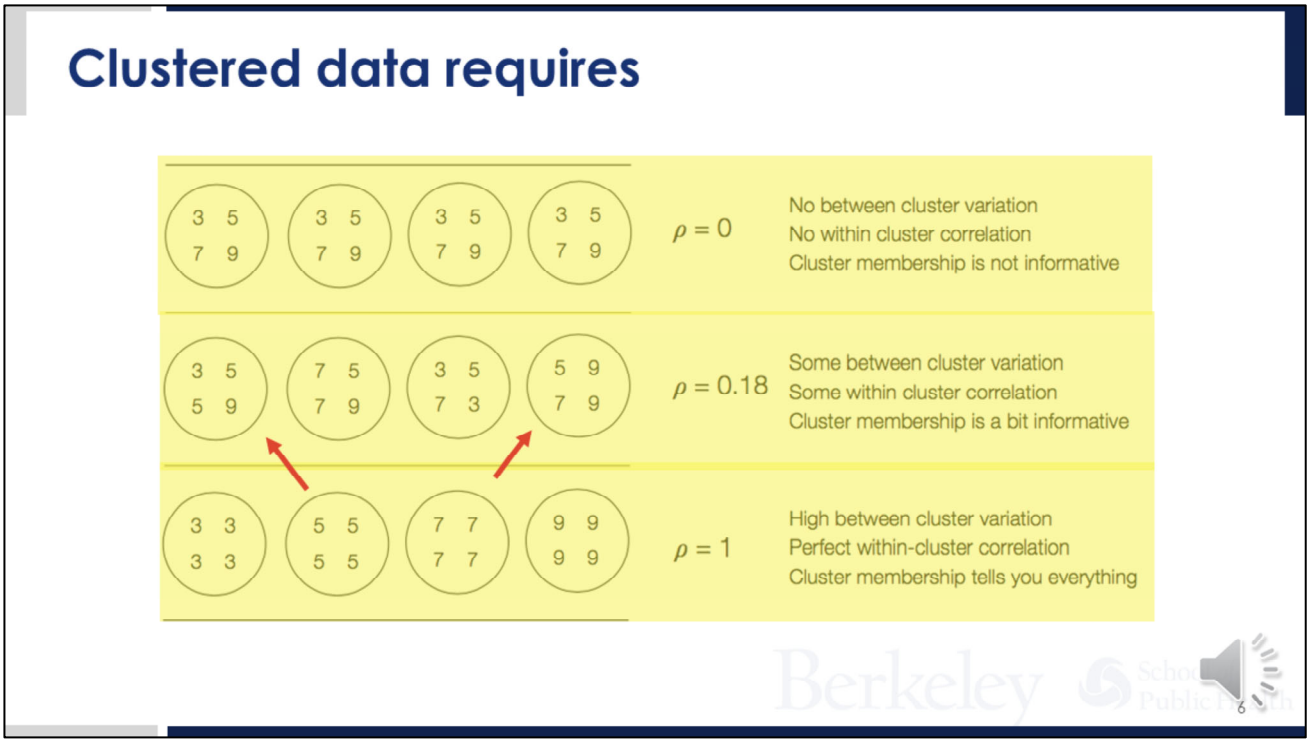
Here are a few examples of applications of cluster randomized trials. *There are community intervention trials, for example, a trial that provides water filtration at the community level and measures its effect on infant health. *In family practice research, providers could use a screening technique on a group of people and assess its impact on patient cholesterol. *In infectious disease research, entire villages could receive mosquito nets to prevent the transmission of malaria. Other villages could be assigned to control, and those villages could be compared to assess the reduced transmission of malaria when all households use bednets. *And in developing country settings, vitamin supplements could be delivered at the community level to assess pregnancy-related mortality.



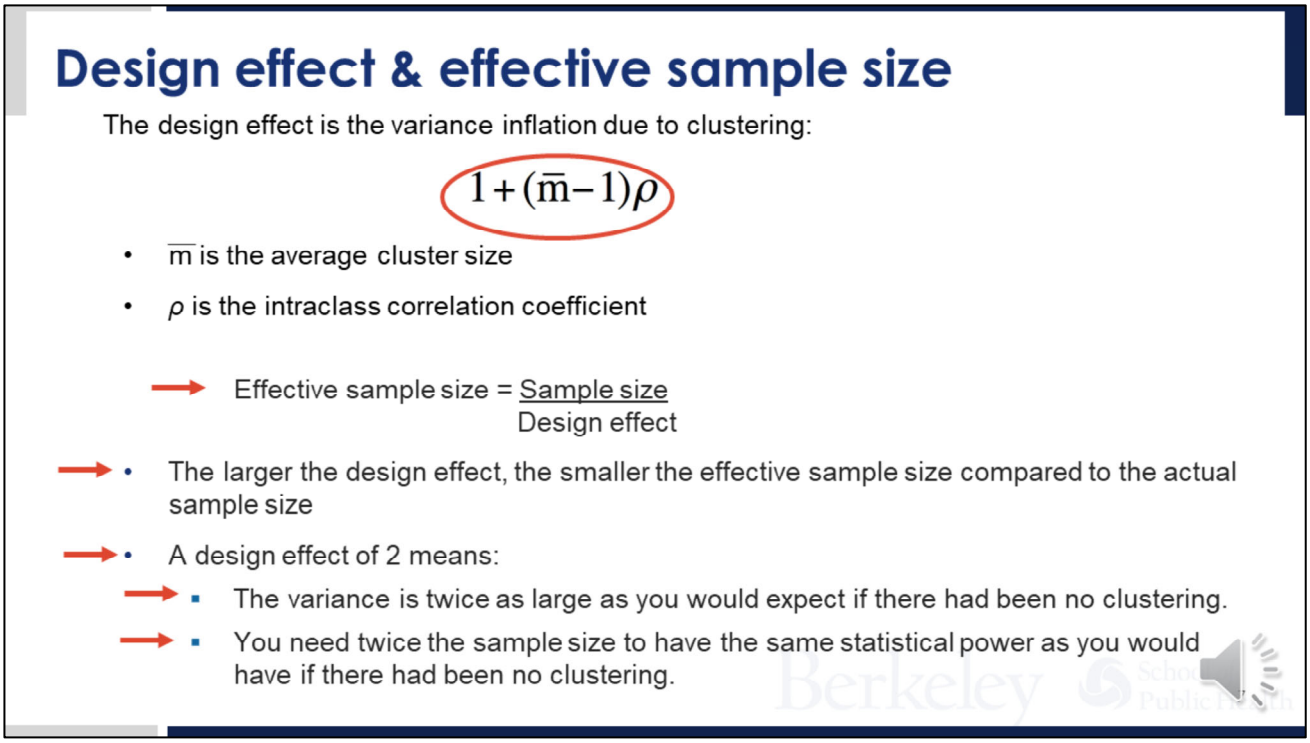
By randomizing groups of people to clusters, we end up having clustered data, and as a result, there are *two statistical requirements unique to cluster-randomized trials or any kind of design that uses clustered data. *The first is that a larger sample size is needed compared to in an individually-randomized study, *and the second is that special analytic methods must be used for clustered data.* So why is this? Well, the intuition is that *people within clusters tend to be more *similar to each other than people in *different clusters. So if we think, for example, about children who go to different schools. Often children who go to the same school and they come from the same community, or there are certain behaviors or cultural practices within certain schools that make children within the same school more similar to each other than to children in other schools. *Another way of saying this is *within cluster variation is typically less than *between cluster variation. There's a bigger difference in people in different clusters than among those within clusters. *Statistically how we'd say this is the outcomes of individuals in the same cluster are correlated, and so they are dependent rather than independent. *However, standard statistical methods assume that outcomes are independent from each other. And so if these methods are used on cluster data they will underestimate variation and standard errors, which will carry forward to have an underestimated or artificially narrow confidence interval.



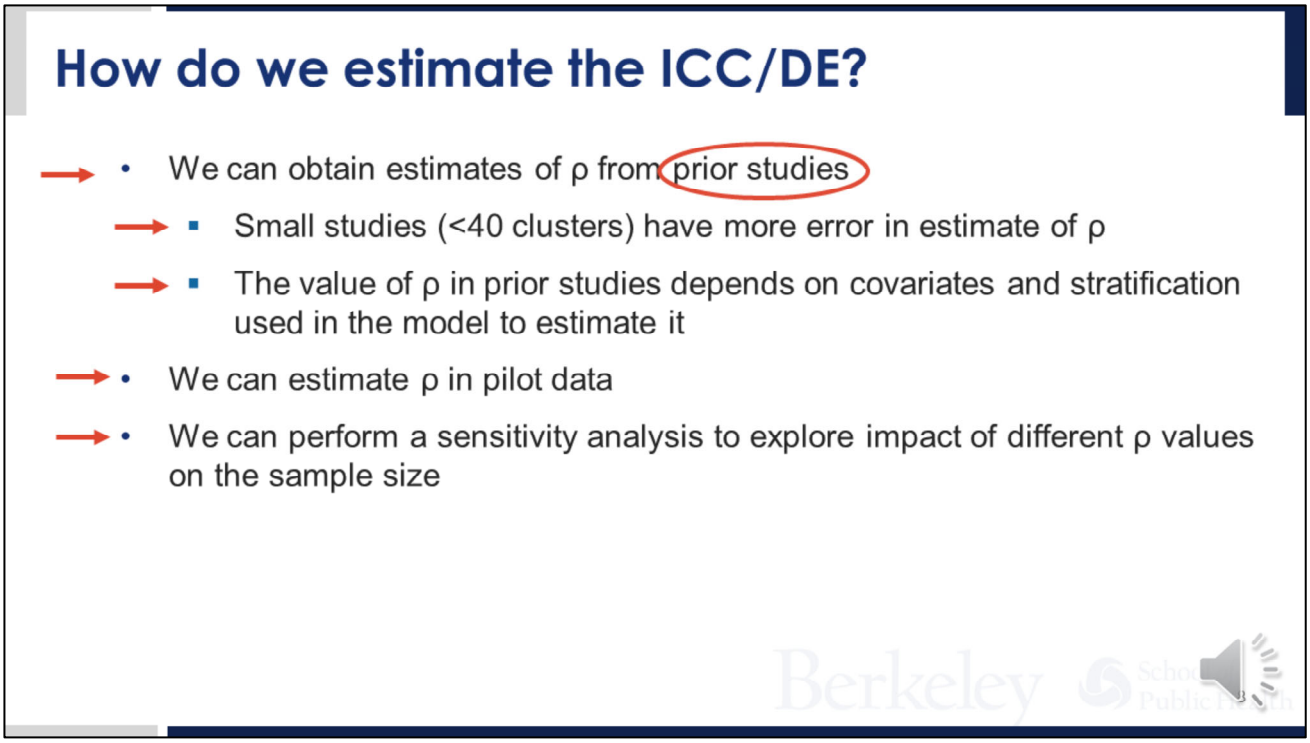
We can measure the extent of clustering using something called the *intracluster correlation coefficient, which is often denoted with the Greek letter rho. Rho is defined as *tao-squared, the between cluster component of the variance, divided by the sum of tao-squared and *sigma squared sub e, which is the within cluster component of variance. It's essentially the fraction of the total variance attributable to the between-cluster component of variance. *Rho can vary from 0 to 1, *and a rho value close to 1 implies that group membership explains a lot of variability in the data.* A small rho value implies that within cluster variance is much greater than between cluster variance. *And the magnitude of rho influences our sample size calculations.



Here's three different examples of a hypothetical study that has four clusters with different levels of rho. *So in the top row, we have an example where rho is equal to 0. The people in each cluster are the same. Three, five, seven, and nine in each cluster, exactly the same pattern and numbers. That means there's no variation between clusters. There's no correlation within clusters. Essentially, knowing the cluster membership is not informative. We could remove the circles around these numbers and have the same information that we had before, rho is equal to 0. *Now let's look at the bottom, where rho is equal to 1. In this example, we see that the values are different in each cluster. So there's a high between cluster variation and a perfect within cluster correlation, meaning the values are the exact same within each circle. So cluster membership tells you everything when rho is 1. Each additional person that you have data on inside a cluster is not really adding anything new, right? If the first circle in the bottom row has another person with a value of 3, we already knew they would have a value of 3. So intuitively you can see how when rho is equal to 1, we actually need a lot of different clusters to measure things about our data. We need a lot of different clusters to have a good amount of information about our data because essentially each cluster is almost like one person. *In the middle row, rho is equal to 0.18. The numbers are similar but do vary between the clusters. There's some between cluster variation, there's some within cluster correlation, but it's not perfect in either direction. So cluster membership is a bit informative. *On the right hand side, we see that that cluster has higher values, *and on the left hand side it has lower values. But it's not going to tell us everything about the values within the cluster just by knowing which cluster we're talking about



We can use rho to estimate something called *the design effect as well as the effective sample size. The design effect is the variance inflation due to clustering. It's equal to 1 plus the average cluster size minus 1 times rho. This is a really useful metric. *If we take the sample size that we have with clustered data, and we divide by the design effect, it gives us our effective sample size. *The larger the design effect, the smaller the effective sample size compared to the actual sample size. *So for example, if our design effect is equal to 2, *the variance we would obtain from our clustered data is twice as large as we would expect if there had been no clustering. *In other words, you would need twice the sample size to have the same statistical power as you would have had if there had been no clustering.

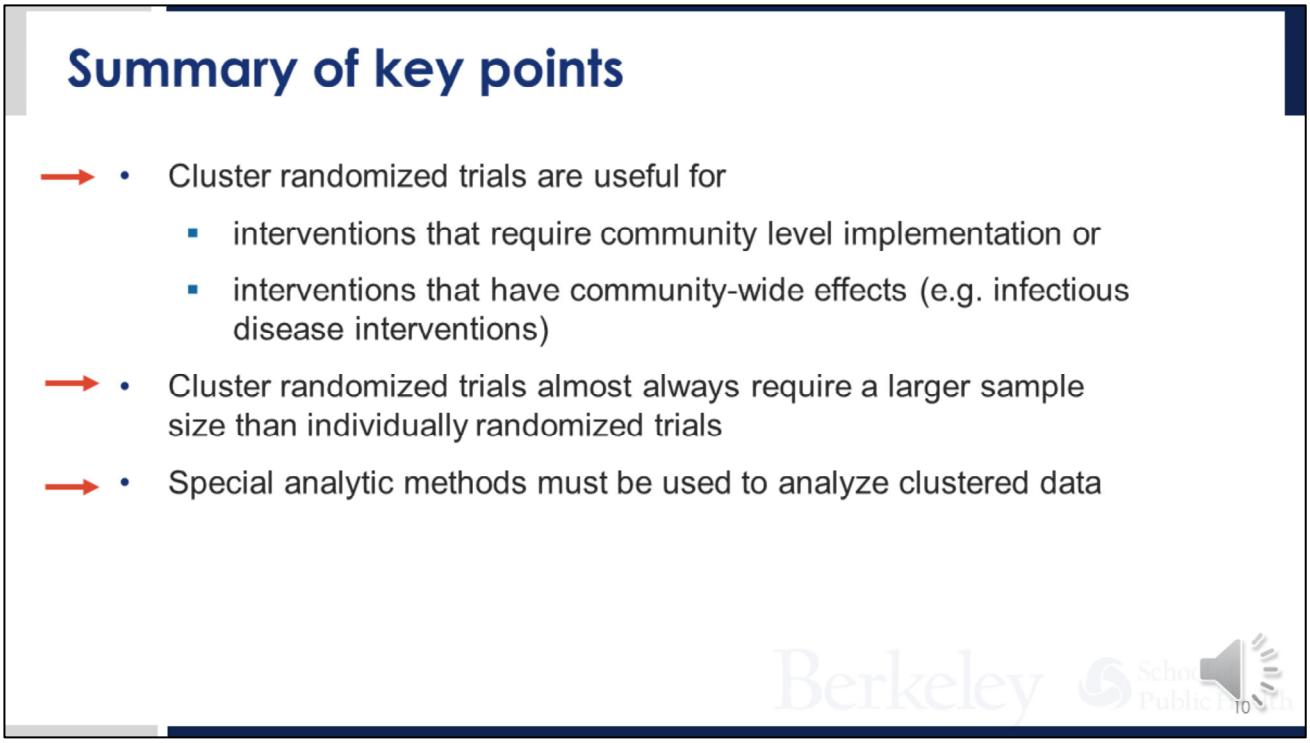


*To estimate the ICC, or the design effect, we need to either use information available from *prior studies in the published literature or we estimate it in data that we have. Talking about prior studies first, it's important to try to find studies that have estimated rho for the same outcome in a similar population and with a decent sample size. *If the number of clusters is less than 40, we're typically concerned that the estimate of rho may be inaccurate. *Also, the value of rho in prior studies will depend on the covariates and the stratification used in the model to estimate rho, so ideally, we want to find a study that used a modeling approach as similar as possible to the one in our study. *Ideally, we would have pilot data where we, for example, conducted a small cluster-randomized trial and estimated rho in that study. Or we could have conducted an observational study, like a cross-sectional study, prior to a cluster-randomized trial, using clustered sampling. We could use that data to then estimate rho in the same population where we'll be doing our research for the randomized trial. *It's always a good idea to do something called a sensitivity analysis, where you explore the impact of different Roe values on the sample size. So at the time you conduct your sample size calculations before the cluster-randomized trial begins, you may have an estimate of rho in your pilot data or from prior research, but we can never be sure that it's really the true rho that we'll see in our study population. So it's good to try repeating the same sample size calculation, just changing the values of rho, to see how much the level of clustering affects our required sample size and to be conservative in choosing a larger sample size in case rho is a different value than the one we used in the original sample size calculation.

Practical implications of the design effect

- Investigators of cluster randomized trials must complete statistical power calculations in advance that use estimates of the ICC and design effect to obtain the proper sample size
  - In most cases, adding more clusters is more important than adding more people per cluster
- Statistical analysis methods must appropriately account for clustering in the data from cluster randomized trials, otherwise standard errors and confidence intervals will be artificially too narrow, suggesting results are more precise than they truly are.

So practically, how does the design effect impact our research? *Well, it's important to always complete a power calculation in advance of starting a trial using estimates of the ICC and the design effect to obtain a proper sample size, and *almost always, it means that we need to add more clusters rather than add more people per cluster. And it just gets back to that slide I showed a few slides back where we look at the impact of the differing values of rho and how adding more people to a cluster with a high rho is really adding very little information, unique information at least. *In addition, during the statistical analysis, we need to use methods that appropriately account for clustering in the data. In other words, these methods can't make assumptions that the outcomes are independent from each other in our data set. Otherwise, if we don't use the appropriate methods, standard errors and confidence intervals will be artificially too small and too narrow, and that will make our results look more precise than they truly are.



To summarize, *cluster-randomized trials are useful for interventions that require community level implementation or interventions that have community-wide effects, like for infectious disease interventions. *They almost always require a larger sample size than individually randomized trials, *and they also require special analytic methods that must be used to analyze clustered data.

reviews, meta-analyses and clinical practice guidelines. The seventh chapter addresses the role of RCTs in health care decisions, their relationship with other types of information, values, preferences and circumstances. This chapter also introduces the basic principles of evidence based decision-making and highlights its strengths and limitations. The last chapter describes 'my wish list'. In it, I highlight what I think are the most important barriers to the optimal use of RCTs in health care, and propose some strategies that could be used to overcome them.

Writing each of the chapters of this book was an extraordinary experience, full of challenges and lessons. The main challenge was to ensure that I could complete the book without affecting my family life or my responsibilities as a new faculty member. The only way in which I could do this was by modifying my sleep pattern. This book was written, essentially, from 11 PM to 2 AM or from 4 AM to 6 AM. Working at these hours gave me more time than I had had in years to think about trials and to put together my thoughts without interruptions. Putting each chapter together forced me to be concise and to use easy-to-read language to describe even the most complex aspects of RCTs. I made every effort to describe the key elements and implications of RCTs without statistics, and in a way that would appeal to busy readers. Each chapter created new questions and opened new avenues for me to explore. During each session I struggled, continuously, to keep the balance between the need to describe the basic elements of RCTs and the urge to express my concerns and expectations about health care research as a whole.

In sum, I did my best to provide you with an enjoyable, balanced, useful and broad view of RCTs and their role in health care. I hope I succeeded.

Alejandro (Alex) R. Jadad
Dundas, 11 June 1998

CHAPTER 1
Randomized controlled trials: the basics

What is a randomized controlled trial?

The randomized controlled trial (RCT) is one of the simplest, most powerful, and revolutionary tools of research.[1,2] In essence, the RCT is a study in which people are allocated 'at random' to receive one of several interventions.

The people who take part in RCTs (the *'study population'*) are called *'participants'* or, irritatingly to some people, *'subjects'*. Participants do not have to be patients, as a study can be conducted in healthy volunteers, in relatives of patients, in members of the general public, in communities, or institutions. The people who design and carry out the study and analyze the results are called the *'investigators'*. The interventions are sometimes called *'clinical maneuvers'*, and include varied actions such as preventive strategies, diagnostic tests, screening programs, and treatments. For instance, if we are conducting a study in which patients with rheumatoid arthritis who are randomized to receive either ibuprofen or a new drug (let us call it 'perfectafen') for the relief of pain, we and our colleagues would be the investigators; the participants the patients with rheumatoid arthritis; and the interventions ibuprofen and perfectafen.

Typically, RCTs seek to measure and compare different events called *'outcomes'* that are present or absent after the participants receive the interventions. Because the outcomes are quantified (or measured), RCTs are regarded as *'quantitative'* studies. In our hypothetical RCT comparing ibuprofen and perfectafen, for instance, the investigators could select pain as the main outcome, measuring it in terms of the number of patients who achieve complete relief 1 week after starting treatment.

Because RCTs are used to compare two or more interventions, they are considered *'comparative'* studies. Usually, one of the interventions is regarded as a standard of comparison or *'control'*, and the

group of participants who receive it is called the *'control group'*. This is why RCTs are referred to as randomized *'controlled'* trials. The control can be conventional practice, a placebo, or no intervention at all. The other groups are called the *'experimental'* or the *'treatment'* groups. In our example, the experimental group is the group that receives 'perfectafen' (the new treatment) and the control group is the one that receives ibuprofen, the standard treatment. Some trials could compare different doses of the same medication, or different ways of administering the intervention as part of either the experimental or control groups.

RCTs are *'experiments'* because the investigators can influence the number and the type of interventions, as well as the regimen (amount, route, and frequency) with which the interventions are applied to the participants. This is in contrast to other types of studies, called *'observational'*, in which the events are not influenced by the investigators. We describe these, briefly, in Chapter 7.

In summary, RCTs are quantitative comparative controlled experiments in which a group of investigators study two or more interventions in a series of individuals who are randomly *'allocated'* (chosen) to receive them.

What does random allocation mean?

Random allocation means that participants are assigned to one of the study groups by chance alone.[3] The decision as to which group they will be in is not determined or influenced by the investigators, the clinicians, or the study participants.

Despite its simplicity, the principle of randomization is often misunderstood by clinicians, researchers, journal reviewers, and even journal editors. Methods to allocate participants according to date of birth (odd or even years), the number of their hospital records, the date in which they are invited to participate in the study (odd or even days), or alternately into the different study groups should not be regarded as really generating random allocation sequences. Although if no one cheats, these 'non-random' or 'quasi-random' studies could produce well-balanced groups, knowledge of the group to which a participant is destined can affect the decision about whether to enter him or her into the trial. This could bias the results of the whole trial.[4]

What is the purpose of random allocation?

By allocating the participants randomly, the characteristics of the participants are likely to be similar across groups at the start of the

comparison (also called *'baseline'*). By keeping the groups *'balanced at baseline'* (as similar as possible at the beginning of the study) the investigators will be more able to isolate and quantify the impact of the interventions they are studying, while minimizing effects from other factors that could influence the outcomes (these are called *'confounding factors'*).

Either known or unknown factors not related directly to the interventions can influence the outcomes of a study. It is fairly easy to match the groups for possible confounding factors, when we know about them. The groups can be kept balanced without randomization as long as all the possible confounding factors have been measured. For example, if 'perfectafen' is evaluated in a retrospective study, the investigators could select a group of patients who received ibuprofen and who took antacids that would match the proportion of patients who took antacids and received 'perfectafen'. But we cannot match groups for factors about which we are not aware. The value of randomization is that if it is done properly, it reduces the risk of serious imbalance in important unknown as well as known factors that could influence the clinical course of the participants. No other study design allows investigators to balance these unknown factors.

The risk of imbalance among the groups is not abolished completely, even if the allocation is perfectly randomized. There are many types of bias that can influence the composition and characteristics of the study groups, even before a trial begins and long after it is completed. We discuss these biases in Chapter 3.

How can randomization be achieved?

We can generate random sequences of allocation in several different ways. Regardless of the method used, investigators should follow two principles: first, they must define the rules that will govern allocation; and second, they should follow those rules strictly throughout the whole study.

In principle, the simplest methods to generate random sequences of allocation are 'flipping a coin' (for studies with two groups) and 'rolling a die' (for studies with two or more groups), although they are rarely used because they do not leave an audit trail.

Investigators can also use 'random number tables' to generate the sequences. Random number tables contain a series of numbers which occur equally often, and that are arranged in a random (therefore unpredictable) fashion. The numbers usually have two or

more digits. The use of a random number table forces investigators to decide the correspondence between the numbers and the groups (e.g. odd corresponding to group A and even to group B; or numbers from 01 to 33 to group A, from 34 to 66 to group B, and from 67 to 99 to group C). Then they have to select the starting point in the table (i.e. the beginning, the end, or any point in the middle of the table marked by a pencil dropped with the eyes closed) and the direction in which the table will be read (e.g. upward or downward). If the numbers in the table contain more than two digits, the investigators have to select the position of the numbers that will determine allocation. For example, if the table contains numbers with four digits (e.g. 2314, 5781, 6703, 8092), the investigators can choose, for example, the last two digits, or the first two, or the first and third. The crucial point is to first define the procedure, and then, once the procedure is defined, do not modify it at any point during the study.

A similar set of numbers may be generated by a computer that is programmed to do so, or by most scientific calculators. The procedures and rules that the investigators must follow are identical to those described for the random number tables.

Regardless of the method the investigators use to generate random sequences of allocation, the number and characteristics of the participants allocated to each of the study groups will probably differ (although slightly) at any given point during the study.[3] To minimize these differences, investigators can use some strategies known as '*restricted* (or *block*) randomization', or '*stratified randomization*'.

Restricted randomization is used to keep the numbers of participants in all the study groups as close as possible. It is achieved by creating 'blocks' of sequences that will ensure that the same number of participants will be allocated to the study groups within each block. For example, in a study with three groups (A, B, and C), the investigators can create six blocks: ABC, ACB, BAC, BCA, CAB, and CBA.

Stratified randomization is used to keep the 'characteristics' of the participants (e.g. age, weight, or functional status) as similar as possible across the study groups. To achieve this, investigators must first identify factors (or 'strata') that are known to be related to the outcome of the study. Once these factors are identified, the next step is to produce a separate block randomization scheme for each factor to ensure that the groups are balanced within each stratum.

On occasion, investigators may not desire the same number of participants in each of the study groups and can decide to allocate

unequal numbers to each group, while preserving the homogeneity of the distribution of the characteristics of the participants across the study groups. This is called '*weighted*' or '*unequal*' randomization. This type of randomization tends to be used by investigators who wish to expose fewer participants to the experimental group because of concerns about unexpected adverse events. In the example of ibuprofen versus perfectafen, the investigators may decide to allocate one patient to perfectafen for every four patients who receive ibuprofen.

Unfortunately, the methods of allocation in studies described as 'randomized' are sometimes poorly reported, and sometimes not reported at all, even when such studies are published in prominent journals.[5,6] Because of these poor descriptions, it is not possible to determine, on most occasions, whether the investigators used a proper method to generate random sequences of allocation. Also, even when the reports of studies described as randomized provide details of the methods of allocation, it has been shown that 5%–10% do not use methods that generate random sequences.[7,8] The reporting of randomization and other aspects of RCTs will be discussed in detail in Chapter 5.

What can be randomized in RCTs?

The most frequent unit of allocation in RCTs is individual people, either patients (the commonest) or caregivers (e.g. treating physicians or nurses). But other units can equally well be randomized to answer specific questions.

Sometimes it is more appropriate to randomize groups of people rather than individuals. This is known as '*cluster*' randomization. Examples of these clusters are hospitals, families, and geographic areas. Investigators frequently use this approach when the RCTs are designed to evaluate interventions that may affect more than one individual within a particular group (e.g. RCTs evaluating the effect of a videotape on smoking cessation on prison inmates, or the effects on parents following a policy of early discharge from hospital after childbirth). It is also used when the way in which the participants in one study group are treated or assessed is likely to modify the treatment or assessment of participants in other groups. This phenomenon is known as '*contamination*'. For example, contamination would be present in an RCT comparing the use of a booklet describing strategies to increase patient participation in treatment decisions versus conventional practice, if patients who have received the booklet shared it with patients who did not.

In other cases, investigators may decide to randomize not only individuals or groups of individuals, but also the order in which the measurements are obtained from each participant. For instance, in an RCT evaluating the effects of morphine on cancer pain, the investigators could randomize the order in which analgesia, adverse effects, and quality of life are assessed.

When are randomized trials needed?

Randomized trials are needed to determine the effects of a health care intervention when these effects are not absolutely clear from observational studies. The effects of some health care interventions, such as antibiotics for pneumonia, or Cesarean section for an obstructed labor, are so dramatic that no further testing is required. More often the effects are less dramatic and may be highly influenced by external factors. Small to moderate effects of interventions can be very important, if the health problem is serious or common.

How are RCTs used?

When reading a trial protocol or a report, it is always wise to consider the purpose of the trial. The theoretical purpose of an RCT is to promote health through a better understanding of the benefits or harms of one or more interventions. A well-conceived, well-performed RCT can inform, enhance, and sometimes change clinical practice or policy. Trials can help individual clinicians to guide their practice, and clinical communities to determine or modify practice patterns. They can provide patients and the public with the information to help them choose what they feel to be the best for them as individuals. Government agencies utilize RCTs for approval of drugs or devices. Insurance agencies, private or government, use them to determine which services or procedures warrant insurance coverage. Institutions can use them to make health policy decisions.

RCTs can, of course, also be used for other purposes. They may be carried out for career advancement or purely for curiosity. They may be funded by companies (most often pharmaceutical, but increasingly also the manufacturing of devices) for regulatory and marketing purposes. They also serve as a powerful form of rhetoric to convince skeptics and doubters, or to control trends that could be considered as too expensive or too disruptive.

How are trials managed and overseen?

Major attention is usually given to when and how RCTs are conceived, designed, and analyzed. All too often, however, too little attention is paid to the actual ongoing meticulous management and oversight of a clinical trial.

Ideally, all activities within a trial must be guided by a '*protocol*', a document that outlines the research question, the rationale for the trial, and the systems that must be set up for recruitment of participants, randomization, data entry, filing, and analysis. These must be clearly established and understood by everyone concerned.

Trials are conducted by research teams led by someone known as the '*principal investigator*', a person who is able to command the respect of fellow collaborators, other clinicians, and the rest of the trial management team. A key member of this team is the '*trial coordinator*', the person responsible for the day-to-day management of the trial and who must be able to respond to the problems that inevitably arise. In addition to the principal investigator (usually known as the 'PI') and the coordinator, the team often includes research assistants, statisticians, data managers, administrative staff, and, increasingly, computer programmers. This team is responsible for ensuring the highest possible levels of quality during patient recruitment, data collection and analysis, and knowledge transfer.[9]

Collecting information and entering it on a computer is relatively simple. Ensuring that the data are valid and sensible is a complicated and detailed process. This often requires lateral thinking, flexibility, good communication, and a great deal of common sense.

The Internet is now playing a larger role in the management of trials, challenging the traditional roles of (and even the need for) each of the members of the management team. An increasing number of tools now allow posting of protocols on the World Wide Web, self-matching by potential trial participants, automatic computer-generated randomization codes, data entry and analysis, results reporting, and audit. Many of these tools are driven by commercial interests and are undergoing rapid transformation under the impetus for market dominance. Governments and academic groups as well are also starting to support the use of online tools. One of the main challenges in the foreseeable future will be to achieve standardized ways to handle each of the components

of a trial online, to promote economies of scale, and efficient and equitable access and exchange of knowledge worldwide.

Can RCTs answer all questions related to health care interventions?

Although RCTs are considered 'the best of all research designs'[10] or 'the most powerful tool in modern clinical research'[11] they are by no means a panacea to answer all health care questions. There are many situations in which they are not feasible, necessary, appropriate, or sufficient to help solve important problems.

The term '*intervention*' is widely used in health care, but infrequently defined. On most occasions the term intervention refers to treatment. However, as we discussed at the beginning of this chapter, this term can be, and often is, used in a much wider sense, to include any health care element offered to the study participants that may have an effect on their health status. Examples include preventive strategies, screening programs, diagnostic tests, the setting in which health care is provided, or educational models. Some of these may be difficult or impossible to study with the methodology of an RCT.

Even when RCT evidence is available, it may not be sufficient to provide all the answers that clinicians, patients, or policy makers need.[12,13] In these cases, we may either require further trials, or use other types of studies to complement the information provided by available RCTs. We discuss other study designs and other types of information, with their advantages and disadvantages, in Chapter 7.

There are many questions for which RCTs are not appropriate. These are usually related to aspects of health care that cannot or should not be influenced by the investigators, such as issues related to the etiology, natural history of diseases, or when the outcomes of interest are adverse effects. It would be unethical and wrong, for instance, to design an RCT in which people would be randomized to smoke or not for decades to compare the prevalence of lung cancer between smokers and non-smokers.

In other circumstances, RCTs may not be worthwhile because of financial constraints, low compliance rates or high drop out rates, or long intervals between the interventions and the outcomes. It would not be possible to carry out an RCT to evaluate the effects of an intervention with very rare outcomes or with effects that take long periods of time to develop. In these cases, other study designs such as case–control studies or cohort studies are more appropriate.

Most RCTs focus on clinical questions and management of disease. Many of the major determinants of health or illness, such as absolute or relative poverty, social class, literacy, transportation or other infrastructure, are not amenable to medical interventions. RCTs can only answer questions for which quantitative results are applicable. A research focus on the types of problems that can be addressed by RCTs can divert our attention and resources from other, equally important health-related problems. Many things that really count cannot be counted.

It follows that before we start reading an RCT, or even searching for one, we should take into account that there are other study designs that may be more appropriate to answer our particular questions. In addition, one RCT in isolation, even when it is appropriate and perfectly designed, is unlikely to provide all the answers we need. We should consider the information provided by a single RCT as an important piece in a puzzle with many empty spaces. This information will have to be assessed and used in conjunction with other types of information (e.g. data from other RCTs or from other study designs, and our own experience), and the values and preferences of the people involved in the decisions, depending on the circumstances in which the decisions are being made.

Our musings

It is very difficult to convey, at the same time, the strengths of RCTs, the value that they have, the risks of over-reliance on them, or their abuse. These concepts swing back and forth as a pendulum. As one of the pioneers of controlled trials, Sir Austin Bradford Hill, put it 'when we think that RCTs can provide all the answers, that doesn't mean just that the pendulum has swung too far, but that it has swung completely off the hook'.[14]

Following the celebration of the 50th anniversary of modern trials, several articles drew attention to the way in which these powerful tools had been hijacked by special interest groups, reducing their ability to provide valid, precise, and relevant answers to important questions.[2,15,16] Since then, these warning calls have been reinforced by highly visible examples of misconduct among funders, policy makers, and researchers, as well as by articles and books by former editors of prominent journals about the current levels of corruption and unethical behavior that exists within the research engine that fuels the drug regulatory process.[17–19]

We now feel that there is an increasing polarization of views about RCTs, along a spectrum of views that ranges from those who put trials on the pedestal of the hierarchy of evidence to those who consider RCTs a dangerous distraction. At the dawn of the 21st century, as the complexity of most health care issues increases[20,21] we have come to realize that trials are valuable sources of knowledge, but not always the most important or even trustworthy ones. One of our greatest challenges will be to learn not only how to carry out scientifically sound and morally ethical RCTs, but why and when to do them.

References

1. Silverman WA, Chalmers I. Sir Austin Bradford Hill: an appreciation. *Controlled Clinical Trials* 1992;13:100–105.
2. Jadad AR, Rennie D. The randomized controlled trial gets a middle-aged checkup. *Journal of American Medical Association* 1998;279(4):319–320.
3. Altman DG. *Practical Statistics for Medical Research*, 2nd edition. London: Chapman & Hall, 2006.
4. Schulz KF. Subverting randomization in controlled trials. *Journal of American Medical Association* 1995;274:1456–1458.
5. Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. *Lancet* 1990;335:149–153.
6. Moher D, Fortin P, Jadad AR, Jüni P, Klassen T, Le Lorier J, Liberati A, Linde K, Penna A. Completeness of reporting of trials in languages other than English: implications for the conduct and reporting of systematic reviews. *Lancet* 1996;347:363–366.
7. Mosteller F, Gilbert JP, McPeek B. Reporting standards and research strategies for controlled trials: agenda for the editor. *Controlled Clinical Trials* 1980;1:37–58.
8. Evans M, Pollock AV. Trials on trial: a review of trials of antibiotic prophylaxis. *Archives of Surgery* 1984;119:109–113.
9. Farrell B. Efficient management of randomized trials: nature or nurture. *British Medical Journal* 2006;317:1236–1239.
10. Norman GR, Streiner DL. *Biostatistics: The Bare Essentials*. Hamilton: Decker, 2000. Smith R. The trouble with medical journals. London: Royal Society of Medicine Press, 2007.
11. Silverman WA. Gnosis and random allotment. *Controlled Clinical Trials* 1981;2:161–166.
12. Naylor CD. Grey zones of clinical practice: some limits to evidence-based medicine. *Lancet* 1995;345:840–842.
13. Freemantle N. Dealing with uncertainty: will science solve the problem of resource allocation in the UK NHS? *Social Sciences and Medicine* 1995;40:1365–1370.

14. Hill AB. Heberden Oration 1965: reflections on the controlled trial. *Annals of Rheumatic Diseases* 1965;25:107–113.
15. Horton R. The clinical trial: deceitful, disputable, unbelievable, unhelpful, and shameful – what next? *Controlled Clinical Trials* 2001;22:593–604.
16. Smith R. Medical journals and pharmaceutical companies: uneasy bedfellows. *British Medical Journal* 2003;326:1202–1205.
17. Angell M. *The Truth About the Drug Companies: How They Deceive Us and What to Do about It*, 1st edition. New York: Random House. 2004.
18. Kassirer JP. *On the Take: How Medicine's Complicity with Big Business Can Endanger Your Health*. USA: Oxford University Press, 2005.
19. Smith R. *The Trouble with Medical Journals*. Royal Society of Medicine Press. 2007.
20. Plsek PE, Greenhalgh T. The challenge of complexity in health care. *British Medical Journal* 2001;323:625–628.
21. www.healthandeverything.org (accessed December 22, 2006).

CHAPTER 2
Types of randomized controlled trials

Randomized controlled trials (RCTs) can be used to evaluate different types of interventions in different populations of participants, in different settings, and for different purposes. Once investigators ensure that allocation of participants to the study groups is random (necessary to call the study an RCT), they can design the study using strategies to match the characteristics of the interventions they want to study, the resources they have available, and their academic, political, marketing, or clinical motivations. Over the years, various jargon terms have been used to describe different types of RCTs. There is no single source with clear and simple definitions for all these terms, so this jargon may be difficult to understand for those who are starting their careers as clinicians or researchers.

In this chapter, we will describe the terms most frequently used to describe different types of RCTs, and do our best to classify them in a way that will be easy to follow, understand, and remember. Some of the terms apply specifically to RCTs, while others may also be applied to other study designs as well. Some terms are mutually exclusive, some overlap considerably, and some complement each other.

RCTs can be classified according to the aspects of the interventions they evaluate, the way in which the participants are exposed to the interventions, the units of analysis, the number of participants included in the study, whether the investigators and participants know which intervention is being assessed, and whether non-randomized individuals and participants' preferences are taken into account in the design of the study (Table 2.1).

RCTs that explore different aspects of the interventions

Depending on the aspects of the interventions that investigators want to evaluate, RCTs can be classified as efficacy, effectiveness, or equivalence trials; and as phase I, II, III, or IV trials.

Table 2.1 Types of RCTs

RCTs that explore different aspects of the interventions they evaluate
- Efficacy and effectiveness trials
- Equivalence trials
- Phase I, II, III, and IV trials

RCTs according to how the participants are exposed to the interventions
- Parallel trials
- Factorial design trials
- Cross-over trials

RCTs by unit of analysis
- Body part
- Individual
- Group (Institution, community, geographic, and location)

RCTs according to the number of participants
- Fixed to variable sample size
- N-of-1 trials to mega-trials

RCTs according to whether the investigators and participants know which intervention is being assessed
- Open trials
- Blinded (masked) trials

RCTs according to whether non-randomized individuals and participants' preferences are taken into account
- Zelen's design
- Comprehensive cohort design
- Wennberg's design

What is the difference between an efficacy and an effectiveness trial?

RCTs are often described in terms of whether they evaluate the *efficacy* or the *effectiveness* of an intervention. These two concepts are sometimes confused because of the similarity of their names.

The term '*efficacy*' refers to whether an intervention works in people who actually receive it.[1] An *efficacy trial* (sometimes referred to as an explanatory trial) aims to address the question of whether or not an intervention *can* work under optimal circumstances, and how. An efficacy trial is designed in such a way that the results are likely to yield a 'clean' evaluation of the intervention. To achieve this, the investigators set strict inclusion criteria that will produce highly homogeneous study groups. For instance, for an efficacy trial of the

effects of a new anti-hypertensive drug they could decide to include only patients between 40 and 50 years of age, with no co-existing diseases. (A hypertensive patient who also had, for example, diabetes would not be eligible for inclusion.) They would also exclude those receiving other relevant interventions (such as beta-blockers).

The investigators will try to include only participants who will follow their instructions and who will actually receive the intervention. The extent to which study participants follow the instructions given by the investigators is called *compliance* or *adherence*. High compliance is easy to achieve when the administration of the interventions can be completely controlled by the investigators or by health professionals who are supportive of the study (e.g. an RCT comparing the effects of coronary artery bypass surgery and those of angioplasty in patients with unstable angina). Compliance is easier to achieve for trials of short duration. It is more difficult to achieve when the interventions are administered by the participants themselves, when the study has a long duration, and when the interventions have to be administered several times a day. Returning to the example of the anti-hypertensive drug previously discussed, compliance will depend on the extent to which the participants take the anti-hypertensive tablets as prescribed for the whole duration of the study. To make high compliance more likely, the investigators may choose to include only patients who have already shown high compliance in other studies or in preliminary tests.

Efficacy trials also tend to use placebos as controls, fixed regimens (such as 20 mg by mouth every 6 hours), long washout periods (if patients have been taking diuretics, for instance, those drugs will be stopped for a period of time long enough to ensure that they are 'washed out' of their bodies), intention-to-treat analysis (see Chapter 3), and focus on 'hard' outcomes (i.e. blood pressure recorded at specific times following a detailed and standardized process).

The term '*effectiveness*' refers to whether an intervention *does* work, rather than whether it *can* work. An *effectiveness trial* (sometimes called a *pragmatic* or *management* trial) typically evaluates an intervention with proven efficacy when it is offered to a heterogeneous group of people under ordinary clinical circumstances.[2] Effectiveness trials are designed to determine not only whether the intervention achieves specific outcomes, but also to describe all the consequences of its use, good and bad, for people to whom it has been offered, under circumstances mimicking clinical practice. To achieve this, effectiveness studies tend to use less restrictive inclusion criteria,

and include participants with heterogeneous characteristics. They tend to use active controls (e.g. the new anti-hypertensive drug versus a beta-blocker) rather than placebo controls, and flexible regimens (e.g. 20 mg orally every 6 hours, reducing or increasing the dose by 5 mg according to the degree of blood pressure control and adverse effects). They are analyzed to include all of the patients who were allocated to receive the intervention (see Chapter 3). Effectiveness trials often include the use of 'soft' outcome measures, such as measures of sexual function or quality of life.

Both efficacy and effectiveness approaches are reasonable and complementary. They often cannot be clearly differentiated. The terms represent a spectrum and most RCTs include elements from each. The key issue is whether the investigators achieved the best combination of elements to answer their (and the readers') questions.

What is an equivalence trial?

On occasions, trials are designed not to detect possible differences in efficacy or effectiveness between two or more interventions, but to show that the interventions are, within certain narrow limits, 'equally effective' or 'equally efficacious'.[3] These trials are called *equivalence trials*. Often, they seek to demonstrate that a new intervention (or a cheaper or more conservative one) is at least as good as the conventional standard treatment. Investigators who engage in equivalence trials should make efforts to minimize the risk of suggesting that the interventions have equivalent effects when in fact they do not. Strategies to minimize this type of risk are described in Chapters 3 and 4.

What are phase I, II, III, and IV trials?

These terms are used to describe the different types of trials that are conducted during the evaluation of a new drug. Only phase III trials are actually RCTs. Phase I, II, and IV trials are not randomized. They are, however, important stages in the development and understanding of the effects of an intervention.

As the name suggests, *phase I trials* are the first studies conducted in humans to evaluate a new drug. They are conducted once the safety and potential efficacy of the new drug have been documented in animals. As the investigators know nothing about the effects of the new drug in humans, phase I trials tend to focus on *safety*, rather than on comparative effectiveness. They are used to establish how much of a new drug can be given to humans without causing serious

adverse effects, and to study how the drug is metabolized by the human body. Phase I trials are mostly conducted on *healthy volunteers*. The typical participant in a phase I study may be one of the investigators who developed the new drug, either an employee of a pharmaceutical company or a member of a research team at a university. People with diseases for which there is no known cure (for instance, certain types of advanced cancer) sometimes participate in phase I trials. As mentioned above, these trials are typically *not randomized*, and even *not controlled*. Usually, they are just series of cases in which the participants are given incremental doses of the drug, without a control group, while they are monitored carefully by the investigators.

After the safety of a new drug has been documented in phase I trials, investigators can proceed to conduct *phase II trials*. These are trials in which the new drug is given to small groups of patients with a given condition. The aim of phase II trials is to establish the relative efficacy of different doses and frequencies of administration. Even though phase II trials focus on efficacy, they can also provide additional information on the safety of the new drug.

Often, phase II trials are not randomized, particularly when the therapeutic effects of the new drug can be measured. For instance, if a new drug has been designed to treat a type of cancer that is associated with a high mortality rate, the investigators will conduct a phase II trial in which about 20 patients will receive the drug while tumor response, mortality, and adverse effects are monitored carefully. If the drug is judged to be ineffective or excessively toxic, no more trials will be conducted. However, if the drug produces a good response (i.e. fewer patients than expected die) and patients tolerate its adverse effects, the investigators can proceed to a phase III trial.

Phase III trials are designed and conducted once a new drug has been shown to be reasonably effective and safe in phase II trials.[4] Phase III trials are typically *effectiveness* trials, because they seek to compare the new drug with an existing drug or intervention known to be effective. This existing drug is usually regarded as the current standard treatment.[4] Most phase III trials are RCTs, the subject of this book.

The term *phase IV* trial is used to represent large studies[3] that seek to monitor adverse effects of a new drug after it has been approved for marketing.[4] These studies are also called post-marketing or post-approval surveillance studies. They are mostly surveys and seldom include comparisons among interventions. Phase IV trials can also

be used to bring a new drug to the attention of a large number of clinicians[4] or to expand the number of indications for clinical use.

Physicians and patients both expect that after a drug has been approved for clinical use it will be both safe and effective for the appropriate indications. This expectation may be naïve. Serious adverse effects of drugs may be quite uncommon, and detecting them accurately can be difficult within the context of most RCTs, or with the existing passive, unsystematic, unregulated, uncoordinated, and non-mandatory system for assessing the risk of adverse events of an intervention following its approval by regulatory agencies. Sometimes because of conflicts of interest, pharmaceutical firms or device manufacturers may neglect to promote studies specifically designed to assess the potential harm of their products in the real world, or to fully acknowledge reports that indicate harm. Developers of new and expensive interventions are strongly motivated to protect and expand their markets, and may use highly defensive articles as well as other tactics to do so. Measures to improve post-approval surveillance are at least under consideration, and we hope that they will succeed in penetrating the dense jungle of competing interests.[5]

RCTs according to how the participants are exposed to the interventions

Depending on the way in which the participants are exposed to the study interventions, RCTs can have parallel, factorial, or cross-over designs.

What is a parallel design?

Parallel design studies (also called parallel trials or RCTs with parallel group design) are trials in which each group of participants is exposed only to one of the study interventions. They are the most frequently used design. For instance, in a parallel trial designed to compare the effects of a new analgesic with those of a placebo in patients with migraine, the investigators would give the new analgesic to one group of patients and placebo to a *different* group of patients.

What is a factorial design?

In a factorial design, investigators can compare two or more experimental interventions in combination as well as individually. For example, in a factorial design trial to compare the effects of low-dose aspirin and vitamin E in preventing cardiovascular events in patients

with diabetes,[6] patients were allocated to receive aspirin only, vitamin E only, both aspirin and vitamin E or placebo. This design allows the investigators to compare the experimental interventions with each other, and with a placebo, and also to investigate interactions between the interventions (i.e. the effects of aspirin and vitamin E given separately with the effects of the combination).

What is a cross-over design?

An RCT has a cross-over design when each of the participants is given *all* the study interventions in successive periods. The order in which the participants receive each of the study interventions is determined at random. Cross-over trials produce *within-participant comparisons*, while parallel designs produce between-participant comparisons. As each participant acts as his/her own control, cross-over can produce statistically significant and clinically important results with fewer participants than would be required with a parallel design.[7,8]

The time during which each of the interventions is administered and evaluated is called a *period*. The simplest cross-over design includes only two periods. Returning to the example of the new analgesic, if the same group of investigators uses a cross-over design, they would randomize *each* patient to receive the new analgesic first and then the placebo, or vice versa, the placebo first and then the new analgesic.

For cross-over trials to be useful, the condition under study must be stable and should not be curable during the first period of the trial (the participants would not be able to receive the control).[9] The effect of the intervention must have a short duration and be reversible. When the effects of one intervention are still present during the evaluation of another, such effects are called *carry-over effects*. If any observed difference between the interventions can be explained by the order in which the interventions were given to the participants, this is called *a treatment–period interaction* and it can invalidate the trial. Carry-over effects can be predicted when the duration of the effects of the interventions are well known. In these cases, carry-over effects can be prevented by separating the study periods by a period of time that is long enough to enable the participants to be free of the influence of the intervention previously used by the time they receive the next intervention.[3] This amount of time is also known as *washout period*.

RCTs according to the unit of analysis

The unit of analysis for an RCT is usually the individual, but sometimes it could be a part of the body (e.g. comparing a treatment used on one limb with a different treatment on the other). On other occasions, the unit has to be larger than the individual.

There are times when randomizing individuals is either technically impossible or may compromise the evaluation. Some interventions can only be delivered to groups. The organization of health care in a community will affect all patients in that community. An intervention aimed at health professions, such as a training package, will (or at least might) modify the care of all patients attended by those professionals. However, randomization in groups has disadvantages in terms of sample size requirement; the number of patients needed will have to be larger because the groups, rather than the individuals will be the unit of analysis. Individuals within each group cannot be considered as independent because, for example, the location of the community or the characteristics of a particular practitioner may attract a particular type of patient. This 'clustering' effect increases the number of subjects required.[10,11]

There may be ethical problems involved with randomization of groups of individuals. Making sure that individual participants are informed of the study may be difficult, and individual patients may not have the opportunity to consent to randomization. This may be acceptable if the potential risk of the intervention is very low, but this would depend on the nature of the intervention (see Chapter 8).

RCTs according to the number of participants

RCTs can have fixed or variable (sequential) numbers of participants. In a fixed-size trial the investigators establish, a priori, the number of participants (the *sample size*) that they will include. This number can be decided arbitrarily or can be calculated using statistical methods. The main goal of using statistical methods to calculate the sample size is to maximize the chance of detecting a statistically and clinically significant difference between the interventions when a difference really exists. In other cases, the investigators do not set the sample size at the outset. Instead, the investigators continue recruiting participants until a clear benefit of one of the interventions is observed, or until they are convinced that there

are no important differences between the interventions.[12] These trials allow a more efficient use of resources than trials with fixed numbers of participants, but they depend on the principal outcome being measured relatively soon after trial entry.

How small can a trial be?

Can a trial be done on one person? Surprisingly, perhaps, the answer is yes. These RCTs are called *'n-of-1 trials'* or *'individual patient trials'*. Basically, they are cross-over trials in which one participant receives the experimental and the control interventions in pairs, on multiple occasions, and in random order. Indeed, many researchers consider the n-of-1 trial to be the strongest form of evidence because what we really want to know is whether the treatment will work for *this* person, rather than what its effects are on average. These trials provide results applicable to that individual, rather than generalizable results.

The n-of-1 trials can be very useful when it is not clear whether a treatment will help a particular person. They are particularly applicable for patients with a rare disease when there are no trials supporting the use of the treatment in that particular disease, or for any condition where the treatment has been evaluated in studies that include very different patients.[13] Typically, the number of pairs of interventions varies from 2 to 7. Usually, the number of pairs is not specified in advance, so that the clinician and the patient can decide to stop when they are convinced that there are (or that there are not) important differences between the interventions.

How big can a trial be?

'Mega-trial' is a term that is being used to describe RCTs with a simple design which include thousands of patients and limited data collection.[14,15] Usually, these trials require the participation of many investigators (sometimes hundreds of them) from multiple centers and from different countries. The main purpose of these large simple trials is to obtain 'increased statistical power' and to achieve wider generalizability.

RCTs according to whether the investigators and the participants know which intervention is being assessed

In addition to randomization (which helps control selection bias), a randomized trial can incorporate other methodological strategies

to reduce the risk of other biases. These biases and the strategies to control them will be discussed in detail in Chapter 3, but we have raised the issue in this chapter because the presence, absence, or degree of one of these strategies have been used to classify RCTs. This strategy is known as *'blinding'* or perhaps more appropriately (but less commonly used) *'masking'*.

In clinical trial jargon, blinding, or masking represents any attempt made by the investigators to keep one or more of the people involved in the trial (e.g. the participant or the person assessing the outcomes) unaware of the intervention that is being given or evaluated. The purpose of blinding is to reduce the risk of *ascertainment or observation bias*. This bias is present when the assessment of the outcomes of an intervention is influenced systematically by knowledge of which intervention a participant is receiving.

The term *'double-blind randomized controlled trial'* is so often used to represent the ultimate in design to produce valid results. This jargon term confuses the issue because the important thing is not the number of people who are blinded (or masked) during a trial, but the number and role of those who are *not* blinded. Anyone of the many people who are involved in a trial can distort its results, if they know the identity of the intervention while it is administered or assessed.

Blinding can be implemented in at least six different levels in an RCT. These levels include the participants, the investigators, or clinicians who administer the interventions, the investigators or clinicians who take care of the participants during the trial, the investigators who assess the outcomes of the interventions, the data analysts and the investigators who write the results of the trial. Of course, in many studies those who administer the interventions, take care of the participants, assess the outcomes, or write the reports are the same. Depending on the extent of blinding, RCTs can be classified as *open* (everyone involved in the trial knows what is happening), or *single-blind, double-blind*, triple-blind, quadruple-blind, and so on.

Successful blinding requires that the interventions be indistinguishable. When the experimental intervention is new and there are no standard effective interventions that could be used as control, the investigators could use an inert substance, or *placebo* that would appear identical to the experimental intervention. These trials are known as *placebo controlled*.

When the RCT is designed to compare a new intervention with a standard treatment, the RCTs are called *active controlled*. Achieving successful blinding in active controlled trials is often difficult and frequently requires the use of what is called a *double-dummy*. In a double-dummy RCT, each group of participants receives one of the active interventions and a placebo (in this case called a dummy) that appears identical to the *other* intervention. The double-dummy technique is particularly useful when the investigators want to compare interventions that are administered by different routes or that require different techniques of administration. For instance, a double-dummy RCT would be the ideal study design to compare one intervention that is given as a tablet with another that is given by injection. In such a trial, the participants in one of the study groups would receive a tablet with the active drug and a placebo injection, while the participants in the other group would receive a placebo tablet and an injection with the active drug.

RCTs that take into account non-randomized individuals and participants' preferences

The preferences of an individual (patient or clinician) for one of the interventions to be compared in a trial could strongly influence the outcome.

If people refuse to participate in a trial because of their preference, the resulting study sample would not be representative of the overall target group. If they participate, despite their preference, they might subvert the protocol either through their compliance or at the time of the assessment of the outcomes. For example, bias may occur when patients who are aware of an experimental treatment not available to them outside a trial decide to join the trial, hoping to receive the treatment, but comply poorly if they receive the control intervention.

The outcomes of the individuals who do not participate in the trials or of those who participate and have strong preferences are rarely recorded. *Preference trials* are designed specifically to overcome this limitation.[16] These trials include at least one group in which the participants are allowed to choose their own preferred treatment from among several options offered.

There are at least three types of preference trials: Zelen's design, trials with a comprehensive cohort design, and Wennberg's design (Figure 2.1).

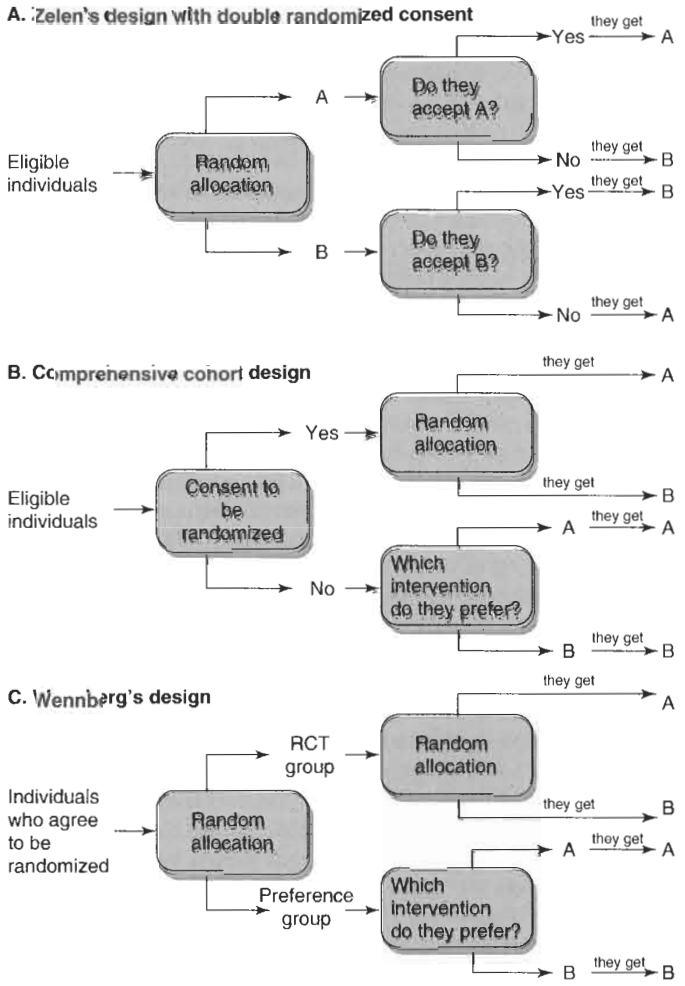


Figure 2.1 Preference trials

What is a trial with Zelen's design?

In a trial with Zelen's design, eligible individuals are randomized to receive either a standard treatment or an experimental intervention *before they give consent* to participate in the trial. Those who are allocated to standard treatment are given the standard treatment

and are not told that they are part of a trial, while those who are allocated to the experimental intervention are offered the experimental intervention and told that they are part of a trial. If they refuse to participate in the trial, they are given the standard intervention but are analyzed as if they had received the experimental intervention.[17]

The main advantages of Zelen's design are that almost all eligible individuals are included in the trial and that the design allows the evaluation of the true effect of *offering* experimental interventions to patients. The main disadvantages are that they have to be open trials and that the statistical power of the study may be affected if a high proportion of participants choose to have the standard treatment.

There are obvious ethical problems in using Zelen's design to randomize patients without their consent, but there may be times when this approach may be kinder to the participants (Chapter 8). To overcome the ethical concerns of not telling patients that they have been randomized to receive the standard treatment, the original approach proposed by Zelen can be modified by informing all participants of the group to which they have been allocated, and by offering them the opportunity to switch to the other group. This design is also known as *double randomized consent design* (Figure 2.1A).

What is a trial with comprehensive cohort design?

A comprehensive cohort trial is a study in which all participants are followed up, regardless of their randomization status (Figure 2.1B). In these trials, if a person agrees to take part in an RCT, she/he is randomized to one of the study interventions. If the person does not agree to be randomized because she/he has a strong preference for one of the interventions that person will be given the preferred intervention and followed up as if she/he was part of a cohort study (Chapter 7).[18,19] At the end, the outcomes of people who participated in the RCT can be compared with those who participated in the cohort studies to assess their similarities and differences.

This type of design is ideal for trials in which a large proportion of eligible individuals are likely to refuse to be randomized because they (or their clinicians) have a strong preference for one of the study interventions.[18] In this case, it could be said that the study is really a prospective cohort study with a small proportion of participants participating in an RCT.

What is a trial with Wennberg's design?

In a trial with Wennberg's design, eligible individuals are randomized to a 'preference group' or to an 'RCT group'[20] (Figure 2.1C). Those individuals in the preference group are given the opportunity to receive the intervention that they choose, while those in the RCT group are allocated randomly to receive any of the study interventions, regardless of their preference. At the end of the study, the outcomes associated with each of the interventions in each of the groups are compared and used to estimate the impact of the participants' preferences on the outcomes.

Preference trials are rarely used in health care research, although they may be more frequently used as consumer participation in health care decisions and research increases.[21] We believe that these study designs are strong, useful, and should be used far more often.

Our musings

With few exceptions, since their introduction into clinical research, trials have followed a simple, basic, and linear recipe: conceive of an intervention that may help cure or alleviate a health problem in a particular population; formulate a question; use chance allocation to divide a sample of the population into two or more groups; use the proposed intervention for one of the groups; and compare the outcomes for that group with the outcomes for the other group or groups who did not receive the intervention. It is a good, robust, recipe that produced a nourishing product. It proved eminently successful for at least half a century.

The recipe is nourishing, but perhaps not delectable. The ingredients are good, but the basic menu may not satisfy the modern palate. Perhaps it is time to look for more varied, more acceptable, and more digestible fare. Recipes that were new and exciting in the middle of the 20th century may not be enough as the new century dawns. Although there have been some innovations in trial design, these have mainly been a tinkering with, rather than a radical look at new possibilities.

In essence, the RCT has followed a linear, cause-and-effect epistemology, an assumption that other things being equal, any differences found between the groups compared will be due to chance, and that they will strictly follow the laws of statistical probability. Sometimes, when problems are fairly simple (such as how to bake a cake, how to treat pneumococcal pneumonia), following a tested

recipe will give us a consistently successful product. Sometimes a problem is more complicated (e.g. how to connect a shuttle with a spaceship in orbit, how to anesthetize a patient for chest surgery), but still if the protocol is followed meticulously we can be reasonably confident of a successful result.

Unfortunately, other things are not always equal. Complex issues (e.g. raising a child, how to treat dementia) do not seem to respond to the straightforward cause-and-effect algorithm. We can do all the right things, but our child grows up to be her own self; dementia follows its own inexorable course. How can traditional trials help us? When we try to address important complex issues related to our health and health care with approaches well suited for simple and complicated problems, they are not likely to work. Being more persistent, doing the same thing over and over again, and using more and more resources in the same way, is not likely to produce different results.

It is not hard to understand the popularity of RCTs as currently conceived. They are very convenient for trialists, for pharmaceutical firms, for funding agencies, for regulatory agencies and publishers. They provide clear paths to follow, from the formulation of the research question, through the recruitment of subjects, collection of data, analysis of the results, to the publication of the findings. They satisfy bureaucratic protocols. The question is how good they still are for our health. The problems that concerned us most in the 20th century were primarily acute problems, ones that could be readily identified and delineated, where the responses to treatments could be readily assessed. More and more of our 21st century problems are complex, chronic, difficult to pinpoint, with long-delayed and uncertain responses to our well-meaning interventions.[22]

The very success to date of RCTs for many problems may be blinding us to the opportunities and need for innovative approaches. The mechanistic approach that was right for the industrial age may even be detrimental for an age in which we must acknowledge that uncertainty and unpredictability are unavoidable, and should be cherished. We believe that it is time for radical new designs for RCTs.

> Why couldn't we give more prominence to patient preference trials, which recognize the effects that choice may have on outcome?

> Why couldn't trials have more democratic designs, in which all who may be affected by the results of a trial are involved in its formulation?
>
> Can we build recursive elements into trials, so that questions be reformulated in response to early findings, or interventions be modified in response to emerging patterns?
>
> Why not investigate qualitative elements concurrently, as integral parts of a randomized trial?
>
> What if we paid more attention to the behavioral aspects of the interventions, and the effects that the trial per se may have on both clinicians and participants?
>
> Is it time to think of the rhetorical aspects of the trial, the effects that they may have on potential users of the information, rather than just the validity of the results?

We know that others will come up with very innovative ideas to further improve RCTs that are more relevant to the needs of today and tomorrow. These ideas may be more difficult to share, they may be resisted, but they may prevail. We hope that they will be given a chance.

It is time for mavericks and academics to leave the assembly line, to recover their role as iconoclasts and innovators. A time to reject bureaucratic mind sets, intolerance of change, and rigid institutional regulations.

References

1. Fletcher RH, Fletcher SW. *Clinical Epidemiology: The Essentials*, 4th edition. Baltimore, MD: Lippincott Williams & Wilkins, 2005.
2. Sackett DL, Gent M. Controversy in counting and attributing events in clinical trials. *New England Journal of Medicine* 1979;301:1410–1412.
3. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research*, 4th edition. Oxford, UK: Blackwell Publishing Professional, 2001.
4. Pocock SJ. *Clinical Trials: A Practical Approach*. Chichester, UK: John Wiley & Sons, 1983.
5. Fontanarosa PB, Rennie D, DeAngelis CD. Postmarketing surveillance – lack of vigilance, lack of trust. *Journal of American Medical Association* 2004; 292:2647–2650.
6. Sacco M, Pellegrini F, Roncaglioni MC, Avanzini F, Tognoni G, Nicolucci A, PPP Collaborative Group. Primary prevention of cardiovascular events with low-dose aspirin and vitamin E in type 2 diabetic patients: results of the Primary Prevention Project (PPP) trial. *Diabetes Care* 2003; 26: 3264–3272.

7. Louis TA, Lavori PW, Bailar III JC, Polansky M. Crossover and self-controlled designs in clinical research. In *Medical Uses of Statistics*, 2nd edition, JC Bailar III and M Frederick (Eds.). Boston: New England Medical Journal Publications, 1992, pp. 83–104.
8. Sibbald B, Roberts C. Understanding controlled trials: crossover trials. *British Medical Journal* 1998;316:1719–1720.
9. Senn S. *Cross-over Trials in Clinical Research*. Chichester, UK: John Wiley & Sons, 2002.
10. Roberts C, Sibbald B. Understanding controlled trials. Randomising groups of patients. *British Medical Journal* 1998;316:1898–1900.
11. Bland JM, Kerry SM. Trials randomized in clusters. *British Medical Journal* 1997;315:600–660.
12. Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall, 1991.
13. Guyatt G, Sackett D, Taylor DW, Chong J, Roberts RS, Pugsley S. Determining optimal therapy – randomized trials in individual patients. *New England Journal of Medicine* 1986;314:889–892.
14. Woods KL. Mega-trials and management of acute myocardial infarction. *Lancet* 1995;346:611–614.
15. Charlton BG. Mega-trials: methodological issues and clinical implications. *Journal of the Royal College of Physicians of London* 1995;29:96–100.
16. King et al. Impact of participant and physician intervention preferences on randomized trials: a systematic review. *Journal of American Medical Association* 2005;293:1089–1099.
17. Zelen M. A new design for randomized clinical trials. *New England Journal of Medicine* 1979;300:1242–1245.
18. Olschewski M, Scheurlen H. Comprehensive cohort study: an alternative to randomized consent design in a breast preservation trial. *Methods of Information in Medicine* 1985;24:131–134.
19. Brewin CR, Bradley C. Patient preferences and randomised clinical trials. *British Medical Journal* 1989;299:684–685.
20. Flood AB, Wennberg JE, Nease Jr RF, Fowler Jr FJ, Ding J, Hynes LM. The importance of patient preference in the decision to screen for prostate cancer. *Journal of General Internal Medicine* 1996;11:342–349.
21. Sackett DL, Wennberg JE. Choosing the best research design for each question. *British Medical Journal* 1997;315:1636.
22. www.healthandeverything.org (accessed December 22, 2006).

CHAPTER 3
Bias in randomized controlled trials

The main appeal of the randomized controlled trial (RCT) in health care comes from its potential to reduce selection bias. Randomization, if done properly, can keep study groups as similar as possible at the outset, so that the investigators can isolate and quantify the effect of the interventions they are studying. No other study design gives us the power to balance unknown prognostic factors at baseline. Random allocation does not, however, protect RCTs against other types of bias.

The existence of most biases related to RCTs is supported mainly by common sense. In recent years, however, important research efforts have used RCTs as the subject rather than the tool of research. These studies are usually designed to generate empirical evidence to improve the design, reporting, dissemination, and use of RCTs in health care.[1] They have confirmed that RCTs are vulnerable to many types of bias throughout their entire life span. Random allocation of the participants to different study groups increases the potential of a study to be free of allocation bias, but has no effect on other important biases.

In this chapter we will discuss the concept of bias in relation to RCTs and highlight some of its sources. We will also list a variety of biases, as well as some strategies that may help us recognize them and minimize their impact on the planning of research and health-related decisions.

What is bias?

An online dictionary[2] defines 'bias' as *'a partiality that prevents objective consideration of an issue or situation'*. In statistics it means *'a tendency of an estimate to deviate in one direction from a true value.*[3] This systematic deviation from the true value can result in either underestimation

or overestimation of the effects of an intervention. Because there is usually more interest in showing that a new intervention works than in showing that it does not work, biases in clinical trials most often lead to an exaggeration in the magnitude or importance of the effects of new interventions.

We should not jump to the conclusion that bias in health research is necessarily associated with a conscious or malicious attempt of investigators, funders, or readers to bend the results of a trial. Indeed, although bias may be introduced into a trial intentionally, it is probably more commonly unintentional, and often unrecognized even by the researchers themselves.

Why does bias in an RCT matter?

The true effects of any health care intervention are unknown. We try to anticipate, detect, quantify, and control bias to produce results from a sample of participants that can be generalized to the target population at large. It is impossible to ever know for sure whether the results of a particular study are biased, simply because it is impossible to establish whether those results depart systematically from a 'truth' that remains unknown.

What are the main types of bias in RCTs?

Most discussions on bias focus on biases that can occur during the actual course of a trial, from the allocation of participants to study groups, through the delivery of interventions, to the measurement of outcomes. Other types of bias can arise, however, even before the trial is carried out, in the choice of problem to study or type of research to use, or after the trial is carried out, in its analysis, and its dissemination. Bias can even be introduced by the person who is reading the report of a trial.[4] These biases, which can also have a profound influence on the way in which the results of RCTs are interpreted and used, tend to receive less attention.

To illustrate how biases can affect the results of an RCT, we invite you to think about the following hypothetical scenario:

> *'Imagine a new drug for the treatment of multiple sclerosis, which has shown promising results in animal studies and in phase I trials. These results, which suggest that the drug can delay the onset of severe motor compromise, have been widely publicized by the media during the past 3 months. Because of these results, patient advocacy groups are putting pressure on the government to make the new drug available as soon as possible. As multiple sclerosis is a debilitating disease that affects millions of people worldwide and for which there is no known cure, the investigators (all clinicians who have dealt with multiple sclerosis patients for years), the company producing the new drug (which has invested millions in developing the drug), the media (interested in confirming the results that they so widely publicized) and the potential participants (patients with multiple sclerosis who have been waiting for an effective treatment to be discovered) are all interested in demonstrating that the new compound is effective. After many intense sessions debating the course of action, a multidisciplinary task force created by the government, including consumer representatives, agrees that the next step should be a randomized clinical trial. A research protocol is produced by another multidisciplinary panel of investigators and consumers, and a well known research group at a large health care center is selected to conduct the study.'*

We discuss the elements of this hypothetical scenario in the following sections.

Selection bias

With true randomization, all participants in the study are given the same opportunity to be allocated or assigned to each of the study groups. But even a perfectly randomized method to allocate participants to the study groups does not protect against selection bias, which can occur both in the way that individuals are accepted or rejected for participation in a trial, and in the way that the interventions are assigned to individuals once they have been accepted into a trial.

Selection bias can occur if some potentially eligible individuals are selectively excluded from the study, because the investigator knows the group to which they would be allocated if they participated. Let us suppose that the investigator in charge of recruiting patients for the multiple sclerosis trial (who at least subconsciously hopes that the drug will be found to be effective) thinks that depressed patients are less likely to respond to the new drug. If he has access to the allocation sequence (which has been generated by computer and is locked in his desk) this investigator could introduce bias into the trial by making it more difficult for depressive

patients to receive the new drug. He could, knowingly or unknowingly, exclude depressive patients who would be allocated to receive the new drug by making them fit the exclusion criteria more easily than if they had been allocated to the placebo group. He could also (again knowingly or unknowingly) present information on the trial to depressive patients allocated to receive the new drug in such a way that they would be discouraged from consenting to participate. At the end of the trial, if the investigator was right, and depressive patients were in truth less likely to respond to the new drug, the trial will show an exaggerated effect of the new drug during the treatment of multiple sclerosis, due to the disproportionate number of depressive patients in the placebo group.

How can selection bias be reduced?

There is empirical evidence to show that effects of new interventions can be exaggerated if the randomization sequence is not concealed from the investigators at the time of obtaining consent from prospective trial participants.[5] One study showed that trials with inadequate allocation concealment can exaggerate the estimate of the effect size of interventions by as much as 40% on average.[6] The irony is that *allocation concealment* is a very simple maneuver that can be incorporated in the design of any trial and that can always be implemented.

Despite its simplicity as a maneuver and its importance to reduce bias, allocation concealment is rarely reported, and perhaps rarely implemented in RCTs. Allocation concealment was reported in less than 10% of articles describing RCTs published in prominent journals in five different languages.[7] This does not necessarily mean that allocation is not concealed in 90% of RCTs; in some cases, allocation may have been concealed, but the authors, peer-reviewers, and journal editors were not aware of how important it is to mention it (it takes about a line in the report, so space limitation is not a good excuse). If, however, allocation concealment was not carried out in most cases in which it was not reported, the majority of RCTs are at risk of exaggerating the effects of the interventions they were designed to evaluate.

Even if the report of an RCT states that efforts were made to conceal the allocation sequence, there are many ways in which randomization can be subverted by investigators who want to break the allocation code before they obtain consent from prospective trial participants.[8] Even when the allocation codes are kept

in sealed opaque envelopes, for instance, investigators can (and sometimes do) look through the envelopes using powerful lights or even open the envelope using steam and reseal it without others noticing. Thus it is very easy to introduce selection bias into RCTs.

Users of RCTs should not get a false sense of security just because a study is randomized.

Ascertainment bias

Ascertainment bias occurs when the results or conclusions of a trial are systematically distorted by knowledge of which intervention each participant is receiving. Ascertainment bias can be introduced by the person administering the interventions, the person receiving the interventions (the participants), the investigator assessing or analyzing the outcomes, and even by the people who write the report describing the trial (Chapter 2).

The best way to protect a trial against ascertainment bias is by keeping the people involved in the trial unaware of the identity of the interventions for as long as possible. This is called blinding or masking. The strategies that can be used to reduce ascertainment bias can be applied during at least two periods of a trial: the time during which data are collected actively (from the administration of the interventions to the gathering of outcome data) and after data have been collected (from data analysis to the reporting of results).

It is important to recognize the difference between biases that are the result of lack of allocation concealment and biases that arise from lack of blinding. *Allocation concealment* helps to prevent selection bias, protects the randomization sequence *before* and *until* the interventions are given to study participants, and can *always* be implemented. *Blinding* helps prevent ascertainment bias, protects the randomization sequence *after* allocation, and cannot always be implemented.[6]

How can ascertainment bias be reduced during data collection?

Ascertainment bias can be introduced in different ways during data collection. For instance, the people administering the interventions can bias the results of a trial by altering systematically the co-interventions given to participants during the trial. Following our

example of the multiple sclerosis trial, the new drug may appear to be more effective at the end of the trial if patients allocated to the new drug received physiotherapy earlier and more intensively than patients allocated to placebo (*co-intervention bias*). If participants know that they have been allocated to the placebo group, they are likely to feel disappointed and less willing to report improvement at each of the study time points (*participant ascertainment bias*). In addition, if the people in charge of assessing and recording the outcomes know which patients are allocated to each of the study groups, they could, consciously or unconsciously, tend to record the outcomes for patients receiving the new drug in a more favorable way than for patients receiving placebo (*observer bias*).

In ideal circumstances, ascertainment bias should be reduced by blinding all concerned: the individuals who administer the interventions, the participants who receive the interventions and the individuals in charge of assessing and recording the outcomes (Chapter 2).

The importance of blinding has been confirmed in empirical studies. It has been shown, for instance, that open studies are more likely to favor experimental interventions over the controls[9] and that studies that are not double-blinded can exaggerate effect estimates by 17%.[6] Despite the empirical evidence available, and common sense, only about half of the trials that could be double-blinded actually were.[10] Even when the trials are described as double-blind, most reports do not provide adequate information on how blinding was achieved or statements on the perceived success (or failure) of double-blinding efforts.[11,12]

The best strategy to achieve blinding during data collection is with the use of placebos. *Placebos* are interventions believed to be inactive, but otherwise identical to the experimental intervention in all aspects other than the postulated specific effect. Placebos are certainly easier to develop and implement successfully in drug trials, in which they should resemble the taste, smell and appearance of the active drug, and should be given using an identical procedure.

Placebo controls can also be used with non-drug interventions, such as psychological, physical, and surgical procedures, although they are more difficult to develop and implement successfully. For example, it is difficult, but not impossible to develop and implement placebo counseling, physiotherapy, acupuncture or electrical stimulation. In some cases it is impossible, unfeasible or simply unethical to use placebos. It would be impossible, for example, to use a placebo intervention in a trial evaluating the effect on mothers and newborns of early versus late discharge from hospital after childbirth. It would be unfeasible or unethical to use a placebo in trials evaluating new or existing surgical interventions (although a strong case can still be made for trials in which sham surgery can successfully challenge the perceived effectiveness of surgical interventions).[13] Placebo controlled studies are not ethical to study a new or existing intervention when there is an effective intervention available (Chapter 8). Even in cases where the use of placebos is impossible, unfeasible or unethical, trials can be at least single blind. In a surgical or acupuncture trial, for instance, single-blinding can be achieved by keeping the investigators in charge of assessing the outcomes unaware of which participants receive which interventions.

How can ascertainment bias be reduced after data collection?

Ascertainment bias can be introduced easily after data collection, if the investigators in charge of analyzing or reporting the results of the trial are aware of which participants are receiving which interventions. The effects of a new intervention can be exaggerated, for instance, if the investigators in charge of analyzing the trial data select the outcomes and the time points that show maximum benefit from the new intervention and ignore outcomes and time points that show either no effect or harm from the new intervention. Similarly, investigators in charge of reporting the trial results can choose to emphasize the outcomes and time points that show the maximum effects of the new intervention, downplaying or ignoring findings that suggest that the new intervention is equivalent or less effective than the control.

This source of bias can be controlled by keeping the data analysts and the people in charge of reporting the trial results unaware of the identity of the study groups. In a study with two groups, for instance, the outcome data could be given to analysts coded as A and B, and once they complete the analysis, the results could be given to the person in charge of writing the report using the same codes. The codes would not be broken until after the data analysis and reporting phases were completed. These valuable strategies should be used and studied more often.

Other important sources of bias

What biases can occur during the planning phase of an RCT?

Choice-of-question bias

Perhaps one of the least recognized forms of bias in an RCT is hidden in the choice of the question that the trial intends to answer. This would not necessarily affect the internal validity of a trial, but may have profound effects on its external validity, or generalizability. This bias can take many forms.

Hidden agenda bias occurs when a trial is mounted, not in order to answer a question, but in order to demonstrate a pre-required answer. The unspoken converse may be 'Don't do a trial if it won't show you what you want to find'. This could be called the *vested interest bias*.[14] Closely related to this is the *self fulfiling prophecy bias*, in which the very carrying out of a trial ensures the desired result.

The *cost and convenience bias* can seriously compromise what we choose to study. When we study what we can afford to study, or what is convenient to study, rather than what we really want to study, or should study, we take resources away from what we know is important. Closely related to this is the *funding availability bias* where studies tend to concentrate on questions that are more readily fundable, often for a vested or commercial interest. We should always look for the *secondary gains search bias* which can influence the choice of study, the methodology used, and the ascertainment and dissemination of the results.

Regulation bias

This is sometimes referred to as the *IRB bias* or the *Bureaucracy bias*. It occurs when institutional review boards are overly restrictive, and block the study of important questions. It also occurs when they are overly permissive and allow or even encourage studies that may not be scientifically or socially valid, but may bring either funding or prestige to the institution. Complicated 'informed consent' regulations may block the participation of many otherwise eligible subjects, and hence bias the results (Chapter 8).

Wrong design bias

The perceived value of an RCT may sometimes induce researchers to use this design for questions that may be better (or can only be answered) with a different design, such as outcome research.[15] The wrong research design can produce misleading answers.

What biases can occur during the course of an RCT?

Population choice bias

The sample population studied can have a major effect on the generalizability of an RCT. If the sample is overly restricted by not including women (*gender bias*) or people over (or under) a specific age group (*age bias*), the results may not be generalizable to people who do not belong to the groups. *Pregnancy bias*, (excluding pregnant women) may sometimes be necessary for reasons of safety to the fetus, but the exclusion must be carefully noted. The same reasoning is required when trials are restricted to, or exclude, people in special circumstances (*special circumstances bias*).

Population choice may be restricted when potential participants are approached (*recruitment bias*) or during registration of participants. Eligible patients may be kept out of a trial because they do not understand the consent form (*informed consent bias, literacy bias, language bias*).

Severity of illness bias is an important subgroup of the sample choice bias. Patients with a mild form of an illness may not respond in the same way as those with a more severe form.

Intervention choice bias

The nature of the intervention chosen can have a major effect on the results obtained. The stage at which an intervention is studied can be very important. The *too early bias* and the *too late bias* can determine the effects found.[16] This holds particularly true for surgical trials where there can be a *learning curve bias* for new operators, or improvements (or regression) in the techniques or contexts in which they are used. Similar concerns may hold for medical interventions, when dose or timing of a medication may be important determinants of the outcome.

Complexity bias can occur when a trial is used to study complex interventions, with a number of components, or where outcomes may depend on multiple contingencies outside of the control of the investigator (e.g. the skill of the surgeons or the resources of the community).[17]

Comparison choice (or control group) bias

If an intervention is compared to a poorly chosen control group, it can erroneously appear to be more (or less) effective than it really is. If a study compares an experimental intervention with a placebo control, the results will only tell us whether the intervention has a specific effect or not. It will not imply that the experimental intervention has a different or better effect than existing alternatives. An obvious way to make an intervention appear to be more effective than it really is would be to choose an ineffective comparison group.

Unfortunately, current regulatory bodies that mandate placebo controls lead to carrying out studies with this limited clinical value.

Outcome choice bias

Sometimes RCTs evaluate outcomes that are easy to measure, rather than the outcomes that are relevant (*measurement bias*). One variant of this is the *time term bias* in which short-term outcomes are measured rather than the important long-term outcomes. It is not surprising that researchers sometimes yield to the temptation to study outcomes that are readily measured rather than those that are important.

What biases can occur during the reporting of a trial?

Withdrawal bias: bias introduced by inappropriate handling of withdrawals, drop outs, and protocol violations

Ideally, all participants in a trial should complete the study, follow the protocol, and provide data on all the outcomes of interest at all time points. In reality, however, most trials have missing data. Data can be missing because some of the participants drop out before the end of the trial, because participants do not follow the protocol either deliberately or accidentally, or because some outcomes are not measured correctly or cannot be measured at all at one or more time points.

Regardless of the cause, inappropriate handling of the missing information can lead to bias. For instance, if in the multiple sclerosis trial patients who do not obtain benefit from the new drug withdraw more frequently because of adverse effects, their exclusion from analysis would lead the investigators to exaggerate the benefit and underestimate the harm of the new drug. This bias can occur independently of whether or not the investigators are aware of the identity of the interventions received by the participants. If the decisions on withdrawals have been made because of knowledge of the interventions received by the participants, this constitutes yet another cause of ascertainment bias.

On occasion, it is impossible to know the status of participants at the times when the missing information should have been collected. This could happen, for example, if participants move to different areas during the study or fail to contact the investigators for an unknown reason. If the reasons for excluding these participants or specific outcome measurements from the final analysis were in any way related to the intervention, this could also lead to bias.

There are two strategies that can confidently be assumed to eliminate bias in these circumstances. One is known as *intention-to-treat analysis*, which means that all the study participants are included in the analyses as part of the groups to which they were randomized, regardless of whether they completed the study or not. The second method is a *worst-case scenario* or *sensitivity analysis*. This is performed by assigning the worst possible outcomes to the missing patients or time points in the group that shows the best results, and the best possible outcomes to the missing patients or timepoints in the group with the worst results. We can then see whether the new analysis contradicts or supports the results of the initial analysis excluding the missing data.

Selective reporting bias

A major and common source of bias in an RCT is selective reporting of results, describing those outcomes with positive results, or which favor the studied intervention. This is not always consciously done. The investigator may even unconsciously be attracted more to certain outcomes than others. Variants of this have been named the *social desirability bias* in which the items that are desired, or the *optimism bias* in which the items hoped for, are more likely to be reported.

The *data dredging bias* is another variant of the selective reporting bias. Having looked at all the data, the investigators can report the outcomes they wish to stress, and not mention the less desirable outcomes. A variant is the *interesting data bias*, in which the authors report the data that they find most interesting. The acme of data dredging can be in the selective analysis of data. If unethically contrived, all trials can be made to appear to have positive results.[18]

Fraud bias

Intentional fraud is perhaps the most important, most serious, and most difficult to detect source of bias. We hope that it is rare, but the extent to which fraudulent results are reported may be underestimated, and may be increasing under the pressure to produce results, no matter how.

What biases can occur during the dissemination of the trials?

What is publication bias?

Investigators and sponsors are more likely to write and submit, and peer-reviewers and editors to accept, manuscripts with positive results for publication. This tendency has been called *publication bias*.[19,20] A systematic review of five empirical methodological studies published mostly during the previous 10 years confirmed that the failure to publish is not a random event, but is heavily influenced by the direction and strength of research findings, whereby manuscripts with statistically significant (positive) results are published preferentially over manuscripts reporting nonsignificant (negative) results.[21] Publication bias may be the main factor behind the systematic differences found between studies funded by industry and their counterparts.[14,22]

Efforts have been made to eliminate publication bias through compulsory registration of trials at inception, and publication of the results of all trials. These have been the focus of intense debate and controversy for several years, fueled by strong ethical and economic interests. Many major journals now refuse to publish the results of studies that had not been registered at inception. Even so, readers must be aware that by relying on published studies to guide their decisions they are always at risk of overestimating the effect of interventions[23–25] (see Chapter 5).

What is language bias?

Recently, a variation of publication bias has been described as *language bias*, to indicate that manuscripts may be submitted to and published by journals in different languages depending on the direction of their results. More studies with positive results may be published in English.[26] A variant of this is the *country of publication bias*, the tendency by some countries to publish a disproportionate number of positive trials.[27]

What is time lag bias?

This bias occurs when the speed of publication depends on the direction and strength of the trial results. In general, it seems that trials with 'negative' results take twice as long to be published as 'positive' trials.[28,29]

What biases can occur during the uptake phase?

Up to this point we have focused on the biases introduced by the investigators who plan and carry out randomized trials, or those who publish and disseminate the results. As this book is primarily a user's guide, rather than a manual for researchers, we felt that we should emphasize the responsibility of the reader of research studies.

Different types of reader biases were described many years ago.[4] At the time in which they were reported, the existence of these biases was supported only by common sense and experience. Recently, there have been empirical studies that support the existence of reader bias, showing that there are systematic differences in the way readers assess the quality of RCTs depending on whether the assessments are conducted under masked or open conditions.[11,30] These studies, however, do not focus on any specific type of reader bias.

The following are some of the biases that we believe are most common and pertinent:

Relation to the author bias, with its subgroups *Rivalry bias* (underrating the strengths or exaggerating the weaknesses of studies published by a rival) and *I owe him one* bias (favoring flawed results from a study by someone who did the same for the reader).

Personal habit bias occurs when readers overrate or underrate a study depending on their own habits (e.g. a reader who enjoys eating animal fat overrating a study that challenges the adverse effects of animal fat on health). This is similar to the *moral bias*, in which readers overrate or underrate a study depending on how much it agrees or disagrees with their moral views (e.g. a reader who regards abortion as immoral overrating a study showing a relationship between abortion and breast cancer). This is closely related to the *values bias* (depending on how important you consider the outcomes of the study to be).

Clinical practice bias takes place when readers judge a study according to whether it supports or challenges their current or past clinical practice (e.g. a clinician who gives lidocaine to patients with acute myocardial infarction underrating a study that suggests that lidocaine may increase mortality in these patients). This is similar to

the *institution bias* (that is, or is not, the way that we do it in our hospital), and the *territory bias* which can occur when readers overrate studies that support their own specialty or profession (e.g. a surgeon favoring a study that suggests that surgery is more effective than medical treatment, or obstetricians underrating a study that suggests that midwives can provide adequate care during uncomplicated pregnancies and deliveries). *Tradition bias* happens when a reader rates a study depending on whether it supports or challenges traditional procedures (e.g. underrating a study that challenges episiotomy during normal vaginal deliveries).

Do something bias means overrating a study that suggests that an intervention is effective, particularly when there is no alternative effective intervention available. This bias may be common among clinicians and patients (e.g. a patient with AIDS overrating a study describing a cure for AIDS).

In this general heading we can include the *technology bias*, which relates to judging a study according to the reader's attraction or aversion for technology in health care. *Resource allocation bias* happens when readers have a strong preference for one type of resource allocation. This bias may be one of the most frequently found in health care, as it can emanate from consumers, clinicians, policy makers, researchers, and fund holders.

Printed word bias occurs when a study is overrated because of undue confidence in published data. Subgroups of the printed word bias include the *prestigious journal bias* (the results of studies published in prestigious journals are overrated), and its opposite, the *non-prestigious journal bias*. Similar to this is the *peer review bias*, which comes into play when readers have an unwarranted belief in the ability of peer review to guarantee the validity of a study.

Prominent author bias occurs when the results of studies published by prominent authors are overrated, and, of course has its converse in the *unknown or non-prominent author bias*. This has been called the '*who is s/he? bias*'.[4] Similar to these are the *famous institution bias*, the *credential or professional background bias* (e.g. physicians underrating research done by nurses or vice versa; basic scientists underrating research done by clinicians or vice versa; PhDs underrating studies published by MDs and vice versa; readers overrating research by authors with many letters after their names and vice versa). Their variants include the *esteemed author bias, esteemed professor bias*, and the *friendship bias*; when the reader overrates results obtained by a close friend or mentor.

We are not through yet!

Geography bias occurs when studies are judged according to the country or region where it was conducted, and is closely related to the *language bias* (e.g. the belief that studies published in languages other than English are of inferior quality than those published in English).[26]

The *trial design bias* can go in either direction. The *favored design bias* occurs when a study that uses a design supported, publicly or privately, by the reader (e.g. a consumer advocate overrating an RCT that takes into account patient preferences). Its converse is the *Disfavored design bias*. Somewhat related are the *large trial bias*, in which the results of large trials are overrated, and the *multicentre trial bias* when the results of multicentre collaborative trials are overrated. The *small trial bias* occurs when the results of trials with small sample size are underrated, particularly when they contradict the opinion of the reader (i.e. attributing to chance any statistically or clinically significant effect found by a small trial, or any lack of significant effects to low power).

Complementary medicine bias refers to the systematic overrating or underrating of studies that describe complementary medicine interventions, particularly when the results suggest that the interventions are effective.

Flashy title bias occurs when the results of studies with attractive titles are overrated (particularly by patients or journalists) or underrated (particularly by academics if they regard them as sensationalist!). Other rather tricky biases include the *substituted question bias*, when a reader substitutes a question for the question that the study is designed to answer and regards the results of the study as invalid if they do not answer the substituted question.

Vested interest bias has a number of subgroups. *Bankbook bias* occurs when a study is rated depending on the impact of its results on the income of the reader (e.g. a surgeon underrating a study that questions the need for surgery to relieve back pain in patients with spinal stenosis, or a pharmaceutical company overrating the results of a study that supports the use of one of its products). *Cherished belief bias* reminds us that there are other competing interests besides the financial ones.

Reader attitude biases include the *Belligerence bias* which results in underrating studies systematically just for the sake of being difficult; the *Empiricism bias* (overrating or underrating a study because it challenges the clinical experience of the reader), or the *I am an*

epidemiologist bias in which the reader repudiates a study that contains any flaw, albeit minor, in its design, analysis or interpretation.

Finally, *careless reading bias* occurs when a study is overrated or underrated because the reader neglected to read a key section. Unfortunately, far too common.

Musings

This has been a difficult chapter to write. We approached it with fear and trepidation, feeling part of a 'no win' situation. We know that the control of bias is the *raison d'etre* for clinical trials, and accept that control of bias is the most important factor in diminishing inevitable error. We know that allocation bias is a major source of potential error in clinical comparison studies, and we know that randomization, if properly done, can control for allocation bias. We want to stress the value of randomization for this purpose, and the vital importance of RCTs.

But we also realize that randomization *per se* can control *only* for allocation bias, and this does not even completely control for selection bias. Other biases can also subvert the validity of conclusions at any stage in the planning, conduct, analysis, or interpretation of the results. As we worked together on this chapter, as we uncovered an increasing number of biases, our fears mounted. We started to feel very discouraged. What is the big deal, if this seemingly powerful tool is so vulnerable? Why should we believe in trials if they can be subverted so easily and at so many levels? If biases cannot be controlled, what is left? We are not sufficiently naïve to think that by finding biases and naming them that we can overcome them. Can we run the risk that by drawing attention to the biases we would attack the very foundation of RCTs, and appear to advocate nihilism?

We believed (and still believe) in the value of RCTs. We felt like heretics, not for the first time.[31] Both of us were, and are, strong and enthusiastic proponents of RCTs. Indeed our support for RCTs has become even stronger as we have become more aware of their limitations. But it is no longer a blind faith, rather one that has been through and survived the crises of doubt.

We are concerned with the danger that RCTs may be perceived as a sort of talisman, to protect us from the evil of bias. But randomized trials are not divine revelations, they are human constructs, and like all human constructs, are fallible. They are valuable, useful

tools that should be used wisely and well. We believe that a strong belief in the strength of randomized trials, without acknowledging their weaknesses, runs the risk of fundamentalism and intolerance of criticism, or alternative views. In this way, it can discourage innovation.

Our list of biases is far from exhaustive. The number of possible biases is practically infinite, as is the names that can be given to them, or the ways in which they can be classified or categorized. RCTs can never be completely objective. They should be carried out with humility; the investigator should be as up front, explicit, and transparent as possible about his or her motivations for choosing to carry out the trial, the methods used, the outcomes looked for as well as the outcomes found. Journalists have an important responsibility to assume, because of their influence on public understanding. At present they tend to bring to public attention the results of trials purporting beneficial effects of a new intervention for incurable diseases, while they ignore the results of previous (or concurrent) trials in which the same intervention showed no benefit.[32] This media coverage may influence the decisions of clinicians and patients who are not aware of the other studies. The same onus must be put on the reader, the one who will be making use of the information gleaned from the trial. It can be far too easy to criticize an RCT, or to read into it what we want, to find rather than what the results actually show.

Our bottom line is that a new sense of freedom can emerge, as we free ourselves from a false sense of objectivity, and can recognize and use RCTs as the valuable tools that they are, when they are the right tool in the right place.

References

1. Jadad AR, Rennie D. The randomized controlled trial gets a middle-aged checkup. *Journal of American Medical Association* 1998;279:319–320.
2. http://www.wordreference.com/definition/bias (accessed December 22, 2006).
3. *Webster's Third New International Dictionary*, Unabridged. G&C Merriam Co, 1976.
4. Owen R. Reader bias. *Journal of American Medical Association* 1982; 247:2533–2534.
5. Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical trials. *New England Journal of Medicine* 1983; 309:1359–1361.

6. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of American Medical Association* 1995;273:408–412.
7. Moher D, Fortin P, Jadad AR, Juni P, Klassen T, Le Lorier J, Liberati A, Linde K, Penna A. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. *Lancet* 1996;347:363–366.
8. Schulz KF. Subverting randomization in controlled trials. *Journal of American Medical Association* 1995;274:1456–1458.
9. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Therapy. *Statistics in Medicine* 1989;8:441–454.
10. Schulz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. *British Medical Journal* 1996;312:742–744.
11. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds JM, Gavaghan DJ, McQuay DM. Assessing the quality of reports on randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials* 1996;17:1–12.
12. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials. *International Journal of Technology Assessment in Health Care* 1996;12:195–208.
13. Freeman TB. Use of placebo surgery in controlled trials of a cellular-based therapy for Parkinson's disease. *New England Journal of Medicine* 1999;34:988–992.
14. Fries JF, Krishnan E. Equipoise, design bias, and randomized controlled trials: the elusive ethics of new drug development. *Arthritis Research and Therapy* 2004;6:R250–R255.
15. Sackett DL, Wennberg JE. Choosing the best research design for each question: It's time to stop squabbling over the 'best' methods. *British Medical Journal* 1997;315:1636.
16. Lilford RJ, Braunholtz DA, Greenhalgh R, Edwards SJL. Trials and fast changing technologies: the case for tracker studies. *British Medical Journal* 2000;320;43–46.
17. Kotaska A. Inappropriate use of randomised trials to evaluate complex phenomena: case study of vaginal breech delivery. *British Medical Journal* 2004;329:1039–1042
18. Martin G. Munchausen's statistical grid, which makes all trials significant. *Lancet* 1984;ii:1457.
19. Dickersin K. The existence of publication bias and risk factors for its occurrence. *Journal of American Medical Association* 1990;263: 1385–1389.
20. Rennie D, Flanagin A. Publication bias – the triumph of hope over experience. *Journal of American Medical Association* 1992;267:411–412.

21. Dickersin K. How important is publication bias? A synthesis of available data. *AIDS Education Prevention* 1997;9(Suppl A):15–21.
22. Smith R. Medical Journals are an extension of the marketing arm of pharmaceutical companies. http://medicine.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pmed.0020124 (accessed November 21, 2006)
23. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of American Medical Association* 1996;276:637–639.
24. Grimes DA. The 'CONSORT' guidelines for randomized controlled trials in Obstetrics and Gynecology. *Obstetrics and Gynecology* 2002;100: 631–632.
25. Rennie D. How to report randomized controlled trials. The CONSORT statement. *Journal of American Medical Association* 1996;276:649.
26. Moher D, Pham B, Lawson ML, Klassen TP. The inclusion of reports of randomised trials published in languages other than English in systematic reviews. *Health Technology Assessment* 2003;7:1–90.
27. Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. *Controlled Clinical Trials* 1998;19:159–166.
28. Ioannidis JPA. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials: a survival analysis. *Journal of American Medical Association* 1998;279:281–286.
29. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *British Medical Journal* 1997; 315:640–645.
30. McNutt RA, Evans AT, Fletcher RH, Fletcher SW. The effects of blinding on the quality of peer review. A randomized trial. *Journal of American Medical Association* 1990;263:1371–1376.
31. Enkin MW, Jadad AR. Using anecdotal information in evidence-based health care: heresy or necessity? *Annals of Oncology* 1998;9:963–966.
32. Koren G, Klein N. Bias against negative studies in newspaper reports of medical research. *Journal of American Medical Association* 1991; 266:1824–1826.

Fundamentals of Clinical Research for Radiologists

Harald O. Stolberg[1]
Geoffrey Norman[2]
Isabelle Trop[3]

Randomized Controlled Trials

Preceding articles in this series have provided a great deal of information concerning research design and methodology, including research protocols, statistical analyses, and assessment of the clinical importance of radiologic research studies. Many methods of research design have already been presented, including descriptive studies (e.g., case reports, case series, and cross-sectional surveys), and some analytical designs (e.g., cohort and case-control studies).

Case-control and cohort studies are also called observational studies, which distinguishes them from interventional (experimental) studies because the decision to seek one treatment or another, or to be exposed to one risk or another, was made by someone other than the experimenter. Consequently, the researcher's role is one of observing the outcome of these exposures. By contrast, in experimental studies, the researcher (experimenter) controls the exposure. The most powerful type of experimental study is the randomized controlled trial. The basic principles of randomized controlled trials will be discussed in this article.

History of Randomized Controlled Trials

The history of clinical trials dates back to approximately 600 B.C. when Daniel of Judah [1] conducted what is probably the earliest recorded clinical trial. He compared the health effects of the vegetarian diet with those of a royal Babylonian diet over a 10-day period. The trial had obvious deficiencies by contemporary medical standards (allocation bias, ascertainment bias, and confounding by divine intervention), but the report has remained influential for more than two millennia [2].

The 19th century saw many major advances in clinical trials. In 1836, the editor of the *American Journal of Medical Sciences* wrote an introduction to an article that he considered "one of the most important medical works of the present century, marking the start of a new era of science," and stated that the article was "the first formal exposition of the results of the only true method of investigation in regard to the therapeutic value of remedial agents." The article that evoked such effusive praise was the French study on bloodletting in treatment of pneumonia by P. C. A. Louis [2, 3].

Credit for the modern randomized trial is usually given to Sir Austin Bradford Hill [4]. The Medical Research Council trials on streptomycin for pulmonary tuberculosis are rightly regarded as a landmark that ushered in a new era of medicine. Since Hill's pioneering achievement, the methodology of the randomized controlled trial has been increasingly accepted and the number of randomized controlled trials reported has grown exponentially. The Cochrane Library already lists more than 150,000 such trials, and they have become the underlying basis for what is currently called "evidence-based medicine" [5].

General Principles of Randomized Controlled Trials

The randomized controlled trial is one of the simplest but most powerful tools of research. In essence, the randomized controlled trial is a study in which people are allocated at random to receive one of several clinical interventions [2]. On most occasions, the term "intervention" refers to treatment, but it should be used in a much wider sense to include any clinical maneuver offered to study participants that may

Received June 14, 2004; accepted after revision July 2, 2004.

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 12th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology*. The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment; staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

[1]Department of Radiology, McMaster University Medical Centre, 1200 Main St. W, Hamilton, ON L8N 3Z5, Canada. Address correspondence to H. O. Stolberg.

[2]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON L8N 3Z5, Canada.

[3]Department of Radiology, Hôpital Saint-Luc, 1058 St. Denis St., Montreal, QC H2X 3J4, Canada.

AJR 2004;183:1539–1544

0361–803X/04/1836–1539

have an effect on their health status. Such clinical maneuvers include prevention strategies, screening programs, diagnostic tests, interventional procedures, the setting in which health care is provided, and educational models [2]. Randomized controlled trials in radiology can play a major role in the assessment of screening programs, diagnostic tests, and procedures in interventional radiology [6–13].

Randomized controlled trials are used to examine the effect of interventions on particular outcomes such as death or the recurrence of disease. Some consider randomized controlled trials to be the best of all research designs [14], or "the most powerful tool in modern clinical research" [15], mainly because the act of randomizing patients to receive or not receive the intervention ensures that, on average, all other possible causes are equal between the two groups. Thus, any significant differences between groups in the outcome event can be attributed to the intervention and not to some other unidentified factor. However, randomized controlled trials are not a panacea to answer all clinical questions; for example, the effect of a risk factor such as smoking cannot ethically be addressed with randomized controlled trials. Furthermore, in many situations randomized controlled trials are not feasible, necessary, appropriate, or even sufficient to help solve important problems [2]. Randomized controlled trials are not appropriate for cancer screening, a situation in which the outcome is rare and frequently occurs only after a long delay. Thus, although the test for appraising the ultimate value of a diagnostic test may be a large well-designed randomized controlled trial that has patient outcomes as the end point [16], the trial should presumably be performed after other smaller studies have examined the predictive value of the test against some accepted standard.

An excellent example of the controversies that can arise with randomized controlled trials is an overview of the publications on mammography screening. The most important references concern the article by Miettinen et al. [17] linking screening for breast cancer with mammography and an apparently substantial reduction in fatalities and the responses that it elicited [18–22].

Randomized controlled trials may not be appropriate for the assessment of interventions that have rare outcomes or effects that take a long time to develop. In such instances, other study designs such as case-control studies or cohort studies are more appropriate. In other cases, randomized controlled trials may not be feasible because of financial constraints or because of the expectation of low compliance or high drop-out rates.

Many randomized controlled trials involve large sample sizes because many treatments have relatively small effects. The size of the expected effect of the intervention is the main determinant of the sample size necessary to conduct a successful randomized controlled trial. Obtaining statistically significant differences between two samples is easy if large differences are expected. However, the smaller the expected effect of the intervention, the larger the sample size needed to be able to conclude, with enough power, that the differences are unlikely to be due to chance. For example, let us assume that we wish to study two groups of patients who will undergo different interventions, one of which is a new procedure. We expect a 10% decrease in the morbidity rate with the new procedure. To be able to detect this difference with a probability (power) of 80%, we need 80 patients in each treatment arm. If the expected difference in effect between the two groups increases to 20%, the number of patient required per arm decreases to 40. Conversely, if the difference between the groups is expected to be only 1%, the study population must increase to 8,000 per treatment arm. The sample size required to achieve power in a study is inversely proportional to the treatment effect squared [23]. Standard formulas are available to calculate the approximate sample size necessary when designing a randomized controlled trial [24–26].

Randomization: The Strength of the Randomized Controlled Trial

The randomization procedure gives the randomized controlled trial its strength. Random allocation means that all participants have the same chance of being assigned to each of the study groups [27]. The allocation, therefore, is not determined by the investigators, the clinicians, or the study participants [2]. The purpose of random allocation of participants is to assure that the characteristics of the participants are as likely to be similar as possible across groups at the start of the comparison (also called the baseline). If randomization is done properly, it reduces the risk of a serious imbalance in known and unknown factors that could influence the clinical course of the participants. No other study design allows investigators to balance these factors.

The investigators should follow two rules to ensure the success of the randomization procedure. They must first define the rules that will govern allocation and then follow those rules strictly throughout the entire study [2]. The crucial issue is that after the procedure for randomization is determined, it should not be modified at any point during the study. There are many adequate methods of randomization, but their common element is that no one should be able to determine ahead of time to which group a given patient will be assigned. Detailed discussion of randomization methods is beyond the scope of this article.

Numerous methods are also available to ensure that the sample of patients is balanced whenever a small predetermined number of patients have been enrolled. Unfortunately, the methods of allocation in studies described as randomized are poorly and infrequently reported [2, 28]. As a result, it is not possible to determine, on most occasions, whether the investigators used proper methods to generate random sequences of allocation [2].

Bias in Randomized Controlled Trials

The main appeal of the randomized controlled trial in health care derives from its potential for reducing allocation bias [2]. No other study design allows researchers to balance unknown prognostic factors at baseline. Random allocation does not, however, protect randomized controlled trials against other types of bias. During the past 10 years, randomized controlled trials have been the subject rather than the tool of important, albeit isolated, research efforts usually designed to generate empiric evidence to improve the design, reporting, dissemination, and use of randomized controlled trials in health care [28]. Such studies have shown that randomized controlled trials are vulnerable to multiple types of bias at all stages of their workspan. A detailed discussion of bias in randomized controlled trials was offered by Jadad [2].

In summary, randomized controlled trials are quantitative, comparative, controlled experiments in which a group of investigators studies two or more interventions by administering them to groups of individuals who have been randomly assigned to receive each intervention. Alternatively, each individual might receive a series of interventions in random order (crossover design) if the outcome can be uniquely associated with each intervention, through, for example, use of a "washout" period. This step ensures that the

effects from one test are not carried over to the next one and subsequently affect the independent evaluation of the second test administered. Apart from random allocation to comparison groups, the elements of a randomized controlled trial are no different from those of any other type of prospective, comparative, quantitative study.

Types of Randomized Controlled Trials

As Jadad observed in his 1998 book *Randomised Controlled Trials* [2]:

> Over the years, multiple terms have been used to describe different types of randomized controlled trials. This terminology has evolved to the point of becoming real jargon. This jargon is not easy to understand for those who are starting their careers as clinicians or researchers because there is no single source with clear and simple definitions of all these terms.

The best classification of frequently used terms was offered by Jadad [2], and we have based our article on his work.

According to Jadad, randomized controlled trials can be classified as to the aspects of intervention that investigators want to explore, the way in which the participants are exposed to the intervention, the number of participants included in the study, whether the investigators and participants know which intervention is being assessed, and whether the preference of nonrandomized individuals and participants has been taken into account in the design of the study. In the context of this article, we can offer only a brief discussion of each of the different types of randomized controlled trials.

Randomized Controlled Trials Classified According to the Different Aspects of Interventions Evaluated

Randomized controlled trials used to evaluate different interventions include explanatory or pragmatic trials; efficacy or equivalence trials; and phase 1, 2, 3, and 4 trials.

Explanatory or pragmatic trials.—Explanatory trials are designed to answer a simple question: Does the intervention work? If it does, then the trial attempts to establish how it works. Pragmatic trials, on the other hand, are designed not only to determine whether the intervention works but also to describe all the consequences of the intervention and its use under circumstances corresponding to daily practice. Although both explanatory and pragmatic approaches are reasonable, and even complementary, it is important to understand that they represent extremes of a spectrum, and most randomized controlled trials combine elements of both.

Efficacy or effectiveness trials.—Randomized controlled trials are also often described in terms of whether they evaluate the efficacy or effectiveness of an intervention. Efficacy refers to interventions carried out under ideal circumstances, whereas effectiveness evaluates the effects of an intervention under circumstances similar to those found in daily practice.

Phase 1, 2, 3, and 4 trials.—These terms describe the different types of trials used for the introduction of a new intervention, traditionally a new drug, but could also encompass trials used for the evaluation of a new embolization material or type of prosthesis, for example. Phase 1 studies are usually conducted after the safety of the new intervention has been documented in animal research, and their purpose is to document the safety of the intervention in humans. Phase 1 studies are usually performed on healthy volunteers. Once the intervention passes phase 1, phase 2 begins. Typically, the intervention is given to a small group of real patients, and the purpose of this study is to evaluate the efficacy of different modes of administration of the intervention to patients. Phase 2 studies focus on efficacy while still providing information on safety. Phase 3 studies are typically effectiveness trials, which are performed after a given procedure has been shown to be safe with a reasonable chance of improving patients' conditions. Most phase 3 trials are randomized controlled trials. Phase 4 studies are equivalent to postmarketing studies of the intervention; they are performed to identify and monitor possible adverse events not yet documented.

Randomized Controlled Trials Classified According to Participants' Exposure and Response to the Intervention

These types of randomized controlled trials include parallel, crossover, and factorial designs.

Parallel design.—Most randomized controlled trials have parallel designs in which each group of participants is exposed to only one of the study interventions.

Crossover design.— Crossover design refers to a study in which each of the participants is given all of the study interventions in successive periods. The order in which the participants receive each of the study interventions is determined at random. This design, obviously, is appropriate only for chronic conditions that are fairly stable over time and for interventions that last a short time within the patient and that do not interfere with one another. Otherwise, false conclusions about the effectiveness of an intervention could be drawn [29].

Factorial design.—A randomized controlled trial has a factorial design when two or more experimental interventions are not only evaluated separately but also in combination and against a control [2]. For example, a 2 × 2 factorial design generates four sets of data to analyze: data on patients who received none of the interventions, patients who received treatment A, patients who received treatment B, and patients who received both A and B. More complex factorial designs, involving multiple factors, are occasionally used. The strength of this design is that it provides more information than parallel designs. In addition to the effects of each treatment, factorial design allows evaluation of the interaction that may exist between two treatments. Because randomized controlled trials are generally expensive to conduct, the more answers that can be obtained, the better.

Randomized Controlled Trials Classified According to the Number of Participants

Randomized controlled trials can be performed in one or many centers and can include from one to thousands of participants, and they can have fixed or variable (sequential) numbers of participants.

"N-of-one trials."—Randomized controlled trials with only one participant are called "*n*-of-one trials" or "individual patient trials." Randomized controlled trials with a simple design that involve thousands of patients and limited data collection are called "megatrials." [30, 31]. Usually, megatrials require the participation of many investigators from multiple centers and from different countries [2].

Sequential trials.—A sequential trial is a study with parallel design in which the number of participants is not specified by the investigators beforehand. Instead, the investigators continue recruiting participants until a clear benefit of one of the interventions is observed or until they become convinced that there are no important differences between the interventions [27]. This element applies to the comparison of some diagnostic interventions and some procedures in interventional radiology. Strict rules govern when trials can be

stopped on the basis of cumulative results, and important statistical considerations come into play.

Fixed trials.—Alternatively, in a fixed trial, the investigators establish deductively the number of participants (sample size) that will be studied. This number can be decided arbitrarily or can be calculated using statistical methods. The latter is a more commonly used method. Even in a fixed trial, the design of the trial usually specifies whether there will be one or more interim analyses of data. If a clear benefit of one intervention over the other can be shown with statistical significance before all participants are recruited, it may not be ethical to pursue the trial, and it may be prematurely terminated.

Randomized Controlled Trials Classified According to the Level of Blinding

In addition to randomization, the investigators can incorporate other methodologic strategies to reduce the risk of other biases. These strategies are known as "blinding." The purpose of blinding is to reduce the risk of ascertainment and observation bias. An open randomized controlled trial is one in which everybody involved in the trial knows which intervention is given to each participant. Many radiology studies are open randomized controlled trials because blinding is not feasible or ethical. One cannot, for example, perform an interventional procedure with its associated risks without revealing to the patient and the treating physician to which group the patient has been randomized. A single-blinded randomized controlled trial is one in which a group of individuals involved in the trial (usually patients) does not know which intervention is given to each participant. A double-blinded randomized controlled trial, on the other hand, is one in which two groups of individuals involved in the trial (usually patients and treating physicians) do not know which intervention is given to each participant. Beyond this, triple-blinded (blinding of patients, treating physicians, and study investigators) and quadruple-blinded randomized controlled trials (blinding of patients, treating physicians, study investigators, and statisticians) have been described but are rarely used.

Randomized Controlled Trials Classified According to Nonrandomized Participant Preferences

Eligible individuals may refuse to participate in a randomized controlled trial. Other eligible individuals may decide to participate in a randomized controlled trial but have a clear preference for one of the study interventions. At least three types of randomized controlled trials take into account the preferences of eligible individuals as to whether or not they take part in the trial. These are called preference trials because they include at least one group in which the participants are allowed to choose their preferred treatment from among several options offered [32, 33]. Such trials can have a Zelen design, comprehensive cohort design, or Wennberg's design [33–36]. For a detailed discussion of these designs of randomized controlled trials, the reader is directed to the excellent detailed discussion offered by Jadad [2].

The Ethics of Randomized Controlled Trials

Despite the claims of some enthusiasts for randomized controlled trials, many important aspects of health care cannot be subjected to a randomized trial for practical and ethical reasons. A randomized controlled trial is the best way of evaluating the effectiveness of an intervention, but before a randomized controlled trial can be conducted, there must be equipoise—genuine doubt about whether one course of action is better than another [16]. Equipoise then refers to that state of knowledge in which no evidence exists that shows that any intervention in the trial is better than another and that any intervention is better than those in the trial. It is not ethical to build a trial in which, before enrollment, evidence suggests that patients in one arm of the study are more likely to benefit from enrollment than patients in the other arm. Equipoise thus refers to the fine balance that exists between being hopeful a new treatment will improve a condition and having enough evidence to know that it does (or does not). Randomized controlled trials can be planned only in areas of uncertainty and can be carried out only as long as the uncertainty remains. Ethical concerns that are unique to randomized controlled trials as well as other research designs will be addressed in subsequent articles in this series. Hellman and Hellman [37] offered a good discussion on this subject.

Reporting of Randomized Controlled Trials

The Quality of Randomized Controlled Trial Reporting

Awareness concerning the quality of reporting randomized controlled trials and the limitations of the research methods of randomized controlled trials is growing. A major barrier hindering the assessment of trial quality is that, in most cases, we must rely on the information contained in the written report. A trial with a biased design, if well reported, could be judged to be of high quality, whereas a well-designed but poorly reported trial could be judged to be of low quality.

Recently, efforts have been made to improve the quality of randomized controlled trials. In 1996, a group of epidemiologists, biostatisticians, and journal editors published "CONSORT (Consolidated Standards of Reporting Trials)" [38], a statement that resulted from an extensive collaborative process to improve the standards of written reports of randomized controlled trials. The CONSORT statement was revised in 2001 [39]. It was designed to assist the reporting of randomized controlled trials with two groups and those with parallel designs. Some modifications will be required to report crossover trials and those with more than two groups [40]. Although the CONSORT statement was not evaluated before its publication, it was expected that it would lead to an improvement in the quality of reporting of randomized controlled trials, at least in the journals that endorse it [41].

Recently, however, Chan et al. [42] pointed out that the interpretation of the results of randomized controlled trials has emphasized statistical significance rather than clinical importance:

> The lack of emphasis on clinical importance has led to frequent misconceptions and disagreements regarding the interpretation of the results of clinical trials and a tendency to equate statistical significance with clinical importance. In some instances, statistically significant results may not be clinically important and, conversely, statistically insignificant results do not completely rule out the possibility of clinically important effects.

Limitations of the Research Methods Used in Randomized Controlled Trials

The evaluation of the methodologic quality of randomized controlled trials is central to the appraisal of individual trials, the conduct of unbiased systematic reviews, and the performance of evidence-based health care. However, important methodologic details may be omitted from published reports, and the quality of reporting is, therefore, often

used as a proxy measure for methodologic quality. High-quality reporting may hide important differences in methodologic quality, and well-conducted trials may be reported badly [43]. As Devereaux et al. [41] observed, "[h]ealth care providers depend upon authors and editors to report essential methodological factors in randomized controlled trials (RCTs) to allow determination of trial validity (i.e., likelihood that the trials' results are unbiased)."

The most important limitations of research methods include the following:

Insufficient power.—A survey of 71 randomized controlled trials showed that most of these trials were too small (i.e., had insufficient power to detect important clinical differences) and that the authors of these trials seemed unaware of these facts [44].

Poor reporting of randomization—A study of 206 randomized controlled trials showed that randomization, one of the main design features necessary to prevent bias in randomized controlled trials, was poorly reported [45].

Other limitations.—Additional limitations identified by Chalmers [46] were inadequate randomization, failure to blind the assessors to the outcomes, and failure to follow up all patients in the trials.

Intent to Treat

A method to correct for differential dropout rates between patients from one arm of the study and another is to analyze data by the intent to treat—that is, data are analyzed in the way patients were randomized, regardless of whether or not they received the intended intervention. The intent to treat correction is a form of protection against bias and strengthens the conclusions of a study. A detailed discussion of the assessment of the quality of randomized controlled trials was offered by Jadad [2].

In the appraisal of randomized controlled trials, a clear distinction should be made between the quality of the reporting and the quality of methodology of the trials [43].

Recent Randomized Controlled Trials in Radiology

In recent years, randomized controlled trials have become increasingly popular in radiology research. In 1997, for instance, there were only a few good randomized studies in diagnostic imaging, such as the one by Jarvik et al. [47]. Since 2000, the number of good randomized controlled trials has significantly increased in both diagnostic and interventional radiology. Examples of randomized controlled trials in diagnostic imaging include the works of Gottlieb et al. [48] and Kaiser et al. [49]. Examples of interventional randomized controlled trials are the studies by Pinto et al. [50] and Lencioni et al. [51].

Randomized controlled trials are equally important in screening for disease. Our initial experience with breast screening was unfortunate, and controversy over this issue continues to this day [52, 53]. On the other hand, positive developments have occurred, such as the work of the American College of Radiology Imaging Network. Writing for this group, Berg [54] has offered a commentary on the rationale for a trial of screening breast sonography.

Radiologists have a great deal to learn about randomized controlled trials. Academic radiologists who perform research and radiologists who translate research results into practice should be familiar with the different types of these trials, including those conducted for diagnostic tests and interventional procedures. Radiologists also must be aware of the limitations and problems associated with the methodologic quality and reporting of the trials. It is our hope that this article proves to be a valuable source of information about randomized controlled trials.

Acknowledgments

We thank Alejandro Jadad for his support and Monika Ferrier for her patience and support in keeping us on track and for preparing the manuscript.

References

1. Book of Daniel 1:1–21
2. Jadad AR. *Randomised controlled trials: a user's guide*. London, England: BMJ Books, 1998
3. Louis PCA. Research into the effects of bloodletting in some inflammatory diseases and on the influence of tartarized antimony and vesication in pneumonitis. *Am J Med Sci* 1836;18:102–111
4. Hill AB. The clinical trial. *N Engl J Med* 1952; 247:113–119
5. Cochrane Library Web site. Available at: www.update-software.com/cochrane. Accessed September 10, 2004
6. Bree RL, Kazerooni EA, Katz SJ. Effect of mandatory radiology consultation on inpatient imaging use. *JAMA* 1996;276:1595–1598
7. DeVore GR. The routine antenatal diagnostic imaging with ultrasound study: another perspective. *Obstet Gynecol* 1994;84:622–626
8. Fontana RS, Sanderson DR, Woolner LB, et al. Screening for lung cancer: a critique of the Mayo Lung Project. *Cancer* 1991;67[suppl 4]:1155–1164
9. [No authors listed]. Impact of follow-up testing on survival and health-related quality of life in breast cancer patients: a multicenter randomized controlled trial—the GIVIO Investigators. *JAMA* 1994;271:1587–1592
10. Jarvik JG, Maravilla KR, Haynor DR, Levitz M, Deyo RA. Rapid MR imaging versus plain radiography in patients with low back pain: initial results of a randomized study. *Radiology* 1997;204:447–454
11. Kinnison ML, Powe NR, Steinberg EP. Results of randomized controlled trials of low-versus high-osmolality contrast media. *Radiology* 1989;170: 381–389
12. Rosselli M, Palli D, Cariddi A, Ciatto S, Pacini P, Distante V. Intensive diagnostic follow-up after treatment of primary breast cancer. *JAMA* 1994; 271:1593–1597
13. Swingler GH, Hussey GD, Zwarenstein M. Randomised controlled trial of clinical outcome after chest radiograph in ambulatory acute lower-respiration infection in children. *Lancet* 1998;351: 404–408
14. Cochrane Library Web site. Available at: www.update-software.com/abstracts/ab001877.htm. Accessed September 10, 2004
15. Nystrom L, Rutqvist LE, Wall S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials. *Lancet* 1993;341: 973–978
16. Duffy SW. Interpretation of the breast screening trials: a commentary on the recent paper by Gotzsche and Olsen. *Breast* 2001;10:209–212
17. Miettinen OS, Henschke CI, Pasmantier MW, Smith JP, Libby DM, Yankelevitz DF. Mammographic screening: no reliable supporting evidence? *Lancet* 2002;359:404–405
18. Tabar L, Vitak B, Chen HHT, Yen MF, Duffy SW, Smith RA. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer* 2001; 91:1724–1731
19. Hoey J. Does mammography save lives? *CMAJ* 2002;166:1187–1188
20. Norman GR, Streiner DL. *Biostatistics: the bare essentials*, 2nd ed. Hamilton, ON, Canada: B. C. Decker, 2000
21. Silerman WA. Gnosis and random allotment. *Control Clin Trials* 1981;2:161–164
22. Gray JAM. *Evidence-based health care*. Edinburgh, Scotland: Churchill Livingstone, 1997
23. Rosner B. *Fundamentals of biostatistics*, 5th ed. Duxbury, England: Thomson Learning, 2000
24. Norman GR, Streiner DL. *PDQ statistics,* 2nd ed. St. Louis, MO: Mosby, 1997
25. Altman DG, Machin D, Bagant TN, Gardner MJ. *Statistics with confidence*, 2nd ed. London, England: BMJ Books, 2000
26. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994;22:122–124
27. Altman DG. Practical statistics for medical research. London, England: Chapman & Hall, 1991
28. Jadad AR, Rennie D. The randomized controlled

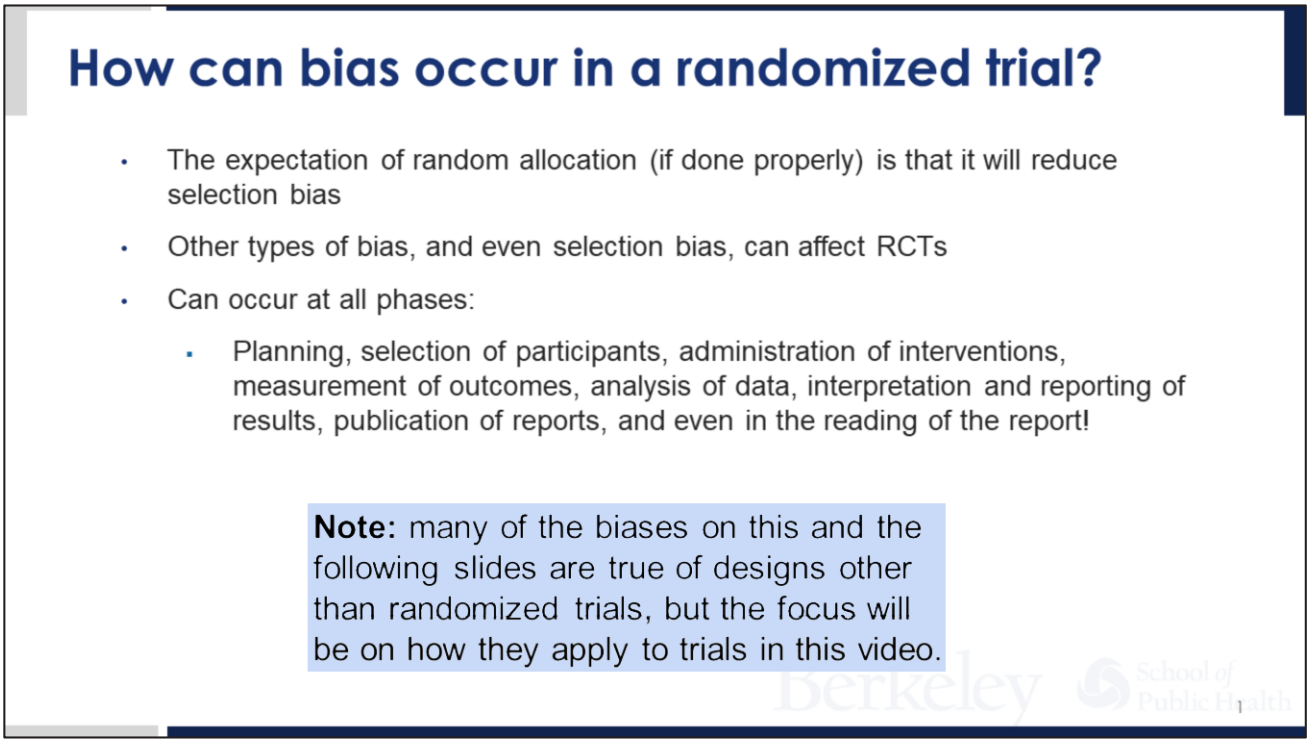
trial gets a middle-aged checkup. *JAMA* 1998; 279:319–320
29. Louis TA, Lavori PW, Bailar JC III, Polansky M. Crossover and self-controlled designs in clinical research. In: Bailar JC III, Mosteller F. eds. *Medical uses of statistics*, 2nd ed. Boston, MA: New England Medical Journal Publications, 1992:83–104
30. Woods KL. Megatrials and management of acute myocardial infarction. *Lancet* 1995;346:611–614
31. Charlton BG. Megatrials: methodological issues and clinical implications. *Coll Phys Lond* 1995; 29:96–100
32. Till JE, Sutherland HJ, Meslin EM. Is there a role for performance assessments in research on quality of life in oncology? *Quality Life Res* 1992; 1:31–40
33. Silverman WA, Altman DG. Patient preferences and randomized trials. *Lancet* 1996;347:171–174
34. Zelen M. A new design for randomized clinical trials. *N Engl J Med* 1979;300:1242–1245
35. Olschewski M, Scheurlen H. Comprehensive Cohort Study: an alternative to randomized consent design in a breast preservation trial. *Methods Inf Med* 1985;24:131–134
36. Brewin CR, Bradley C. Patient preferences and randomized clinical trials. *BMJ* 1989;299:684–685
37. Hellman S, Hellman DS. Of mice but not men: problems of the randomized trial. *N Engl J Med* 1991;324:1585–1592
38. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996;276:7–9
39. Moher D, Schulz KF, Altman DG, CONSORT Group (Consolidated Standards of Reporting Trials). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* 2001;285:1987–1991
40. Altman DG. Better reporting of randomized controlled trials: the CONSORT statement. *BMJ* 1996;313:570–571
41. Devereaux PJ, Manns BJ, Ghali WA, Quan H, Guyatt GH. The reporting of methodological factors in randomized controlled trials and the association with a journal policy to promote adherence to the Consolidated Standards of Reporting Trials (CONSORT) checklist. *Control Clin Trials* 2002;23:380–388
42. Chan KBY, Man-Son-Hing M, Molnar FJ, Laupacis A. How well is the clinical importance of study results reported? an assessment of randomized controlled trials. *CMAJ* 2001;165:1197–1202
43. Huwiler-Müntener K, Jüni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. *JAMA* 2002;287:2801–2804
44. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of Beta, the type 2 error, and sample size in design and interpretation of randomized controlled trials. *N Engl J Med* 1978;299:690–694
45. Schulz KF, Chalmers I, Hayes RJ, Altman DJ. Empirical evidence of bias: dimensions of the methodologic quality associated with estimates of treatment efforts in controlled trials. *JAMA* 1995;273:408–412
46. Chalmers I. Applying overviews and meta-analysis at the bedside: discussion. *J Clin Epidemiol* 1995;48:67–70
47. Jarvik JG, Maravilla KR, Haynor DR, Levitz M, Deyo RA. Rapid MR imaging versus plain radiography in patients with low back pain: initial results of a randomized study. *Radiology* 1997;204: 447–454
48. Gottlieb RH, Voci SL, Syed L, et al. Randomized prospective study comparing routine versus selective use of sonography of the complete calf in patients with suspected deep venous thrombosis. *AJR* 2003;180:241–245
49. Kaiser S, Frenckner B, Jorulf HK. Suspected appendicitis in children: US and CT–a prospective randomized study. *Radiology* 2002;223:633–638
50. Pinto I, Chimeno P, Romo A, et al. Uterine fibroids: uterine artery embolization versus abdominal hysterectomy for treatment—a prospective, randomized, and controlled clinical trial. *Radiology* 2003;226:425–431
51. Lencioni RA, Allgaier HP, Cioni D, et al. Small hepatocellular carcinoma in cirrhosis: randomized comparison of radio-frequency thermal ablation versus percutaneous ethanol injection. *Radiology* 2003;228:235–240
52. Dean PB. Gotzsche's quixotic antiscreening campaign: nonscientific and contrary to Cochrane principles. *JACR* 2004;1:3–7
53. Gotzsche PC. The debate on breast cancer screening with mammography is important. *JACR* 2004;1:8–14
54. Berg WA. Rationale for a trial of screening breast ultrasound: American College of Radiology Imaging Network (ACRIN) 6666. *AJR* 2003;180: 1225–1228

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004

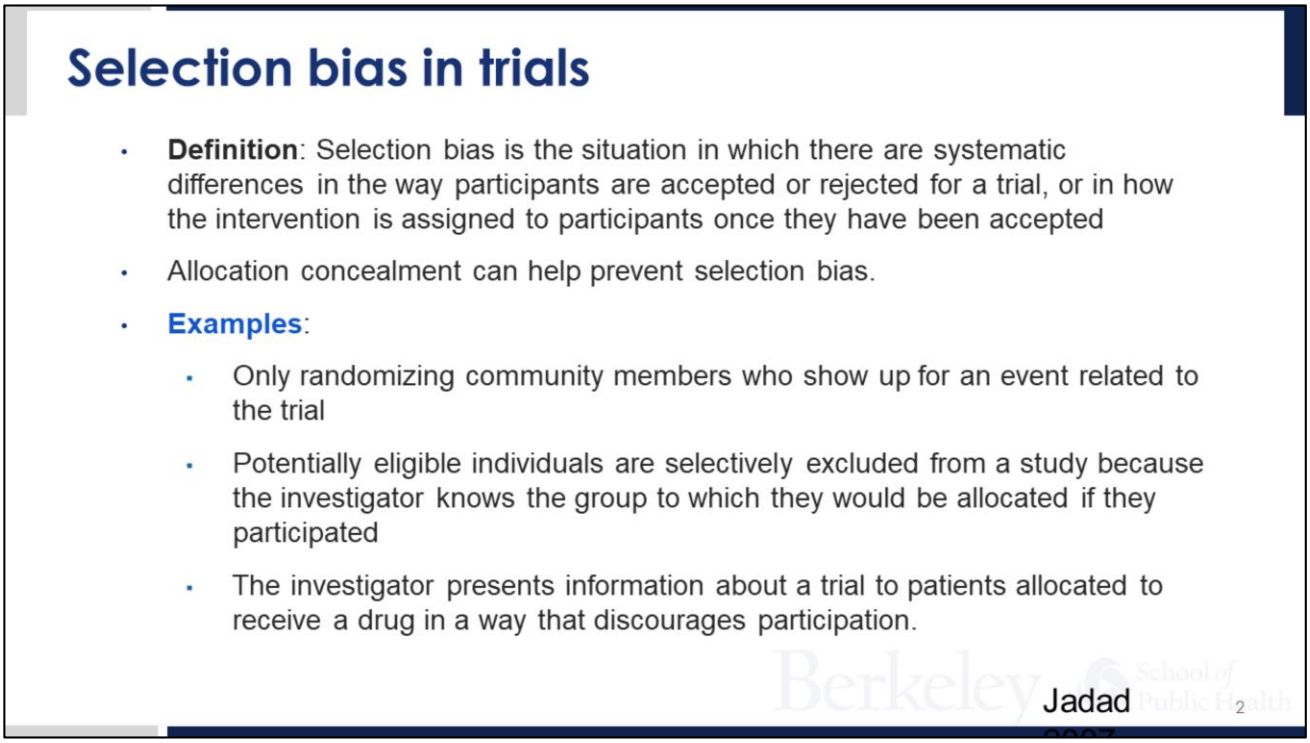
Lecture: Bias in Randomized Trials

PRESENTER: I'd like now to discuss how bias can occur in randomized trials.



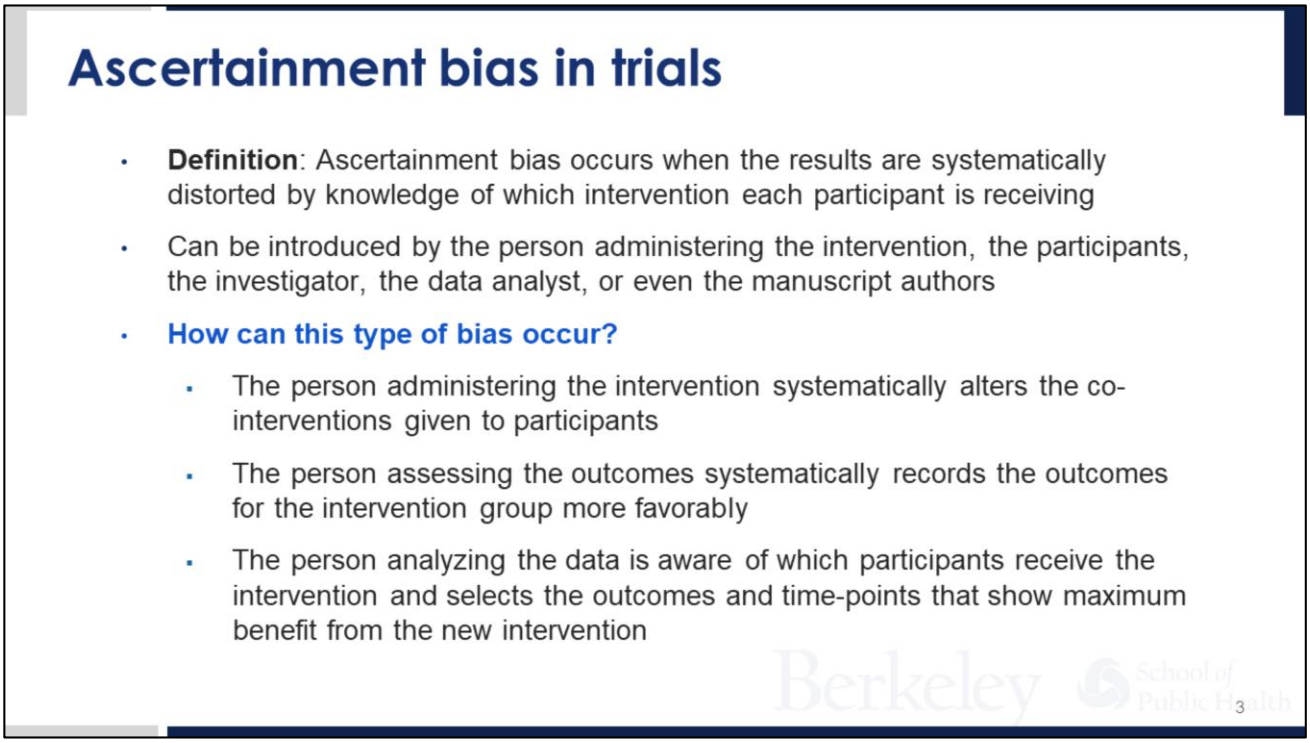
Bias can occur at all phases of a study. It can occur in the planning of the study, the selection of participants, when administration of the interventions is done, when outcomes are measured. During the analysis of data, during interpretation and reporting of results, during the publication of reports, and even in the reading of the report itself, bias can creep in.

Many of the biases on this and following slides in this talk are true and can occur in designs other than randomized trials. But our focus for now will be on how they apply to trials.



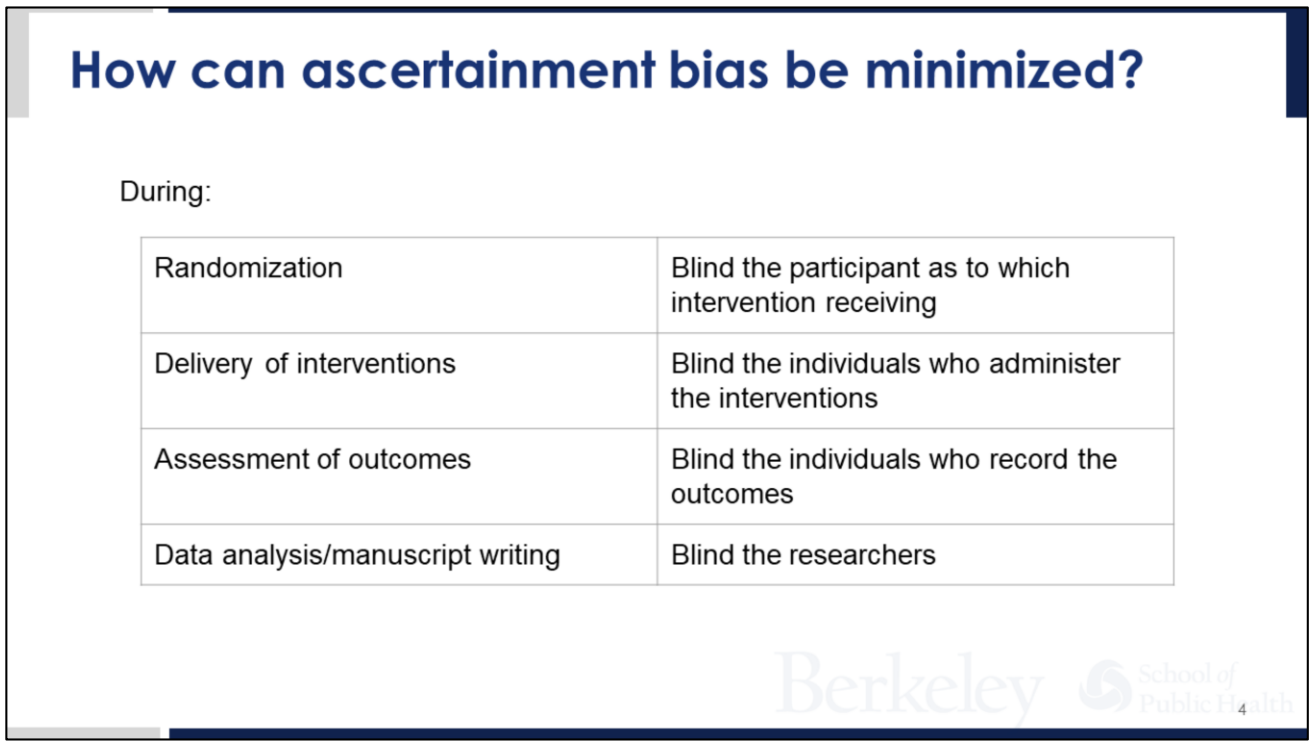
In a trial, selection bias is the situation in which there are systematic differences in the way that participants are accepted or rejected for a trial or in how the intervention is assigned to participants once they have been accepted. Allocation concealment is one tool that can help prevent selection bias.

There are many examples. Imagine only randomizing community members who show up for an event related to the trial. That could create selection bias. Or if potentially eligible individuals are selectively excluded from a study because the investigator knows the group to which they would be allocated if they participated and the investigator didn't want them assigned to that group, that could introduce selection bias. Or if the investigator presents information about a trial to patients allocated to receive a drug in a way that discourages participation, this could also create selection bias.



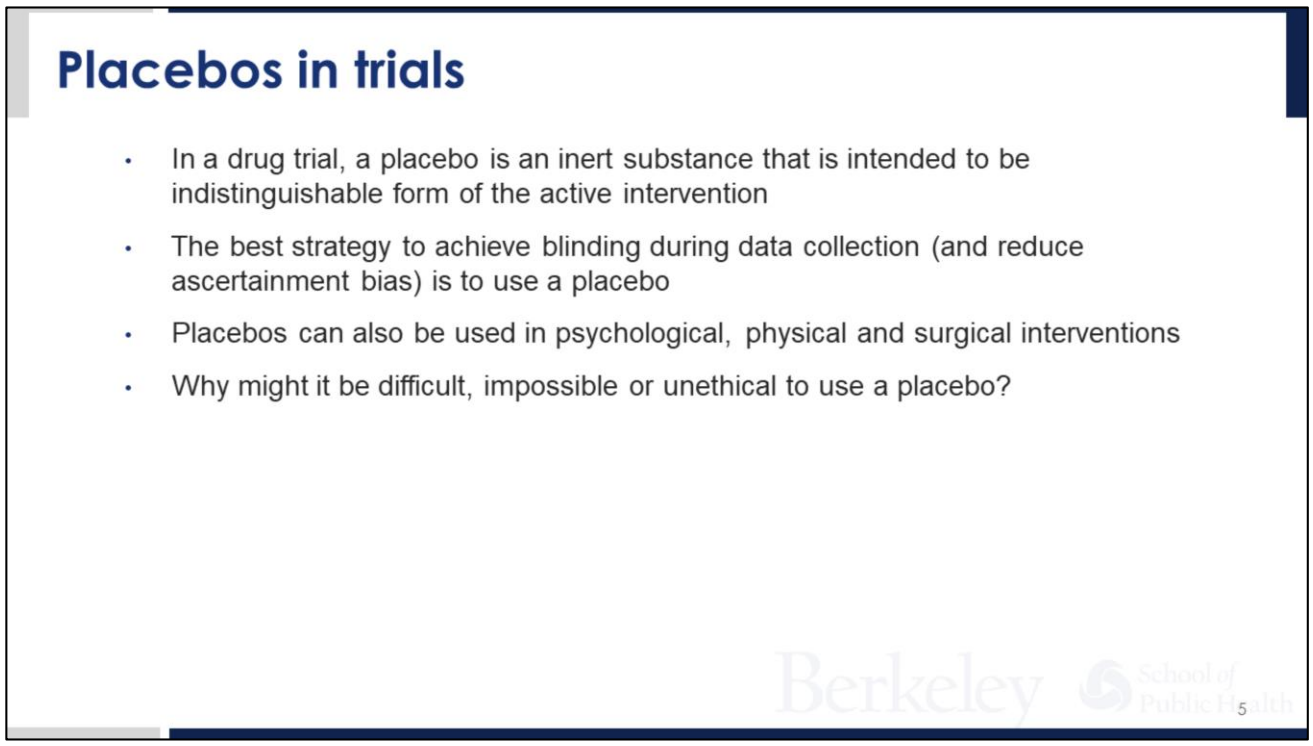
Ascertainment bias is a different situation. This occurs when the results are systematically distorted by knowledge of which intervention each participant is receiving. This can be introduced by the person administering the intervention, the participants themselves, the investigator, the data analyst, or even the manuscript authors.

How can this ascertainment bias occur? If the person administering the intervention systematically alters the co-interventions given to participants because the person administering the intervention knows the assignment, that could introduce ascertainment bias. Or if the person assessing the outcomes systematically records the outcomes for the intervention group more favorably because the person assessing the outcomes is aware what was assigned to that person and wishes that the intervention worked or didn't work, that could lead to ascertainment bias. Another example could be that the person analyzing the data is aware of which participants received the intervention and then selects the outcomes and time points that show the maximum benefit from the new intervention. This might occur if the person analyzing the data subconsciously or consciously really wants the intervention to work.

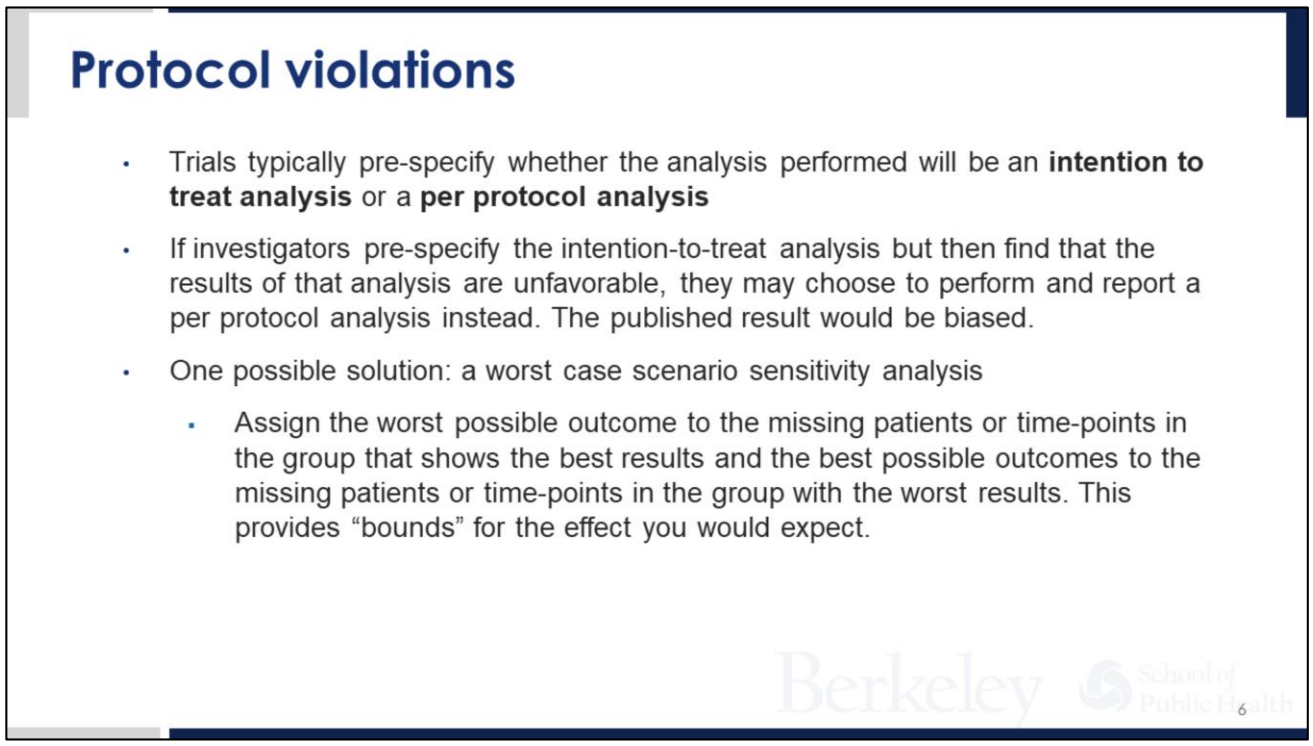


What can be done to minimize ascertainment bias during a trial? During randomization, blinding the participants as to which intervention is being received by the participants can help a lot. In the step of delivery of the interventions, blinding the interventions to who administers the intervention can help a lot.

Assessment of outcomes, the stage in which we measure what happened to each subject during the study-- here, we want to blind the individuals who record the outcomes so they can't be biased by knowledge of what exposure or what assignment each participant received. Finally, during data analysis and manuscript writing, blinding the researchers to what was done can help to minimize bias.

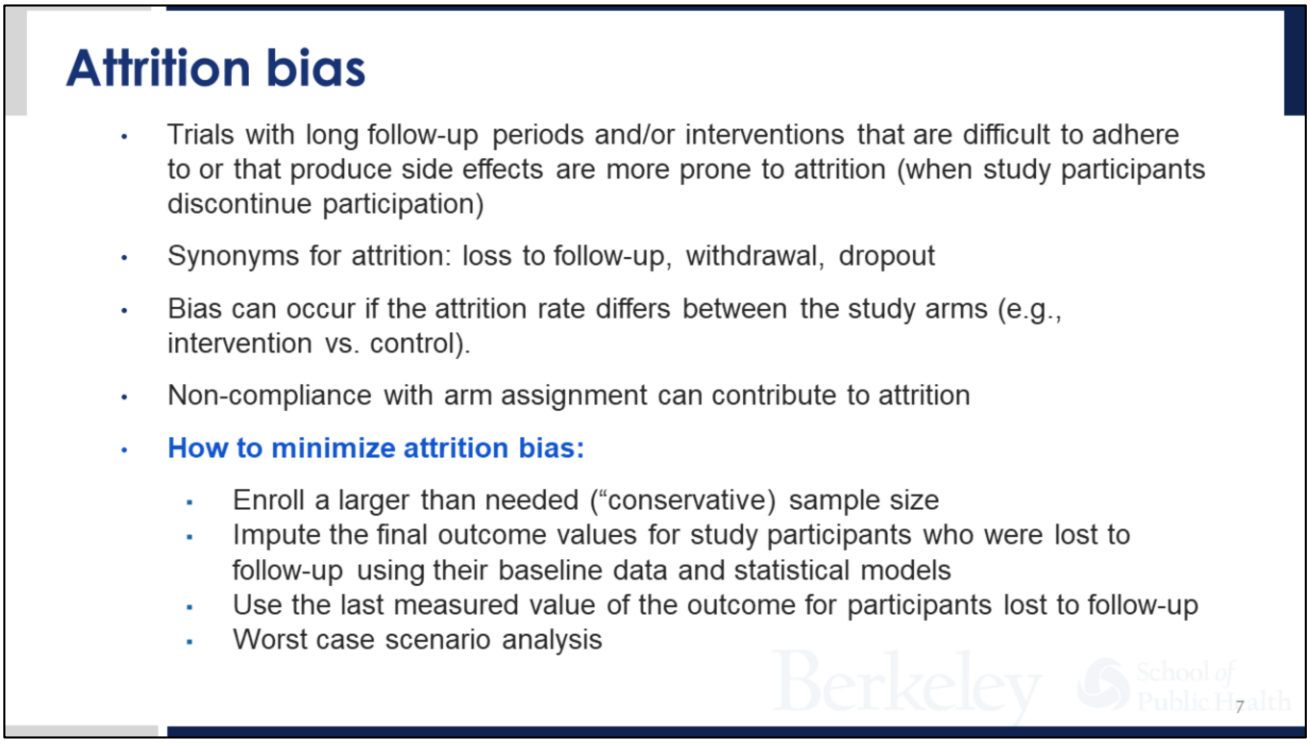


Placebos are an important topic in trials. And in a drug trial, a placebo is an inert substance that's intended to be indistinguishable from the active intervention. It's an indistinguishable form of the pill or whatever is being administered that looks, tastes, and seems insane, but is not the same. It's inert or inactive. The best strategy to achieve blinding during data collection and thereby reduce ascertainment bias is to use a placebo. Placebos can also be used in psychological, physical, and surgical interventions. Think about why it might be difficult, impossible, or unethical to use a placebo.



Trials typically pre-specify whether the analysis performed will be an intention to treat or per protocol analysis. These are two types of analysis you should understand. If investigators pre-specify the intention to treat analysis, but then find that the results of that analysis are unfavorable, they may choose to perform and report a protocol analysis instead. This would result in a biased reporting or a biased result of the study.

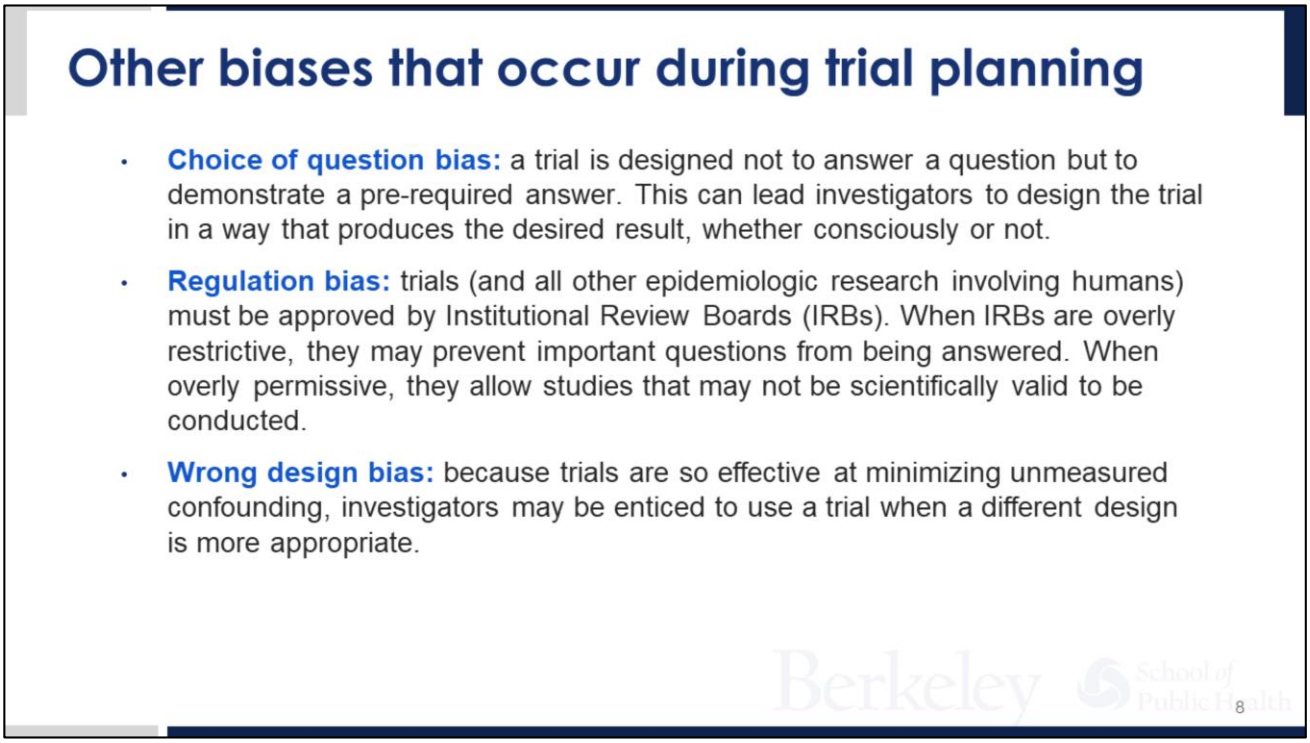
One possible solution, a worst case scenario, would be a sensitivity analysis. Here, we would assign the worst possible outcome to the missing patients or time-points in the group that shows the best results and the best possible outcomes to the missing patients or time-points in the group with the worst results. This provides bounds for the effect you would expect because by making these two groups go in the opposite direction of an effective intervention, it brings the two groups together and leads to minimizing the outcome of the trial, that is, giving you, in a sense, the most conservative result.



Attrition bias is another type of bias. And here, trials with long follow-up periods and/or interventions that are difficult to adhere to, or that produce side effects are more prone to attrition. This is when study participants discontinue participation. There are synonyms for attrition bias, or attrition itself, which include loss to follow-up, withdrawal, dropout.

Bias can occur in these situations if the attrition differs between the steady arms-- that is, intervention versus control. One would look at the attrition rate in each of the arms and see if it differs. Non-compliance with arm assignment can contribute to attrition.

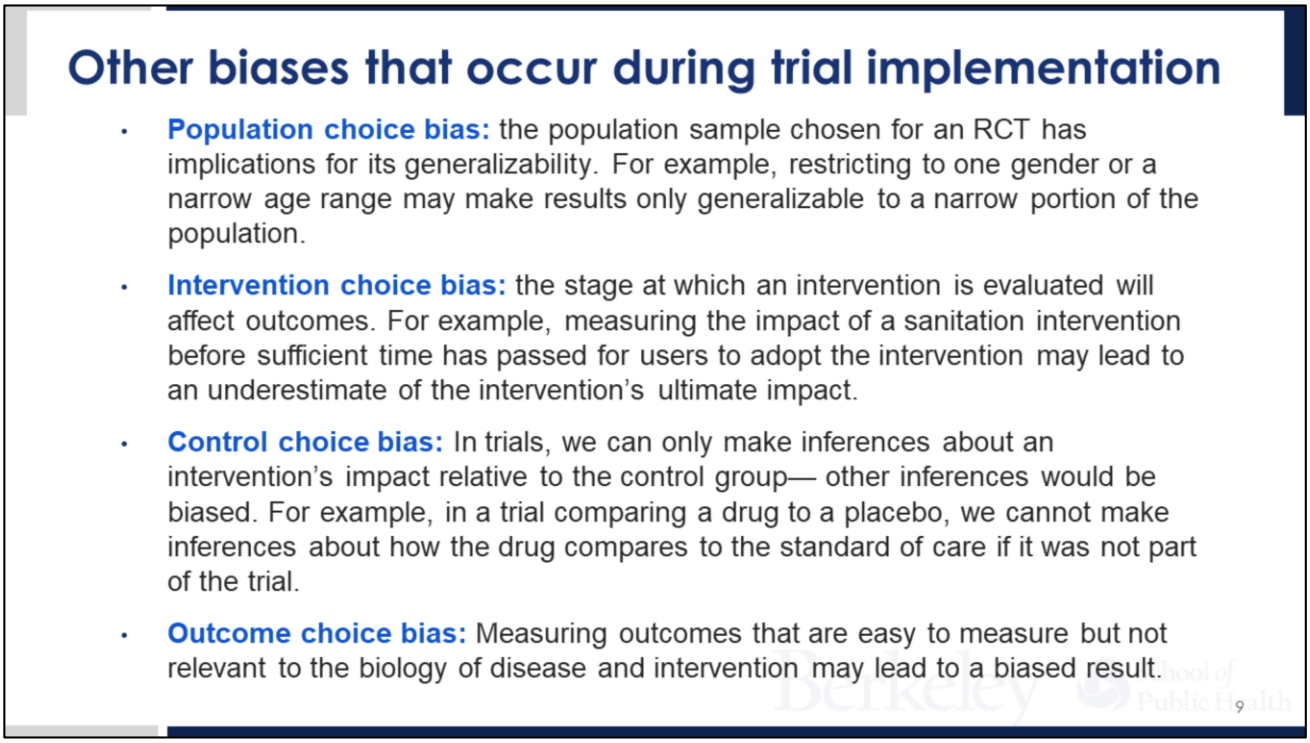
Several ideas or ways to minimize attrition bias include enrolling a larger than needed sample size. We could call that a conservative sample size. It's bigger than we need. Imputing the final outcome values for study participants who were lost to follow-up using their baseline data and statistical models, using the last measured value of the outcome participants who were lost to follow-up as a way to assign them an actual value, and the worst case scenario analysis that I discussed just a little earlier.



A few other biases that can occur during the planning of a trial-- the choice of question bias. In this situation, a trial is designed not to answer a question, but to demonstrate a pre-required answer. This can lead investigators to design the trial in a way that produces the desired result, whether that's conscious or not.

Regulation bias is another type of bias that can occur during trial planning. And this can occur in all epi research involving humans that require approval by the Institutional Review Board. When IRBs are overly restrictive, they may prevent important questions from being answered. When overly permissive, the IRBs might allow studies that may not be scientifically valid to be conducted. All of these situations could be thought about as types of regulation bias.

Finally, wrong design bias-- because trials are so effective at minimizing unmeasured confounding, investigators may be enticed to use a trial when a different design might be more appropriate.

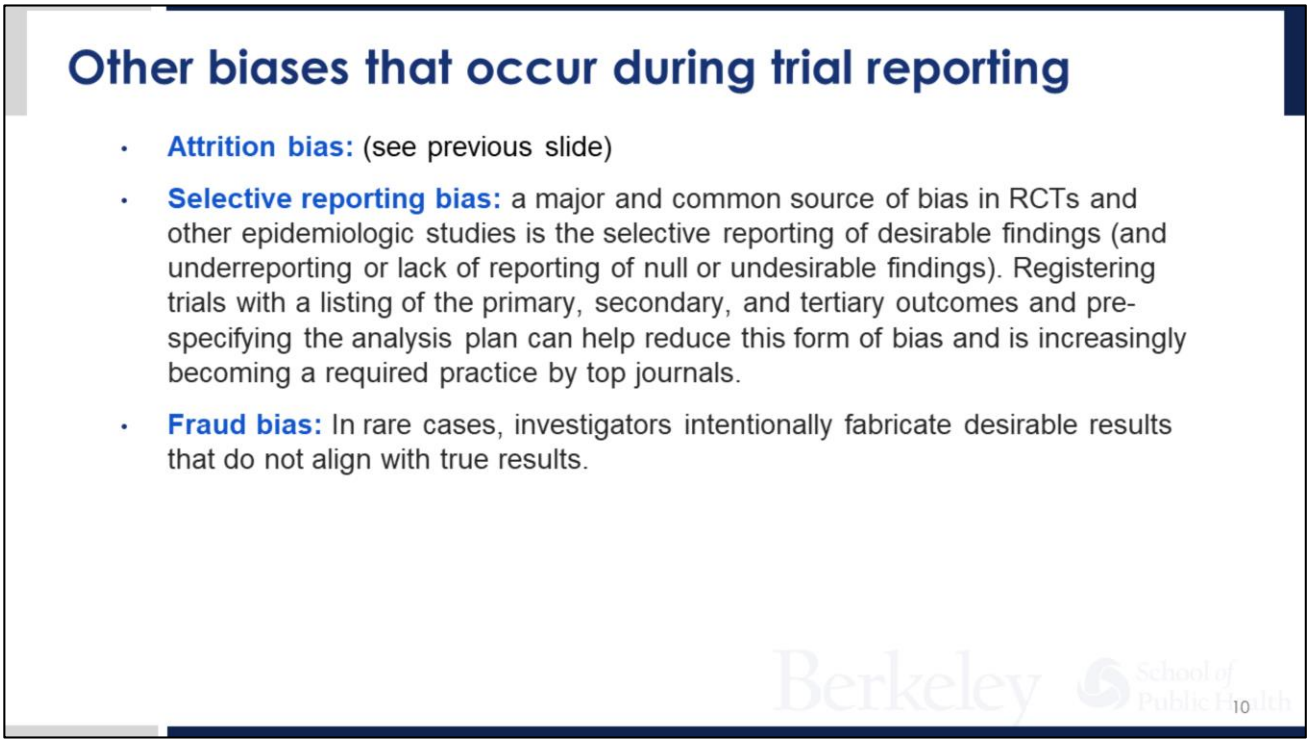


During trial implementation, several types of bias might occur. Population choice bias is the situation in which the population sample chosen for an RCT has implications for its generalizability. For example, restricting to one gender or narrow age range may make the results only generalizable to a narrow portion of the population, thereby introducing population choice bias.

Another type of bias, intervention choice bias, and here, the stage at which an intervention is evaluated will affect outcomes. For example, measuring the impact of a sanitation intervention before sufficient time has passed for the users to adopt the intervention may lead to an underestimate of the intervention's ultimate impact, because had it gone longer, you would have seen more adoption and use and impact.

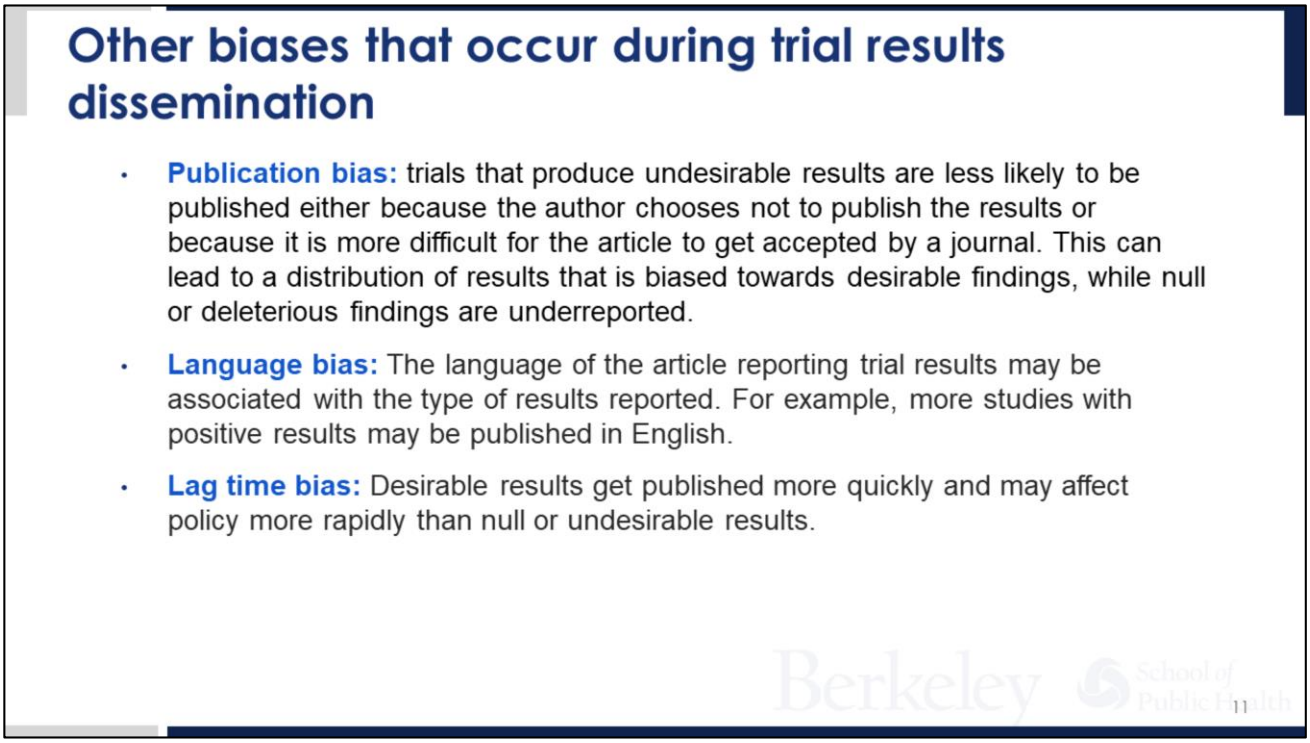
Control choice bias-- in trials, we can only make inferences about an intervention's impact relative to the control group. Other inferences would be biased. For example, in a trial comparing a drug to a placebo, we can't make inferences about how the drug compares to the standard of care if it was not part of the trial. But if we compare it to a placebo, that's not the standard of care. There might be some bias in our conclusions, our inferences, because of whom we chose as controls.

And finally, outcome choice bias-- measuring outcomes that are easy to measure but not relevant to the biology of disease intervention can itself lead to a biased result.



Let's talk about selective reporting bias. This is a major and common source of bias in randomized controlled trials and other epi studies. It's the reporting of desirable findings and underreporting or lack of reporting of null or undesirable findings. Registering trials with a listing of the primary, secondary, and tertiary outcomes and pre-specifying the analysis plan can help reduce this form of bias and is increasingly becoming a required practice by top journals.

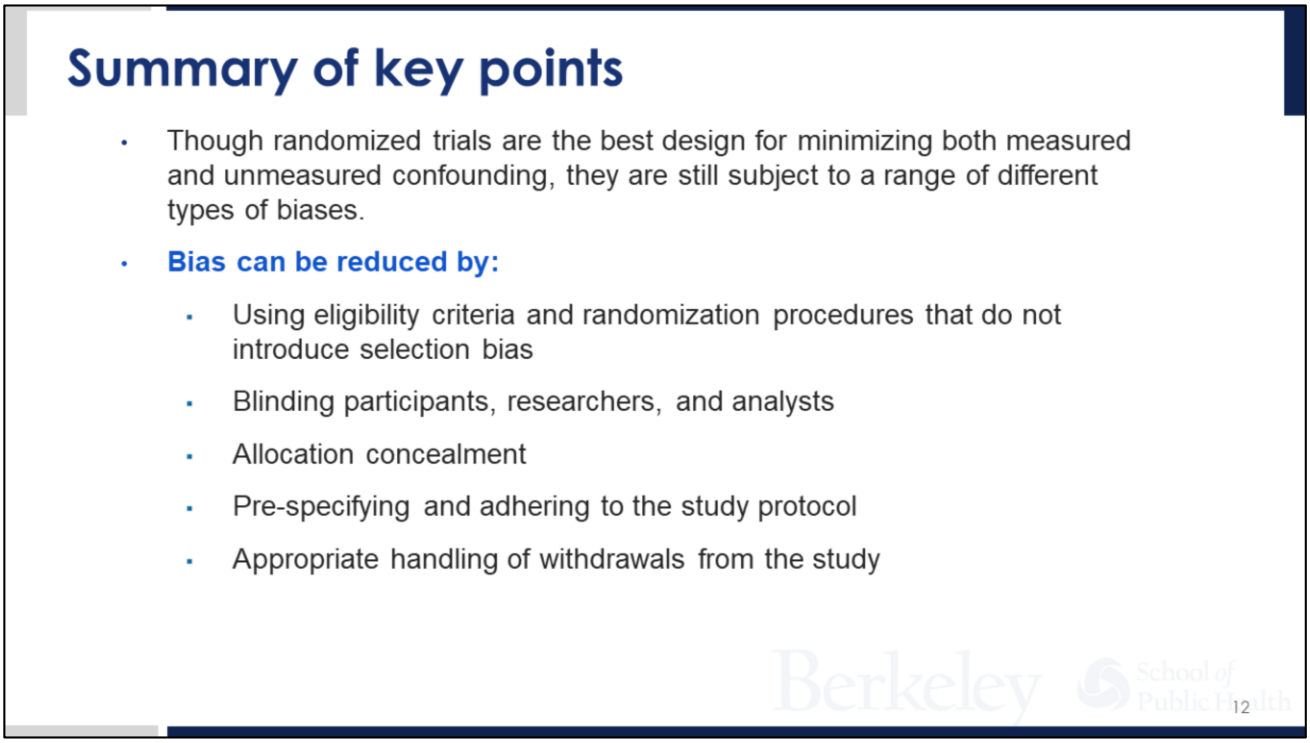
Also fraud bias-- in rare cases, investigators intentionally fabricate desirable results that do not align with the true results.



During dissemination of trial results, a few types of bias can occur. Publication bias is one. Trials that produce undesirable results are less likely to be published, either because the author chooses not to publish the results or because it's more difficult for the article to get accepted by a journal. This can lead to a distribution of results that is biased toward desirable findings, while null or deleterious findings are under-reported.

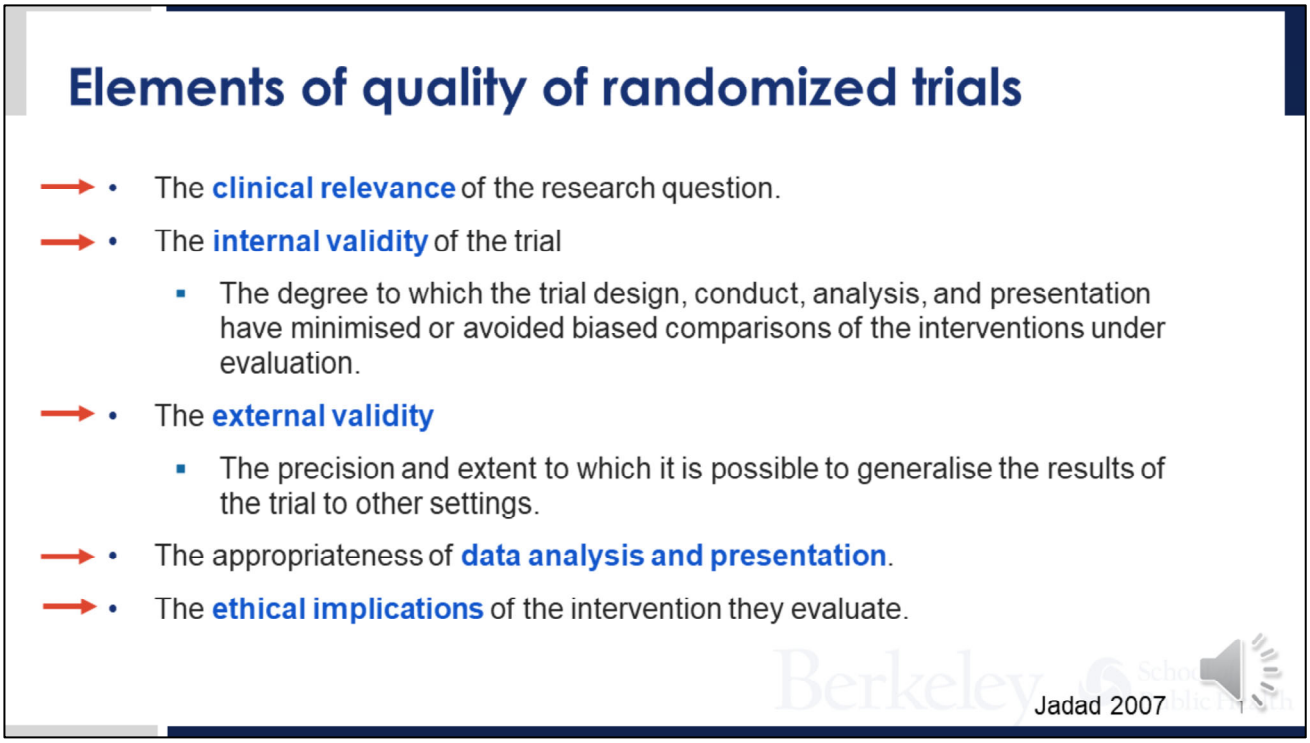
The language bias, the language of the article reporting the trial results may be associated with the type of results that are reported. For example, more studies with positive results seem to be published in English.

Lag time bias-- desirable results get published more quickly and may affect policy more rapidly than null or undesirable results. And you can think of why this happens because editors want to be out there in their journal with important finding.



Though randomized trials are the best design for minimizing both measured and unmeasured confounding, they are still subject to a range of different types of biases. Bias can be reduced by using eligibility criteria and randomization procedures that do not introduce selection bias, blinding participants, researchers, and analysts, choosing to do allocation concealment, pre-specifying and adhering to the study protocol, and making sure to exercise appropriate handling of withdrawals from the study.

In this video, we'll talk about how to evaluate the quality of randomized trials.



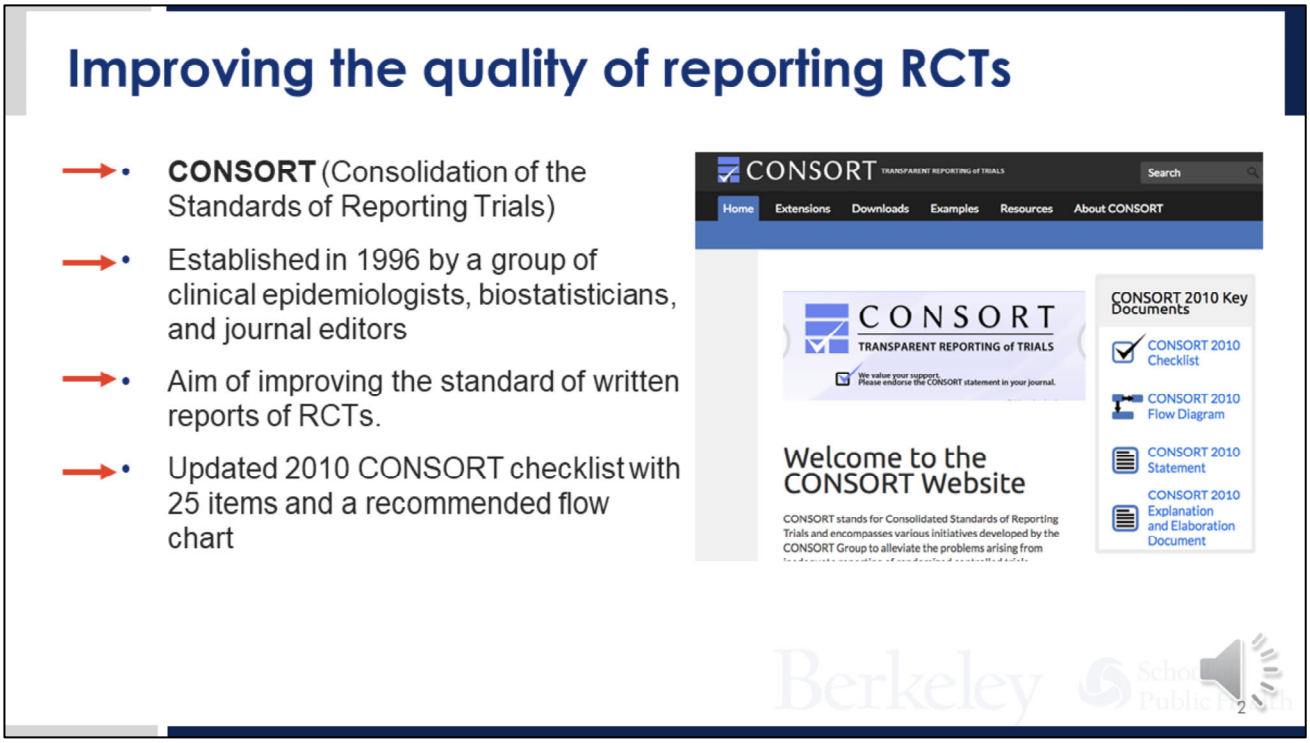
Here's a list of different elements of quality of randomized trials. *First is the clinical relevance of the research question. We only want to perform trials for questions that really need to be answered to improve medical practice or a public health practice. That relevance is an element of quality.

*Internal validity of trials is something we spend a great deal of time thinking about as epidemiologists. This includes the degree to which the trial design, conduct, analysis, and presentation of results have minimized or avoided biased comparisons of the interventions under evaluation. In other words, we want to design and implement our trial in a way that reduces both confounding and numerous types of biases.

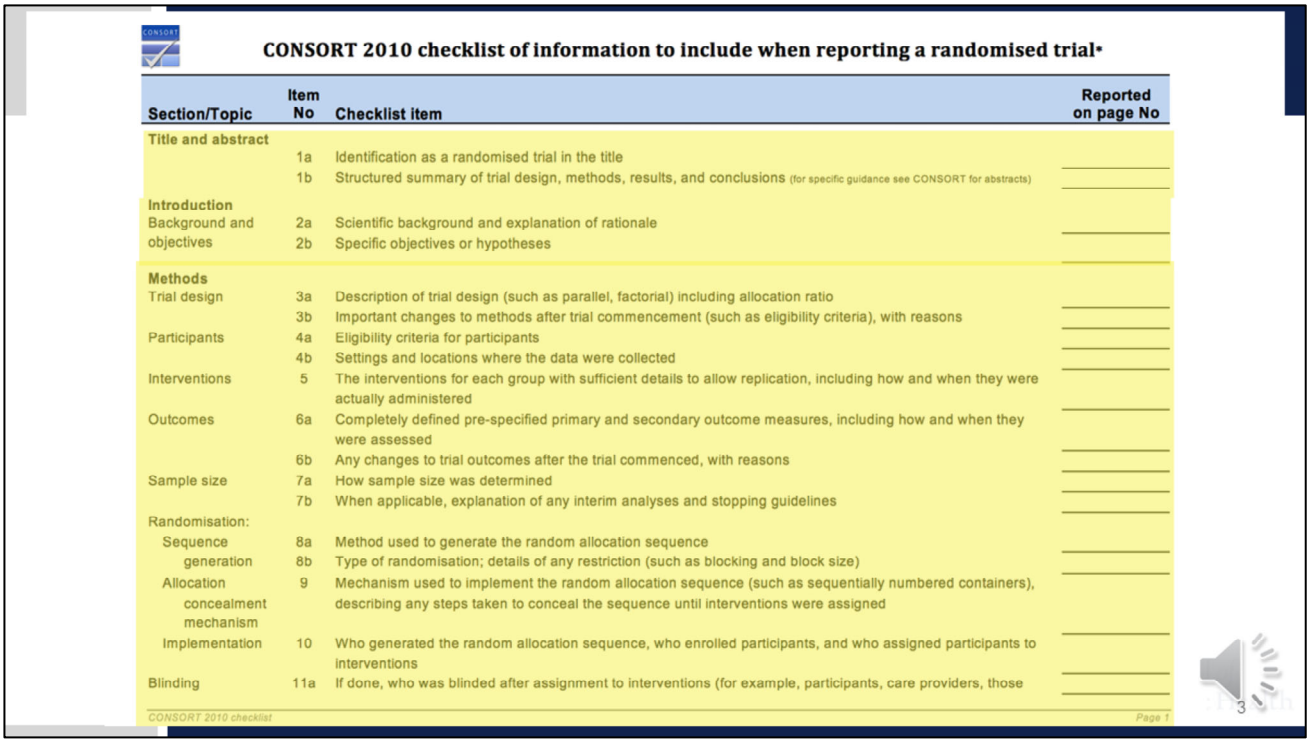
*External validity on the other hand, focuses on the extent to which we can generalize our results from our study population to other important populations.

*The appropriateness of data analysis and presentation are an element of quality. Presentation is usually made in the form of a published manuscript or a presentation in person at a conference.

*Finally, ethical implications of the interventions that are evaluated are another element of quality in randomized trials. But this video will primarily focus on internal and external validity.



*The CONSORT statement stands for the Consolidation Of the Standards Of Reporting Trials. *In 1996, a group of clinical epidemiologists, bio statisticians, and journal editors came together and created the CONSORT to standardize trial reporting. *Prior to this, there was a really wide range of different types of formats used to report trial results. Different studies would be reported with different types of information. So it was difficult for example, if you wanted to know how many people were randomized, how many people were followed up, and how many people were included in the analysis dataset. Prior to CONSORT, that information could take you an hour to find in a paper because it wasn't organized in a standard way across all trials. CONSORT really helped to standardize reporting of trials. *The most updated version of the CONSORT checklist is from 2010. And it includes 25 different items and a recommended flow chart for all authors of randomized trials to use. And many journals these days require that for a trial to be published, they must adhere to the CONSORT checklist and flowchart.

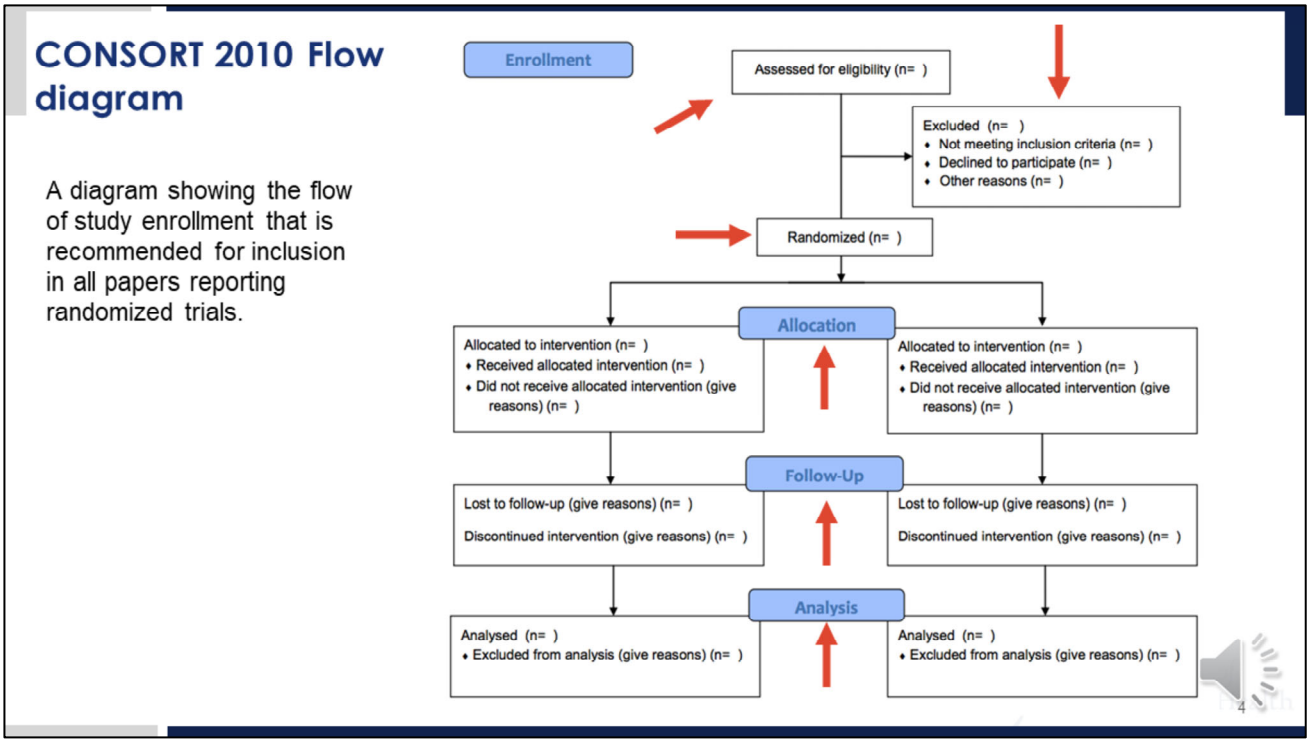


CONSORT 2010 checklist of information to include when reporting a randomised trial•

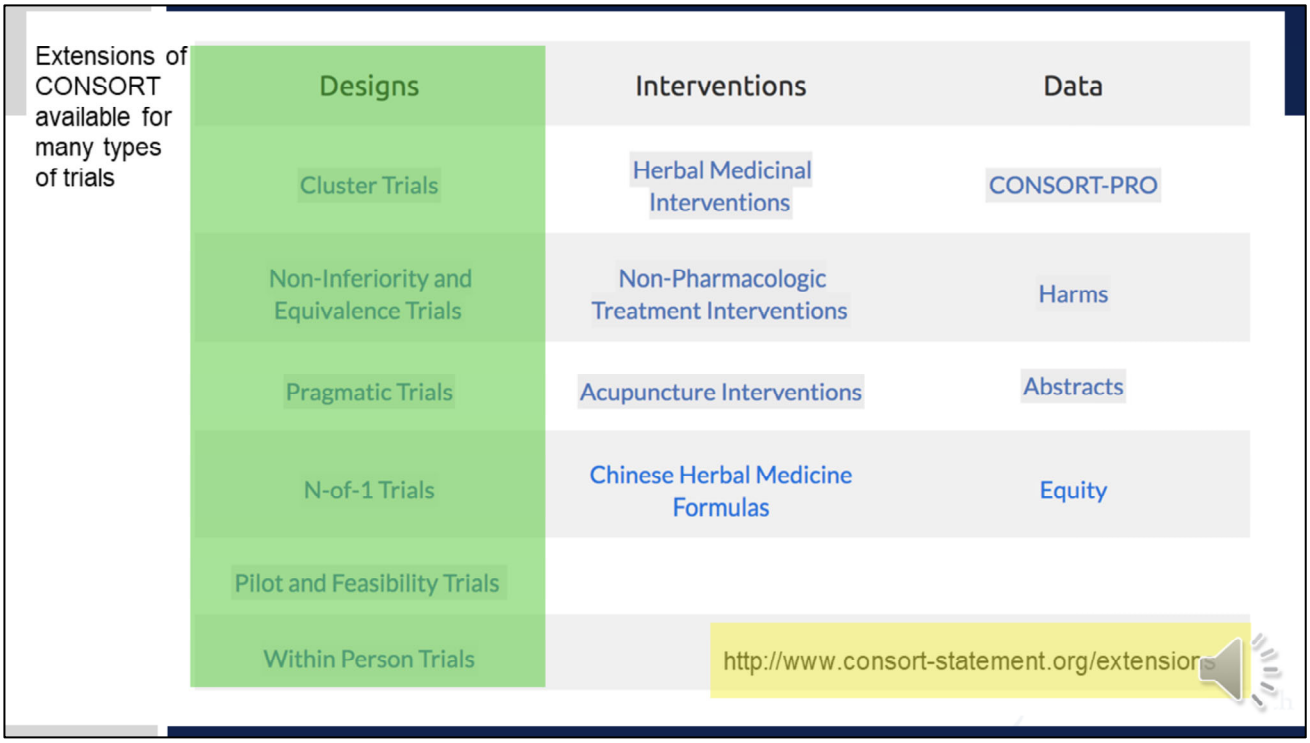
Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	
	2b	Specific objectives or hypotheses	
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	
Participants	4a	Eligibility criteria for participants	
	4b	Settings and locations where the data were collected	
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	
	6b	Any changes to trial outcomes after the trial commenced, with reasons	
Sample size	7a	How sample size was determined	
	7b	When applicable, explanation of any interim analyses and stopping guidelines	
Randomisation:			
Sequence generation	8a	Method used to generate the random allocation sequence	
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those	

CONSORT 2010 checklist Page 1

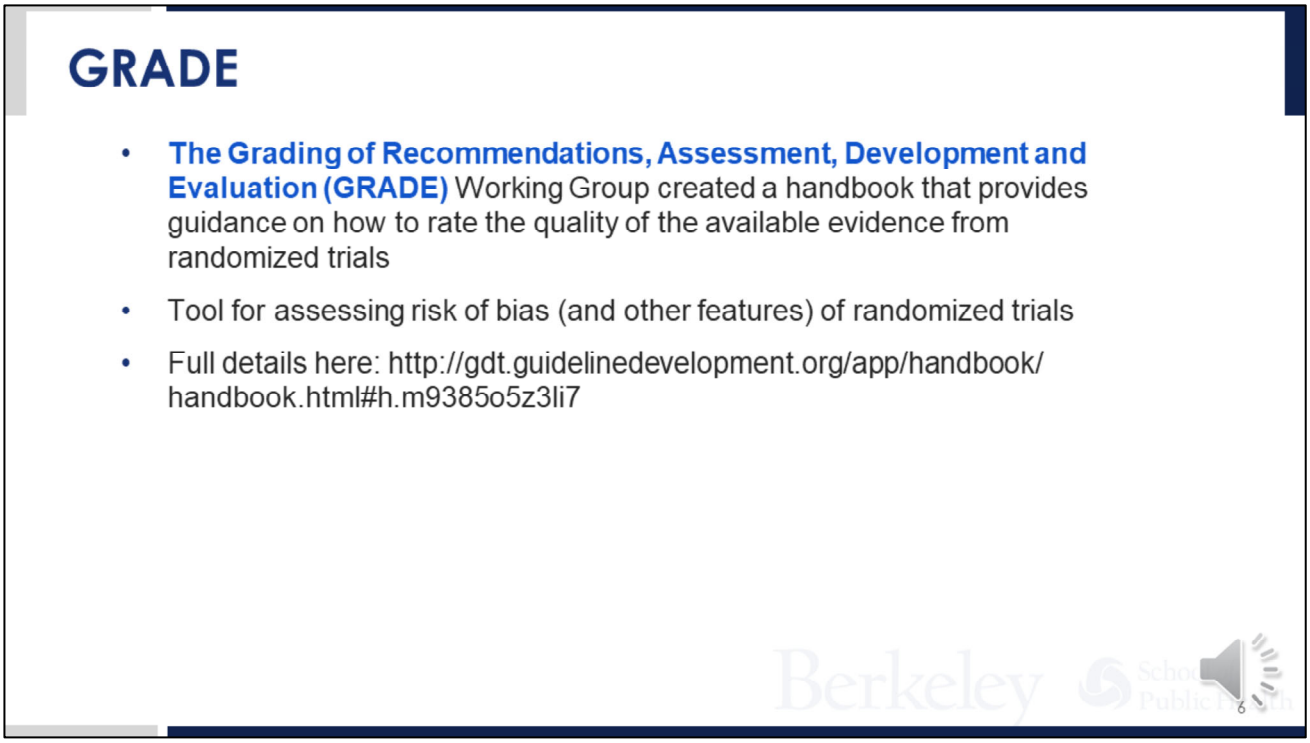
Here's a glimpse of the CONSORT 2010 checklist. We can see that it includes the title and abstract. For example, it's important that the study design be identified in the title and the abstract of the paper should include a certain set of topics. It includes details about the introduction, many details about the methods.They're asking us to report what type of trial design was used. For example, parallel versus factorial or crossover. The allocation ratio, which would be, for example is it 1 to 1, an equal number of intervention and control or it could be 2 to 1. And then the list includes lots of other details to be reported. This checklist is published on the course website and we encourage you to take a look at it in detail on your own.



Here's the flow diagram that is recommended for all randomized trial publications. At the beginning, we see at the top the number of people assessed for eligibility, the number who were excluded, and the number for each type of exclusion criteria listed below, the number randomized, and then the number allocated to each group, those who received the intervention those who did not. *And ideally the author would include the reasons that people who did not have the outcome measured, the number lost to follow up, the number who discontinued, and *then the number analyzed. Having this kind of information in a standard format in all randomized trials that are published today makes it much easier for readers to quickly understand what happened when a study was implemented. It also makes it a lot easier to perform systematic reviews and meta-analyses of trials, which was another major motivation for CONSORT.



And there have been many extensions of CONSORT. This URL here at the bottom right takes you to this page where you can see that they have CONSORT tailored to different design. For example, cluster trials, pragmatic trials, N-of-1 trials. CONSORT for different interventions and then CONSORT for different types of data. And then there are also different reporting guidelines for study designs other than trials, for observational studies and meta-analyses, etc. And we'll come to those at different points in the course as they are relevant

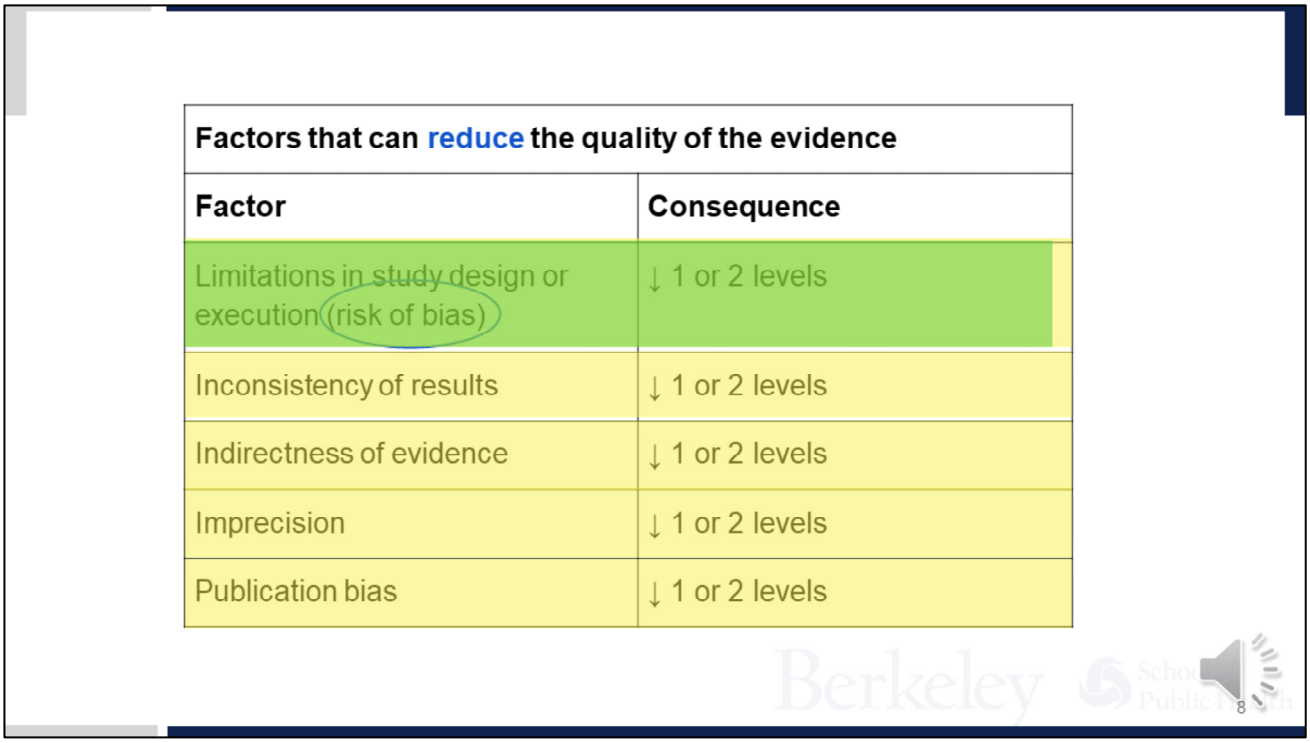


Another really helpful tool that was generated to assess the quality of randomized trials is called up GRADE, Grading of Recommendations Assessment Development and Evaluation. This is essentially a workbook or a handbook that provides guidance to researchers about how to assess the quality of randomized trials in a very standardized way. We're going to focus on the tool they created to assess the risk of bias in randomized trials but there's lots of other features that you can assess. And the URL for the full guidebook is available here.

Quality of Evidence Grades

Grade	Definition
High	We are **very confident** that the true effect lies close to that of the estimate of the effect.
Moderate	We are **moderately confident** in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low	Our **confidence** in the effect estimate is **limited**: The true effect may be substantially different from the estimate of the effect.
Very Low	We have **very little confidence** in the effect estimate: The true effect is likely to be substantially different from the estimate of effect
Uncertain	The article did not provide enough detail to evaluate this criterion.

They use a five point scale. A grade is high quality if we're very confident that the true effect lies close to that of the estimate of the effect in the trial. It's essentially saying we're very confident that the trial results are not biased. And then there's moderate, low, and very low grades as well as an uncertain



Here are the main factors that are assessed in the GRADE system. *The first is limitation in study design or execution, which can increase the risk of bias. And the consequence of identifying these kinds of limitations is that if the study started off at a high level of quality, you might get downgraded one or two levels to moderate or low.

*Inconsistency of results in between different studies on the same topic is definitely an element of quality. And this is more related to systematic reviews, where you're looking at many trials at the same time. You're seeing very inconsistent results with what the prior literature has shown. This can suggest that there might be something really specific to your trial that makes it difficult to generalize findings to other populations.

*Indirectness of evidence is also a factor related to quality. A study may be considered to have indirect evidence if the particular outcomes that were measured in a study are not closely aligned with the true outcomes of interest that we really expect to be affected by an intervention.

*Imprecision is often due to too small of a sample size. So it essentially means that the confidence intervals are too wide to really be able to feel confident about results.
*And then finally, publication bias is a concern that in a particular field of research, only these studies with desirable or positive results have been published. Or even

within a single study, if some results that were prespecified ultimately are not included in the final publication. A study can be downgraded because of this type of bias.

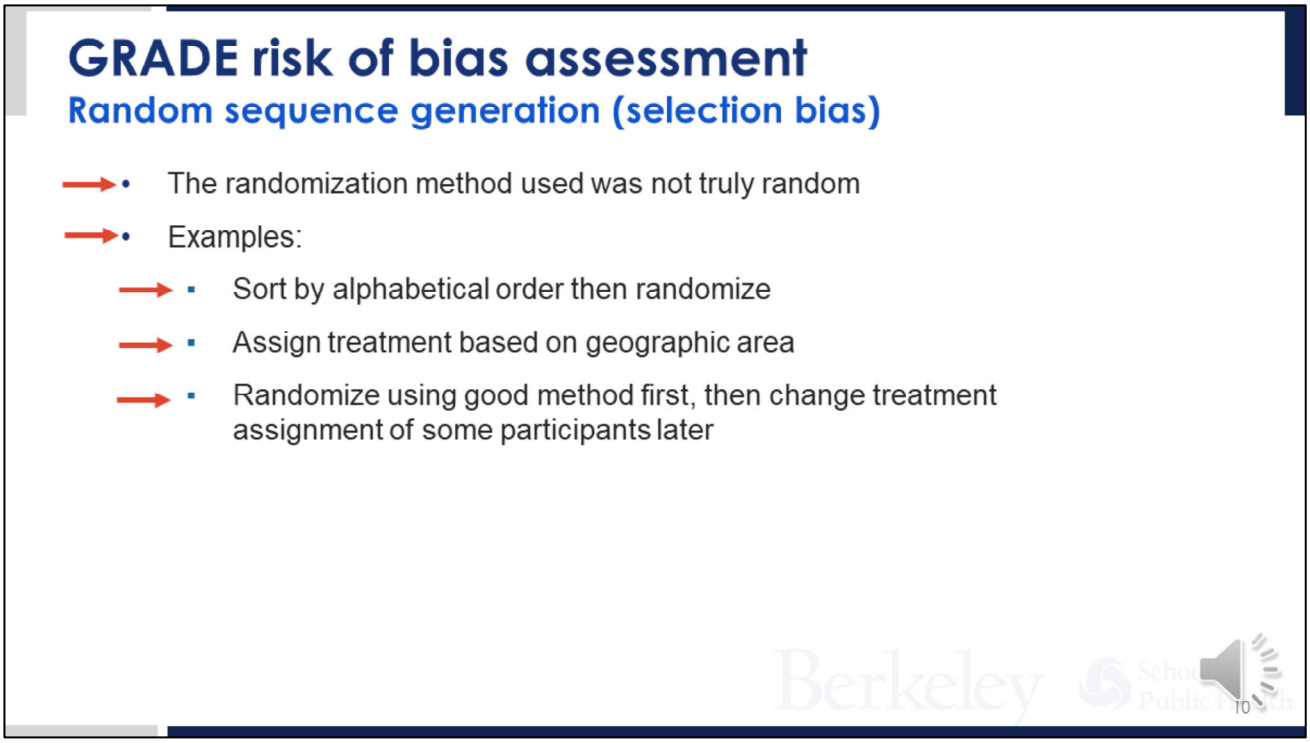
But in this video, we're really focusing on the risk of bias--*this top row here. The remaining criteria are very relevant when it comes to systematic reviews and meta-analyzes and so we'll come back to this later in the course.

Factors that can increase the quality of the evidence	
Factor	**Consequence**
Large magnitude of effect	↑ 1 or 2 levels
All plausible confounding would reduce the demonstrated effect or increase the effect if no effect was observed	↑ 1 level
Dose-response gradient	↑ 1 level

There are also factors that can increase the quality of evidence. *The first is having a large magnitude of effect. This is something that's well-known in epidemiology, and it goes back to Hill's causal criteria. The larger the magnitude of effect, assuming that this study was done well otherwise, generally suggests that it's a higher quality study. This can actually increase the grade by one or two levels. It's difficult to make policies and public health programming decisions off of a borderline or null or small effect, so the quality of evidence is increased if an intervention has a large effect.

*The second is related to what we expect would happen if confounding was present. If we believe that any confounding that occurred would reduce the demonstrated effect or increase the effect if there was no effect observed, than this can increase the quality grading.

*And finally, if there's a dose response pattern, meaning that a higher degree of exposure or a higher level of intervention is associated with either worse or better health outcomes and that pattern follows a linear or other type of association, then we can increase the quality of evidence because it usually signals that there's good biological plausibility.

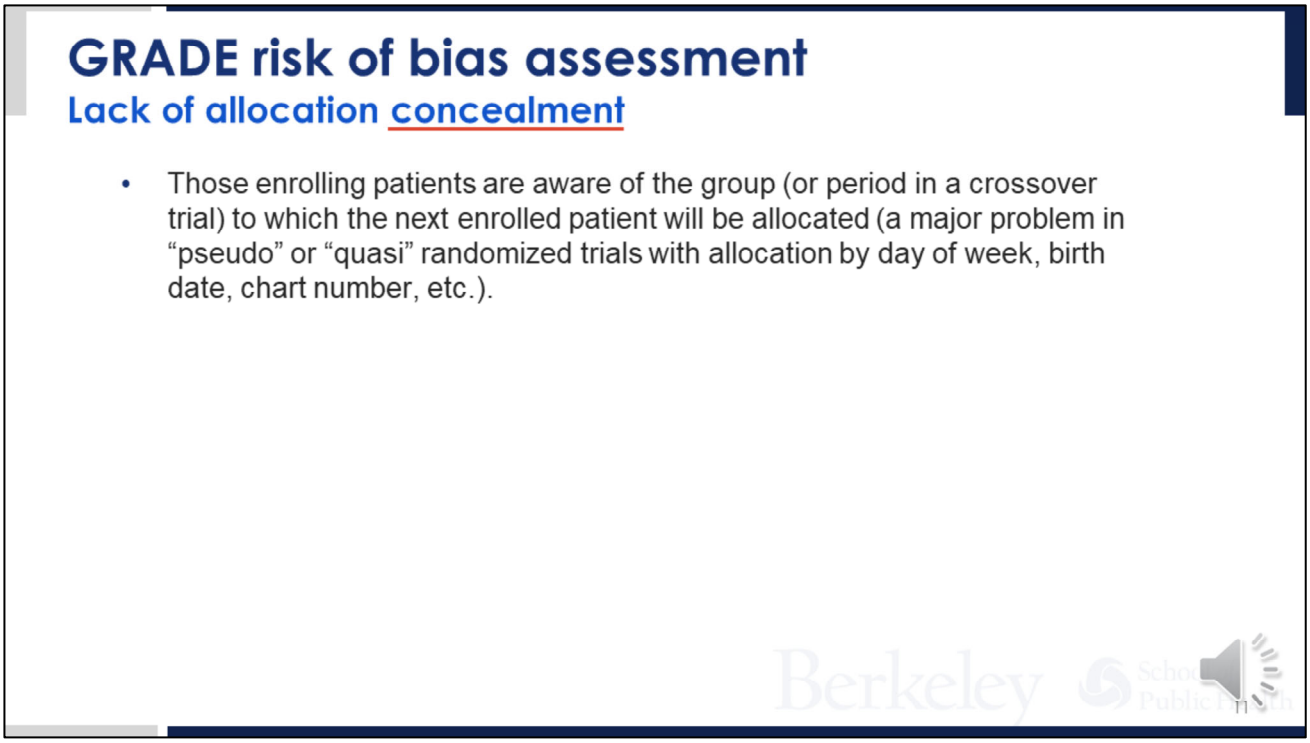


Now, we'll go into detail about the GRADE risk of bias assessment. This is just one of several different criteria that GRADE looks at. *The first item within the risk of bias assessment is related to the random sequence generation and this is also related to selection bias. Essentially, under the GRADE system, we need to evaluate whether a trial used an appropriateness method of randomization that was truly random. If they didn't, then there's concern that there's selection bias in the trial.

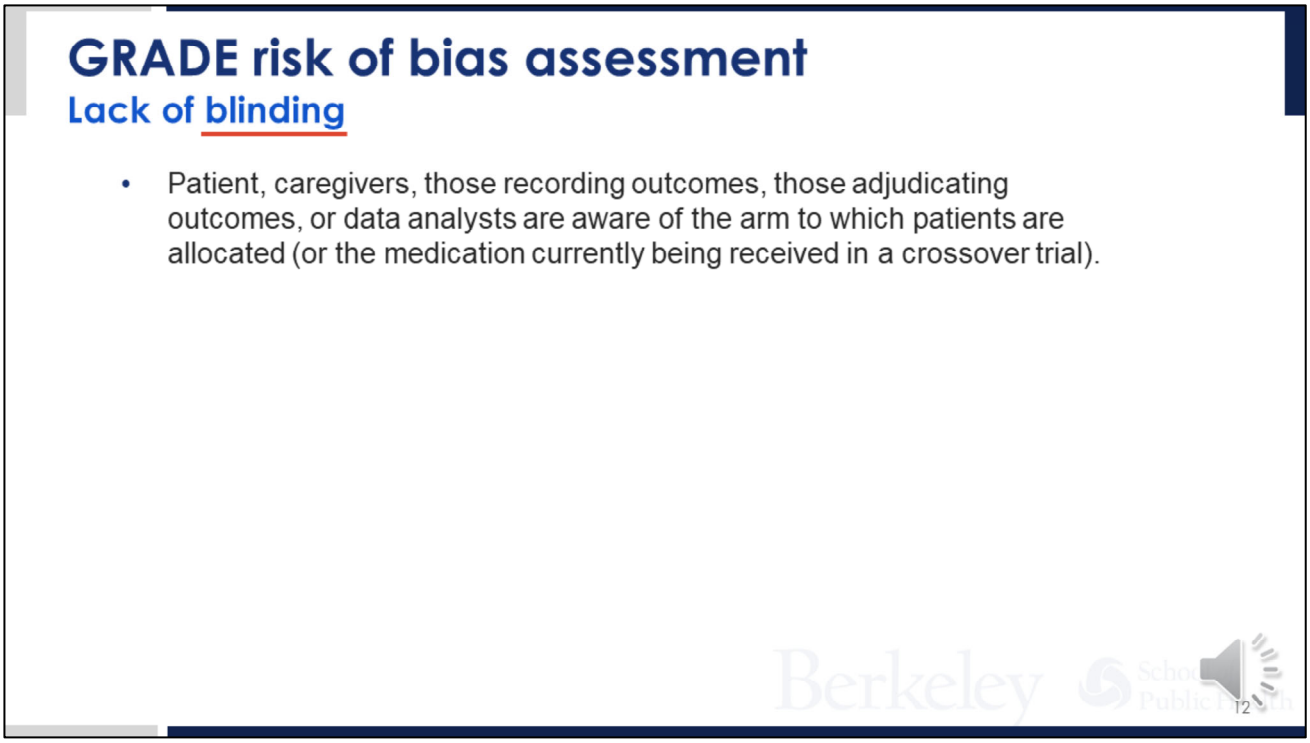
*Examples of things that are not truly random include *sorting the study population by name in alphabetical order and then randomizing. And this is because people with certain first or last names may come from certain ethnicity groups or other types of groups. Randomizing them in this order could cause there to be systematic differences between groups.

*Assigning treatment based on geographic area, which is typically considered convenience sampling, without using any randomization would not be valid.

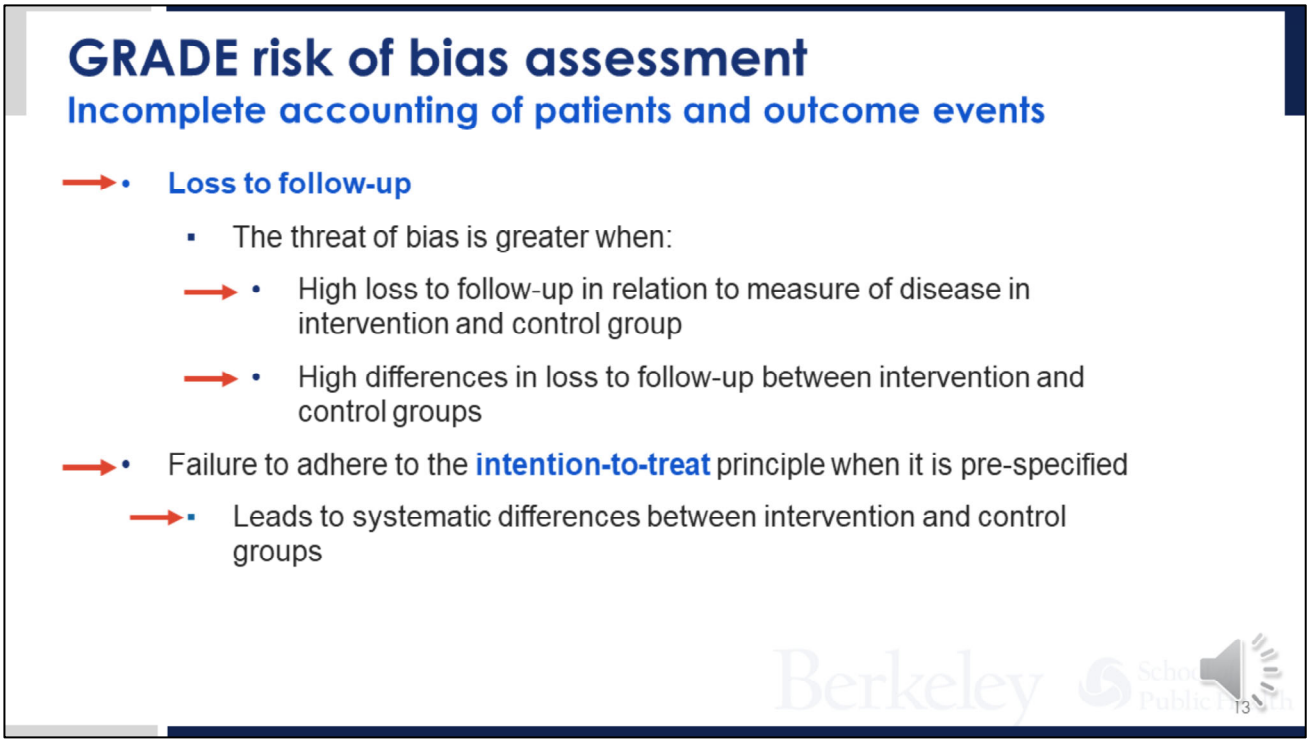
*And then the most common thing we see is a set of investigators randomize participants using a good method--an appropriate, truly random method but then later changes the treatment assignment of some participants for one reason or another. This is not valid and can lead to selection bias.



The second factor is lack of allocation concealment. Once patients or study participants are randomly assigned to treatment groups, it's important that that information be concealed from the person who's enrolling those people into the study. For example, in a clinical setting, it's often common to create a series of envelopes that contain a random sequence of treatment assignment. It wouldn't necessarily be ordered treatment control, treatment control in different envelopes, it would be a random sequence of intervention assignments. And then a clinician would be in charge of identifying potential patients, enrolling them, and then letting them know what their treatment assignment is. If those envelopes were not appropriately concealed, the physician could see which treatment group the next patient might be in and they may decide to not invite that person to participate in the study based on the particular treatment group that's coming up next or they may follow some other path that leads to selection bias.



A lack of blinding can also increase the risk of bias. And this is among patients, study participants, caregivers, people who are assessing outcomes, people who are adjudicating outcomes, data analysts, and all the investigators in a trial. It's really ideal to blind as much as possible, but as we know this isn't always feasible to do. This is a common criterion in which studies may be downgraded in the GRADE system.

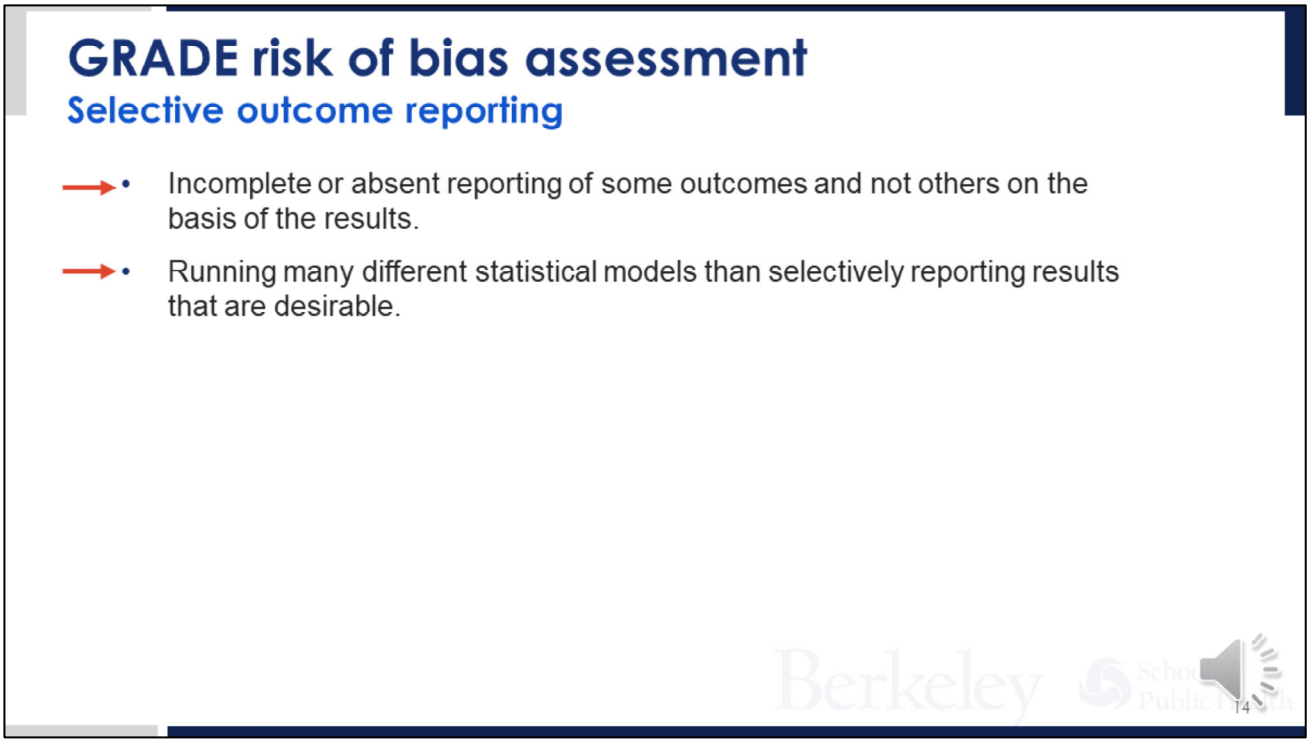


Another risk of bias is related to incomplete accounting of patients and outcome events, *usually due to a loss to follow-up. The threat of bias is greater when *there is a large amount of loss to follow-up in relation to the number of measures of disease in the intervention and control group and *high differences in loss to follow-up between intervention and control. We might see, for example, that people in the control group are much more likely to stop responding to our surveys or stop coming into clinic. And note that those people who stop coming or stop participating are likely to be systematically different from those who stay in the trial. And then, because people were more likely to drop out of the control group, we'll end up seeing differences in characteristics that make it difficult to make valid comparisons between intervention and control groups at the end of the study.

Ideally, a trial would prespecified how they will handle lots of follow-up before it occurs and then follow those steps at the time. This could include doing some kind of imputation, where they take the values that were measured for outcomes of people before they were lost to follow-up and use those during the time periods when they were lost to follow-up. And then there are lots of other different methods and we won't go into detail right now, but the key point is that something must be done to account for this, we can't simply ignore loss to follow-up in the analysis phase.

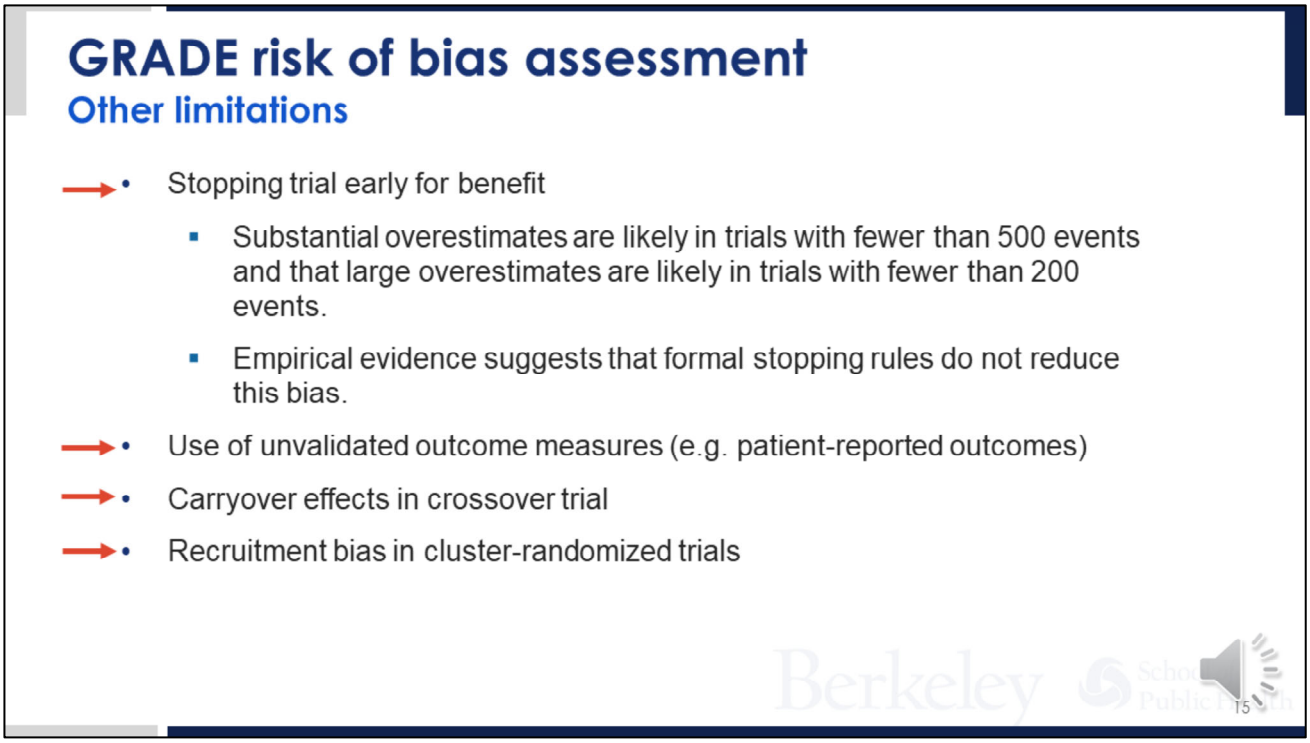
*And then another potential threat to validity is that if we prespecify for example, that we want to use an intention to treat analysis but then we fail to adhere to this at

the time of statistical analysis, this could lead to systematic differences between intervention and control groups. Again, intention to treat means that we're analyzing study participants based on the original randomization. *If we aren't analyzing them based on the original randomization, then there may be confounding and systematic differences between people in the intervention group who adhered to intervention and people in the control group. This essentially is making the study into an observational study, and then we need to use different analytic methods than we would probably use with an intention to treat analysis.



Selective outcome reporting and selective reporting in general is a major threat to validity that's been receiving increasing attention in recent years. *This includes incomplete or absent reporting of certain outcomes but not others on the basis of their results. In other words, if you have, for example, five different outcomes of interest and three of them show a finding that's consistent with your expectation but two show something very confusing that surprises you. If you choose to only report the three that are consistent with your expectation, then there's selected outcome reporting.

*Another really common form of this is to run many different types of statistical models using different sets of adjustment covariates and then selectively reporting the results of the model that are most desirable. This in general, is not a good practice for a randomized trial. For exploratory studies it may be more appropriate.Because randomized trials are designed to test a very specific hypothesis and are generally not exploratory in nature, it's generally a good practice to prespecify the exact statistical model you plan to use and adhere to that regardless of whether it provides you with the result that you want to see.This type of selective outcome reporting can cause the literature to be biased towards desirable findings when there may be harmful effects of interventions or simply null effects of interventions that don't make it to the final publication but that nevertheless are important and should be influencing potential policy and public health programming decisions



Here's a list of other potential limitations– so when stopping trials early for intervention benefits, over estimates of findings are likely in trials with fewer than 500 events, and large over estimates are likely in trials with fewer than 200 events. We want to make sure that in this type of trial, we've enrolled enough people and followed enough events before doing this kind of early stopping. Even when we have formal stopping rules, studies have shown that this kind of bias can still occur.

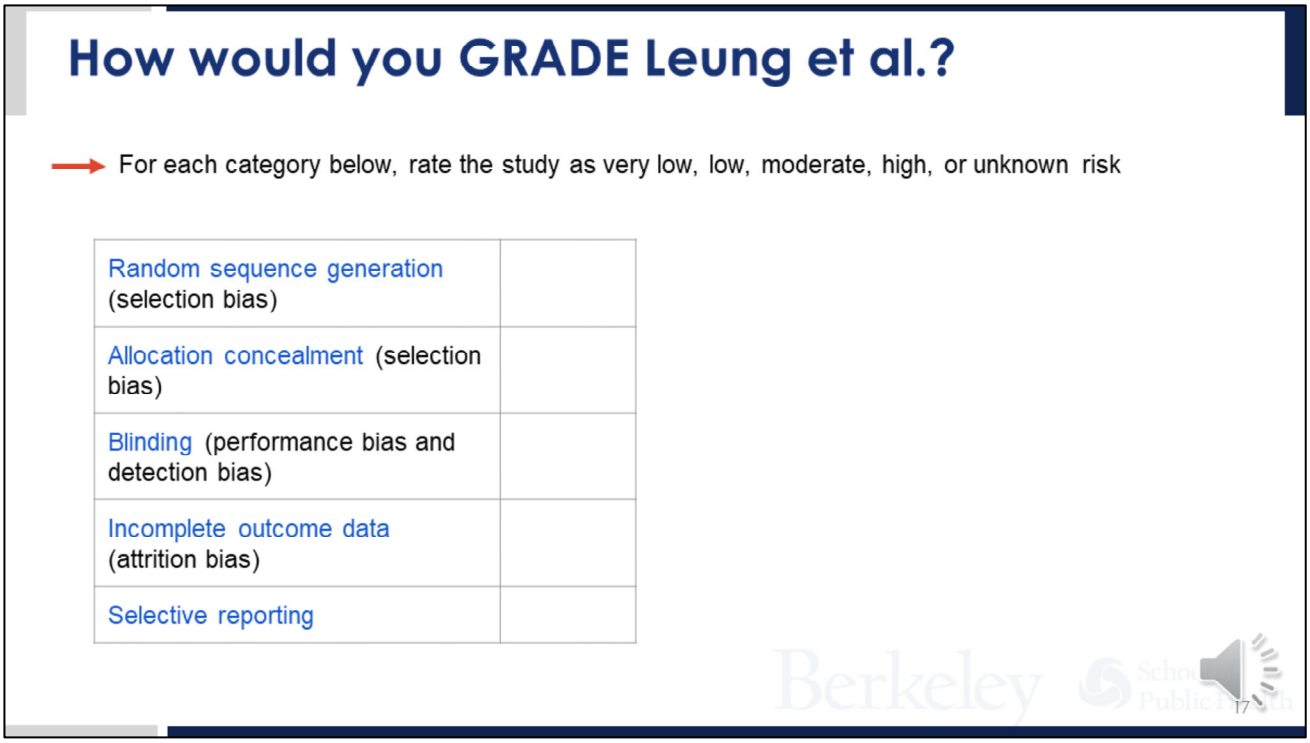
*Another potential limitation is unvalidated outcome measures. For example, patient reported outcomes.In the water and sanitation field of research for many years it's been common to use self-reported diarrhea as a primary outcome in a trial. And this is problematic because people may have imperfect recall of their diarrhea status. They may also feel some pressure to report something desirable to the investigator.Generally speaking, trials that use validated outcome measures, such as measures that are based on biological sample collection, are of higher quality than those that use unvalidated or self-reported outcome measures.

*Another type of limitation would be carryover effects in crossover trials, which we discussed in a previous video. This can occur when there isn't sufficient space or sufficient time between intervention and control periods--one period of time is affecting the next.

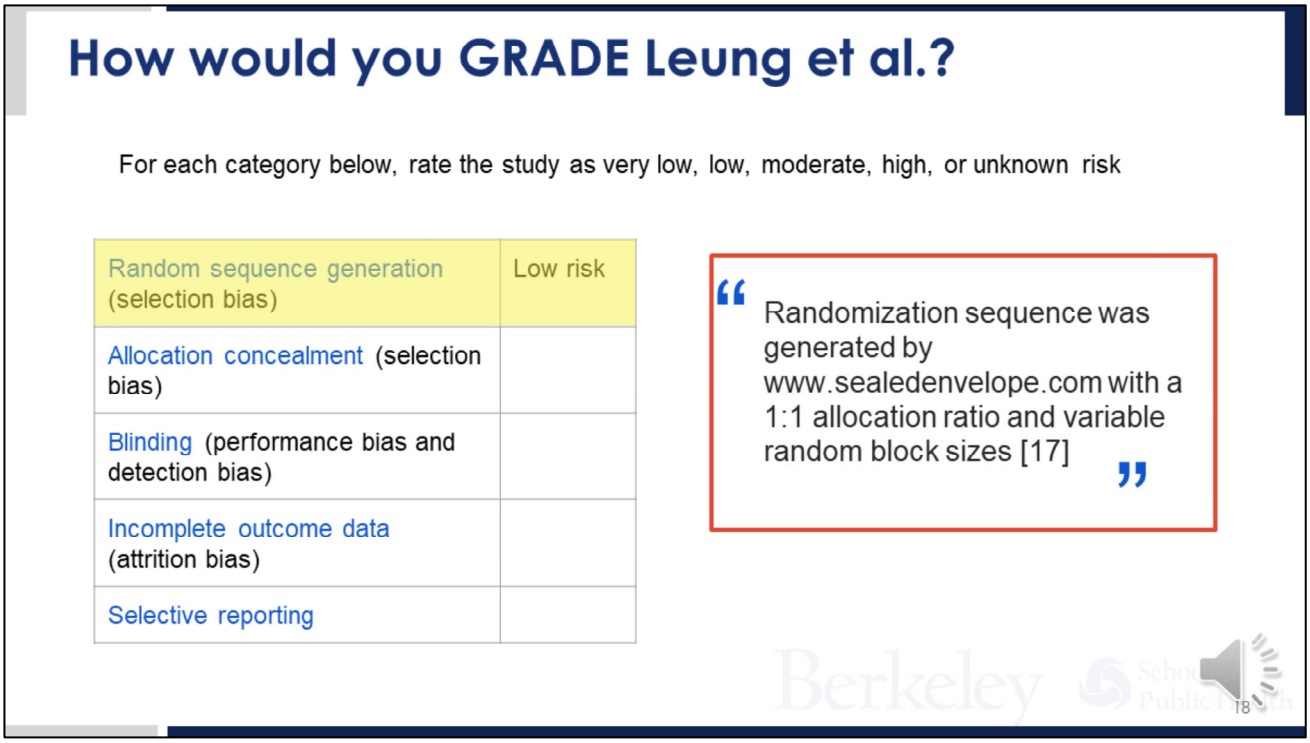
*Finally, in cluster randomized trials we have to be careful of recruitment bias. These kinds of trials sometimes will enroll people at the individual level, but at other

times they'll enroll people in groups. We just have to be careful that if enrollment is occurring at the group level, for example, any town hall style meeting or a village meeting, that the recruitment process is done in a minimally biased way and in a way that's consistent across different treatment and control arms.

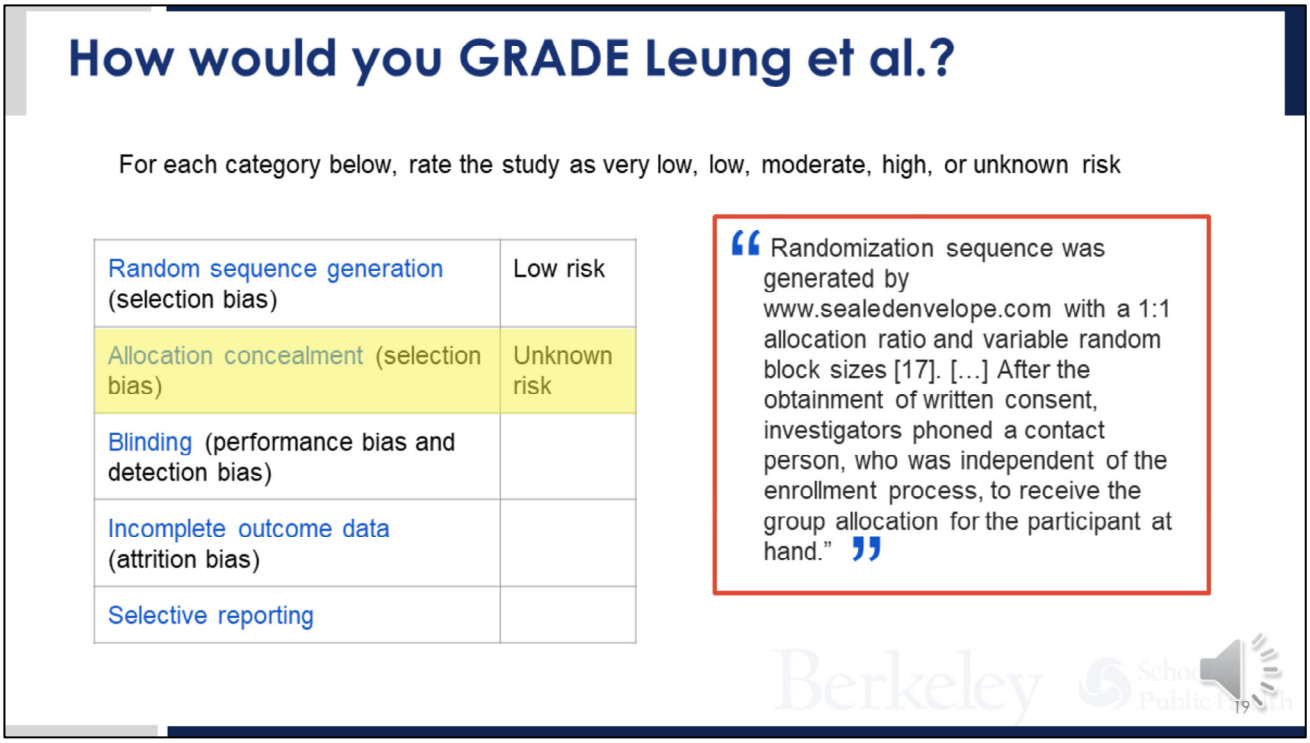
Now, let's practice applying these GRADE criteria for risk of bias to a randomized control trial by Leung et al. This trial was called the impact of patient education on influenza vaccine uptake among community-dwelling elderly. Take a few minutes to review this paper for yourself and think about the different GRADE criteria that we just went over as you're reading the paper. And then in this video, I'll discuss my opinion about what each criterion should be graded.



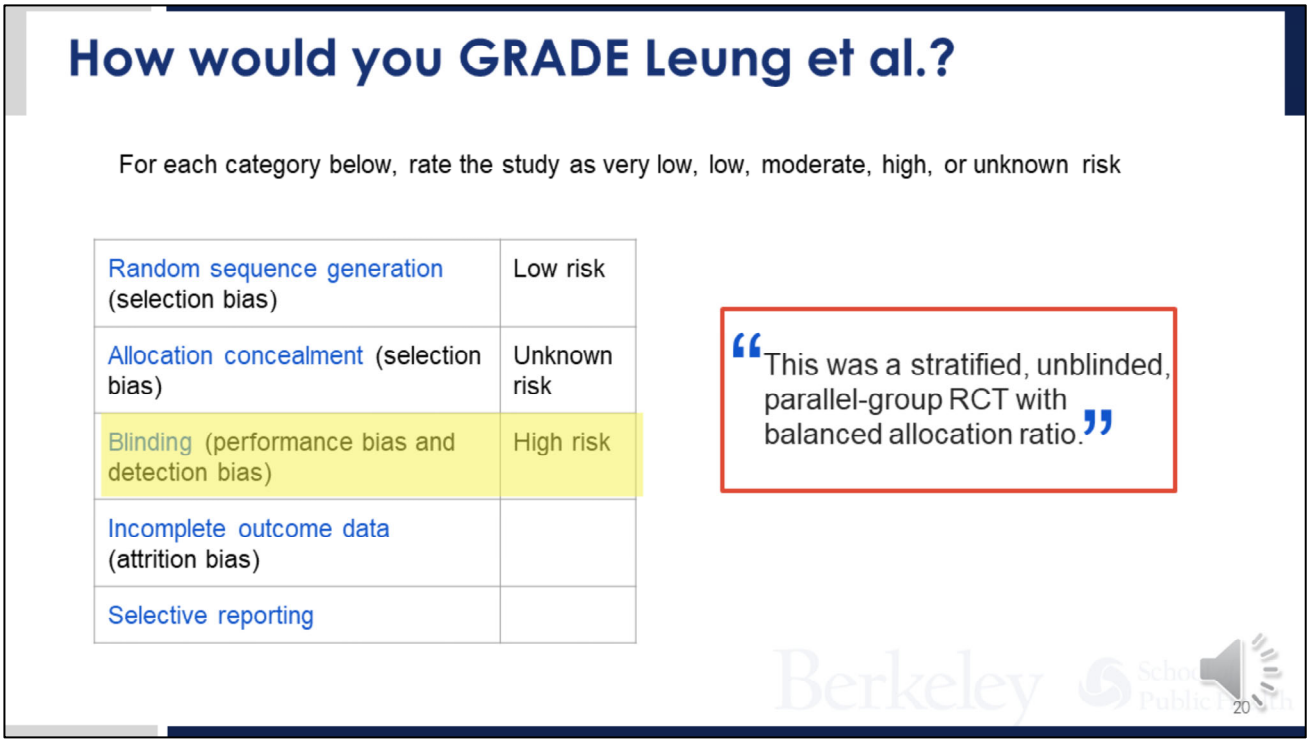
In each of these following categories we need to rate it as very low, low, moderate, high or unknown risk. Hit pause for now, read the article. And then when you're ready, press play again and we'll go over the grading in each of these categories.



For random sequence generation I classified this as low risk. Here's a quote from the paper that helped me with my determination--"Randomisation sequence was generated by www.sealedenvelope.com a 1 to 1 allocation ratio and variable random block sizes.“ So there's a lot of different software available online; the study used one, and there a different software packages, such as R or Stata. And generally speaking, these are very valid ways of generating a random sequence. I consider this to be low risk. It also could potentially be very low risk. I personally haven't investigated sealedenvelope.com. But I would need to do so to decide whether to make it very low instead of low risk.



For allocation concealment I classified this as unknown risk. And this is because the article doesn't provide enough detail for me to assess whether this is low or high risk. Here's a little more of that sentence from the last slide. It says, "After the obtainment of written consent, investigators found a contact person who is independent of the enrollment process to receive the group allocation for the participant at hand."It doesn't say much about who this contact person was, what kind of access to information they had. A higher degree of detail would be needed in order to assess if the allocation concealment was valid.

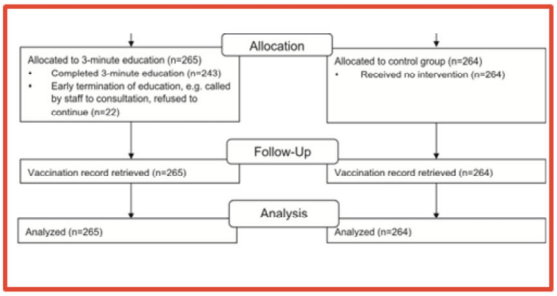


For blinding, I classify this as high risk because the trial was not blind at all. And the authors are very transparent about this. Here is a sentence--they say it's a stratified, unblinded, parallel-group RCT. As I mentioned before, blinding is a category in which trials often get a poor quality rating, but it's something that, from a logistical standpoint, isn't always possible. So we just do the best that we can.

How would you GRADE Leung et al.?

For each category below, rate the study as very low, low, moderate, high, or unknown risk

Random sequence generation (selection bias)	Low risk
Allocation concealment (selection bias)	Unknown risk
Blinding (performance bias and detection bias)	High risk
Incomplete outcome data (attrition bias)	Low risk
Selective reporting	



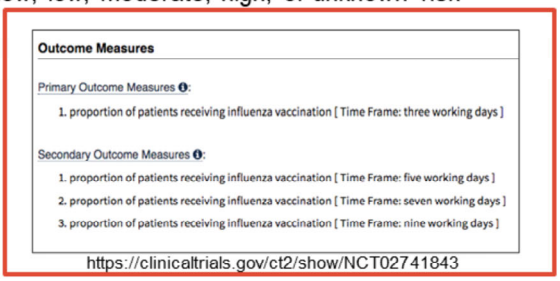
21

Next is incomplete outcome data. Here's a snapshot of the CONSORT flow diagram from this article. And I classified this as low risk because you can see that the number of people who were allocated to intervention or control is the same number of people who are ultimately analyzed. It's 265 on the left hand side of the flow chart for the intervention group and 264 on the control group. There was no loss to follow-up, which means they have complete outcome data, which is great. It's a low risk of bias.

How would you GRADE Leung et al.?

For each category below, rate the study as very low, low, moderate, high, or unknown risk

Random sequence generation (selection bias)	Low risk
Allocation concealment (selection bias)	Unknown risk
Blinding (performance bias and detection bias)	High risk
Incomplete outcome data (attrition bias)	Low risk
Selective reporting	Low risk



Outcome Measures

Primary Outcome Measures:

1. proportion of patients receiving influenza vaccination [Time Frame: three working days]

Secondary Outcome Measures:

1. proportion of patients receiving influenza vaccination [Time Frame: five working days]
2. proportion of patients receiving influenza vaccination [Time Frame: seven working days]
3. proportion of patients receiving influenza vaccination [Time Frame: nine working days]

https://clinicaltrials.gov/ct2/show/NCT02741843

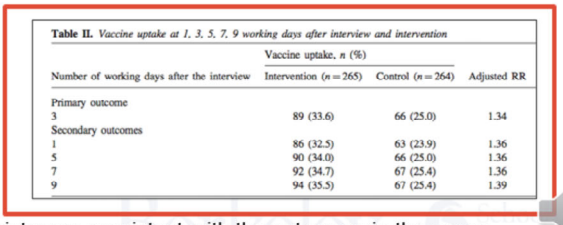


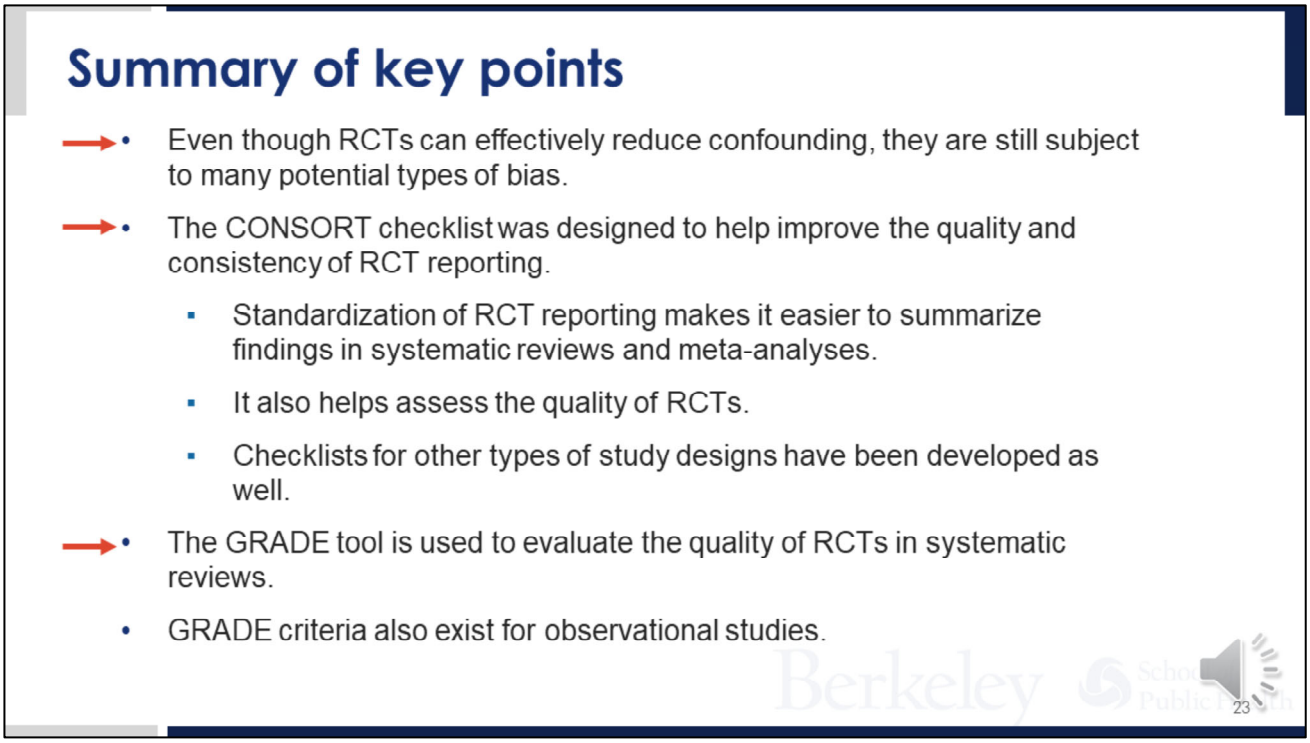
Table II. *Vaccine uptake at 1, 3, 5, 7, 9 working days after interview and intervention*

Number of working days after the interview	Vaccine uptake, *n* (%) Intervention (*n* = 265)	Control (*n* = 264)	Adjusted RR
Primary outcome			
3	89 (33.6)	66 (25.0)	1.34
Secondary outcomes			
1	86 (32.5)	63 (23.9)	1.36
5	90 (34.0)	66 (25.0)	1.36
7	92 (34.7)	67 (25.4)	1.36
9	94 (35.5)	67 (25.4)	1.39

The outcomes listed in the ClinicalTrials.gov registry are consistent with the outcomes in the paper.

22

And then finally, with respect to selective outcome reporting, a great way to assess this is for trials to look to see if they've registered the trial somewhere. There is a clinicaltrials.gov website. And this is the most common place for studies to register themselves. It's for trials only. There's also other types of sites available online for other study designs.This is a place where you can list the study design, the interventions, the outcomes, the investigators, and the timeline publicly prior to completion of the trial. And that website will stay alive even when the study is completed. Someone like me can come along and say, OK, what were the primary outcomes and secondary outcomes at the time they designed the study?Here we can see that the primary outcome was the proportion of patients receiving the influenza vaccine and the time frame was three working days. And then the secondary outcome was the same thing with different time frames--5, 7, and 9 working days. And then *here is a snapshot from the paper table two. So we can see it says, vaccine uptake at 1, 3, 5, 7, and 9 working days. They actually included yet another time frame that wasn't prespecified.Overreporting is not a bad thing. What we'd be concerned about is if they for example, didn't list the three working day result and only listed the 5, 7 and 9 working day results. That would suggest that perhaps there was an undesirable finding for that outcome that they didn't want to include in the publication. But in this case, this information is consistent with the registration. And so this is considered low risk or very low risk. Now, if I had taken the time to go through and compare everything in the clinicaltrials.gov registry against the paper, I may decide to make it a very low risk classification but I've just focused on the primary outcomes here for the purpose of this video.



To summarize, even though randomized trials can effectively reduce both measured and unmeasured confounding, they're still subject to many potential types of biases as we covered in a previous video. The CONSORT checklist was designed to help improve the quality and consistency of trial reporting. The standardization of trial reporting makes it easier to summarize findings in systematic reviews and meta-analyzes and it also helps us assess the quality of trials.There's checklists for other types of study designs as well. The GRADE tool is used to evaluate the quality of randomized trials and systematic reviews. We've done it for just a single trial in this video as an example. And GRADE criteria also exists for observational studies.

CONSORT 2010 checklist of information to include when reporting a randomised trial*

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	
	2b	Specific objectives or hypotheses	
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	
Participants	4a	Eligibility criteria for participants	
	4b	Settings and locations where the data were collected	
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	
	6b	Any changes to trial outcomes after the trial commenced, with reasons	
Sample size	7a	How sample size was determined	
	7b	When applicable, explanation of any interim analyses and stopping guidelines	
Randomisation:			
Sequence generation	8a	Method used to generate the random allocation sequence	
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those	

		assessing outcomes) and how	
	11b	If relevant, description of the similarity of interventions	
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	
Results			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	
	13b	For each group, losses and exclusions after randomisation, together with reasons	
Recruitment	14a	Dates defining the periods of recruitment and follow-up	
	14b	Why the trial ended or was stopped	
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	
Discussion			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	
Other information			
Registration	23	Registration number and name of trial registry	
Protocol	24	Where the full trial protocol can be accessed, if available	
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	

*We strongly recommend reading this statement in conjunction with the CONSORT 2010 Explanation and Elaboration for important clarifications on all the items. If relevant, we also recommend reading CONSORT extensions for cluster randomised trials, non-inferiority and equivalence trials, non-pharmacological treatments, herbal interventions, and pragmatic trials. Additional extensions are forthcoming: for those and for up to date references relevant to this checklist, see www.consort-statement.org.

Consensus statement

The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials

David Moher[1], Kenneth F Schulz[2] and Douglas G Altman[3]

Address: [1]University of Ottawa, Thomas C. Chalmers Centre for Systematic Reviews, Ottawa, Canada, [2]Family Health International and Dept. of Obstetrics and Gynecology, School of Medicine, University of North Carolina at Chapel Hill, North Carolina, USA and [3]ICRF Medical Statistics Group and Centre for Statistics in Medicine, Institute of Health Sciences, Oxford, UK for the CONSORT Group

E-mail: David Moher - dmoher@uottawa.ca; Kenneth F Schulz - kschulz@fhi.org; Douglas G Altman - d.altman@icrf.icnet.uk

Published: 20 April 2001

BMC Medical Research Methodology 2001, **1**:2

Received: 30 March 2001
Accepted: 20 April 2001

This article is available from: http://www.biomedcentral.com/1471-2288/1/2

Abstract

To comprehend the results of a randomized controlled trial (RCT), readers must understand its design, conduct, analysis and interpretation. That goal can only be achieved through complete transparency from authors. Despite several decades of educational efforts, the reporting of RCTs needs improvement. Investigators and editors developed the original CONSORT (Consolidated Standards of Reporting Trials) statement to help authors improve reporting by using a checklist and flow diagram. The revised CONSORT statement presented in this paper incorporates new evidence and addresses some criticisms of the original statement.

The checklist items pertain to the content of the Title, Abstract, Introduction, Methods, Results and Discussion. The revised checklist includes 22-items selected because empirical evidence indicates that not reporting the information is associated with biasedestimates of treatment effect or the information is essential to judge the reliability or relevance of the findings. We intended the flow diagram to depict the passage of participants through an RCT. The revised flow diagram depicts information from four stages of a trial (enrolment, intervention allocation, follow-up, and analysis). The diagram explicitly includes the number of participants, for each intervention group, included in the primary data analysis. Inclusion of these numbers allows the reader to judge whether the authors have performed an intention-to-treat analysis.

In sum, the CONSORT statement is intended to improve the reporting of an RCT, enabling readers to understand a trial's conduct and to assess the validity of its results.

Contributors

Frank Davidoff, MD, *Annals of Internal Medicine,* (Philadelphia, PA); Susan Eastwood, ELS(D), University of California at San Francisco, (San Francisco, CA); Matthias Egger, MD, Department of Social Medicine, University of Bristol, (Bristol, UK); Diana Elbourne, PhD, London School of Hygiene and Tropical Medicine, (London, UK); Peter Gøtzsche, MD, Nordic Cochrane Centre, (Copenhagen, Denmark); Sylvan B. Green, PhD, MD, School of Medicine, Case Western Reserve University, (Cleveland, OH); Leni Grossman, BA, Merck & Co., Inc., (Whitehouse Station, NJ); Barbara S. Hawkins, MD, Wilmer Ophthalmological Institute, Johns Hopkins University, (Baltimore, MD); Richard Horton, MB, *The Lancet,* (London, UK); Wayne B. Jonas, MD, Uniformed Services University of the Health Sciences, (Bethesda, MD); Terry Klassen, MD, Department of Pediatrics, University of Alberta, (Edmonton, Alberta); Leah Lepage,

PhD, Thomas C. Chalmers Centre for Systematic Reviews, (Ottawa, ON); Thomas Lang, MA, Tom Lang Communications, (Lakewood, OH); Jeroen Lijmer, MD, Dept. of Clinical Epidemiology, University of Amsterdam, (Amsterdam, The Netherlands); Rick Malone, BS, TAP Pharmaceuticals, (Lake Forest, IIL); Curtis L. Meinert, PhD, Johns Hopkins University, (Baltimore, MD); Mary Mosley, BS, Life Science Publishing, (Tokyo, Japan); Stuart Pocock, PhD, London School of Hygiene and Tropical Medicine, (London, UK); Drummond Rennie, *JAMA,* Chicago, IIL); David S. Riley, MD, University of New Mexico Medical School, (Santa Fe, NM); Roberta W. Scherer, MD, Epidemiology & Preventive Medicine, University of Maryland School of Medicine, (Baltimore, MD); Ida Sim, MD, PhD, University of California at San Francisco, (San Francisco, CA); Donna Stroup, PhD, MSc, Epidemiology Program Office, Center for Disease Control & Prevention, (Atlanta, GA).

David Moher, Ken Schulz, and Doug Altman participated in regular conference calls, identified participants, contributed in the CONSORT meetings and drafted the manuscript. David Moher and Leah Lepage planned the CONSORT meetings, identified and secured funding, invited the participants and planned the meeting agenda. The members of the CONSORT group listed above attended the consort meetings and provided input in the revised checklist, flow diagram and/or text of this manuscript. David Moher is the Guarantor of the manuscript.

Introduction

A report of a randomized controlled trial (RCT) should convey to the reader, in a transparent manner, why the study was undertaken, and how it was conducted and analyzed. For example, a lack of adequately reported randomization has been associated with bias in estimating the effectiveness of interventions, [1,2]. To assess the strengths and limitations of an RCT, readers need and deserve to know the quality of its methodology.

Despite several decades of educational efforts, RCTs still are not being reported adequately, [3–6]. For example, a review, [5] of 122 recently published RCTs that evaluated the effectiveness of selective serotonin reuptake inhibitors (SSRI) as first-line management strategy for depression found that only one (0.8%) paper described randomization adequately. Inadequate reporting makes the interpretation of RCTs difficult if not impossible. Moreover, inadequate reporting borders on unethical practice when biased results receive false credibility.

History of CONSORT

In the mid 1990s, two independent initiatives to improve the quality of reports of RCTs led to the publication of the CONSORT statement, [7] which was developed by an international group of clinical trialists, statisticians, epidemiologists and biomedical editors. CONSORT has been supported by a growing number of medical and health care journals, [8–11] and editorial groups, including the International Committee of Medical Journal Editors, [12] (ICMJE, The Vancouver Group), the Council of Science Editors (CSE), and the World Association of Medical Editors (WAME). CONSORT is also published in Dutch, English, French, German, Japanese, and Spanish. It can be accessed together with other information about the CONSORT group on the Internet, [13].

The CONSORT statement comprises a checklist and flow diagram for reporting an RCT. For convenience, the checklist and diagram together are called simply CONSORT. They are primarily intended for use in writing, reviewing, or evaluating reports of simple two- group parallel RCTs.

Preliminary indications are that the use of CONSORT does indeed help to improve the quality of reports of RCTs, [14,15]. In an evaluation, [14] of 71 published RCTs, in three journals in 1994, allocation concealment was not clearly reported in 61% (n = 43) of the RCTs. Four years later, after these three journals required authors reporting an RCT to use CONSORT, the proportion of papers in which allocation concealment was not clearly reported had dropped to 39% (30 of 77, mean difference = -22%; 95% confidence interval of the difference: -38%, -6%).

The usefulness of CONSORT is enhanced by continuous monitoring of the biomedical literature that permits it to be modified depending on the merits of maintaining, or dropping current items and including new items. For example, when Meinert, [16] observed that the flow diagram did not provide important information about the number of participants who entered each phase of an RCT (i.e., enrollment, treatment allocation, follow-up, and data analysis), the diagram could be modified to accommodate the information. The checklist is similarly flexible.

This iterative process makes the CONSORT statement a continually evolving instrument. While participants in the CONSORT group and their degree of involvement vary over time, members meet regularly to review the need to refine CONSORT. At the 1999 meeting a decision was made to revise the original statement. This report reflects changes determined by consensus of the CONSORT group, partly in response to emerging evidence on the importance of various elements of RCTs.

Table 1: Checklist of items to include when reporting a randomized trial

PAPER SECTION And topic	Item #	Descriptor	Reported on page #
TITLE & ABSTRACT	1	How participants were allocated to interventions (e.g., random allocation", "randomized", or "randomly assigned").	
INTRODUCTION			
Background	2	Scientific background and explanation of rationale.	
METHODS			
Participants	3	Eligibility criteria for participants and the settings and locations where the data were collected.	
Interventions	4	Precise details of the interventions intended for each group and how and when they were actually administered.	
Objectives	5	Specific objectives and hypotheses.	
Outcomes	6	Clearly defined primary and secondary outcome measures and, when applicable, any methods used to enhance the quality of measurements (e.g., multiple observations, training of assessors).	
Sample size	7	How sample size was determined and, when applicable, explanation of any interim analyses and stopping rules.	
Randomization:			
Sequence generation	8	Method used to generate the random allocation sequence, including details of any restriction (e.g., blocking, stratification).	
Allocation concealment	9	Method used to implement the random allocation sequence (e.g., numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned.	
Implementation	10	Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups.	
Blinding (Masking)	11	Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. If done, how the success of blinding was evaluated.	
Statistical methods	12	Statistical methods used to compare groups for primary outcome(s); Methods for additional analyses, such as subgroup analyses and adjusted analyses.	
RESULTS			
Participant flow	13	Flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group report the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analyzed for the primary outcome. Describe protocol deviations from study as planned, together with reasons.	
Recruitment	14	Dates defining the periods of recruitment and follow-up.	
Baseline data	15	Baseline demographic and clinical characteristics of each group.	
Numbers analyzed	16	Number of participants (denominator) in each group included in each analysis and whether the analysis was by intention-to-treat". State the results in absolute numbers when feasible (e.g., 10/20, not 50%).	
Outcomes and Estimation	17	For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision (e.g., 95% confidence interval).	
Ancillary analyses	18	Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those pre-specified and those exploratory.	
Adverse events	19	All important adverse events or side effects in each intervention group.	
DISCUSSION			
Interpretation	20	Interpretation of the results, taking into account study hypotheses, sources of potential bias or imprecision and the dangers associated with multiplicity of analyses and outcomes.	
Generalizability	21	Generalizability (external validity) of the trial findings.	
Overall evidence	22	General interpretation of the results in the context of current evidence.	

Revision of the CONSORT statement

Thirteen members of the CONSORT group met in May 1999 with the primary objective of revising the original CONSORT checklist and flow diagram, as needed. The merits of including each item were discussed by the group in the light of current evidence. As in developing the original CONSORT statement, our intention was to keep only those items deemed fundamental to reporting standards for an RCT. Some items not considered essential may well be highly desirable and should still be included in an RCT report even though they are not included in CONSORT. Such items include institutional ethical review board approval, sources of funding for the trial, and a trial registry number (as, for example, the International Standard Randomized Controlled Trial Number (ISRCTN) used to register the RCT at its inception [17].

Shortly after the meeting a revised version of the checklist was circulated to the group for additional comments and feedback. Revisions to the flow diagram similarly were made. All of these changes were discussed when CONSORT participants met in May 2000 and the revised statement finalized shortly afterwards.

The revised CONSORT statement includes a 22-item checklist (Table 1) and a flow diagram (Figure 1). Its primary aim is helping authors improve the quality of reports of simple two-group parallel RCTs. However, the basic philosophy underlying the development of the statement can be applied to any design. In this regard additional statements for other designs will be forthcoming from the group, [13]. CONSORT can also be used by peer reviewers and editors to identify reports with inadequate description of trials and those with potentially biased results, [1,2].

During the 1999 meeting the group also discussed the benefits of developing an explanatory document to enhance the use and dissemination of CONSORT. The document is patterned on reporting of statistical aspects of clinical research, [18] which was developed to help facilitate the recommendations of the ICMJEs Uniform Requirements for Manuscripts Submitted to Biomedical Journals. Three members of the CONSORT group (DGA, KFS, DM), with assistance from members on some checklist items, drafted an explanation and elaboration document. That document, [19] was circulated to the group for additions and revisions and was last revised after review at the latest CONSORT group meeting.

Changes to CONSORT

• In the revised checklist a new column for "paper section and topic" integrates information from the "subheading" column that was contained in the original statement.

• The "Was it reported?" column has been integrated into a "reported on page #" column, as requested by some journals.

• Each item of the checklist is now numbered and the syntax and order have been revised to improve the flow of information.

• The "Title" and "Abstract" are now combined in the first item.

• While the content of the revised checklist is similar to the original one some items that previously were combined are now separate. For example, previously authors were asked to describe "primary and secondary outcome(s) measure(s) and the minimum important difference(s), and indicate how the target sample size was projected". In the new version issues pertaining to outcomes (item 6) and sample size (item 7) are separate, enabling authors to be more explicit about each. Moreover, some items request additional information. For example, for outcomes (item 6) authors are asked to report any methods used to enhance the quality of measurements, such as multiple observations.

• The item asking for the unit of randomization (e.g., cluster) has been dropped because specific checklists have been developed for reporting cluster RCTs, [20] and other design types, [13] since publication of the original checklist.

• Whenever possible new evidence is incorporated into the revised checklist. For example, authors are asked to be explicit about whether the analysis reported is by intention-to-treat (item 16). This request is based in part on the observations, [21] that (a) authors do not adequately describe and apply intention-to-treat analysis and (b) reports that do not provide this information are less likely to report other relevant information, such as losses to follow-up, [22].

• The revised flow diagram depicts information from four stages of a trial (enrolment, intervention allocation, follow-up, and analysis). The revised diagram explicitly includes the number of participants, for each intervention group, included in the primary data analysis. Inclusion of these numbers lets the reader know whether the authors have performed an intention to treat analysis, [21–23]. Because some of the information may not always be known and to accommodate other information, the structure of the flow diagram may need to be modified for a particular trial. Inclusion of the participant flow diagram in the report is strongly recommended but may be unnecessary for simple trials, such as those without any participant withdrawals or dropouts.

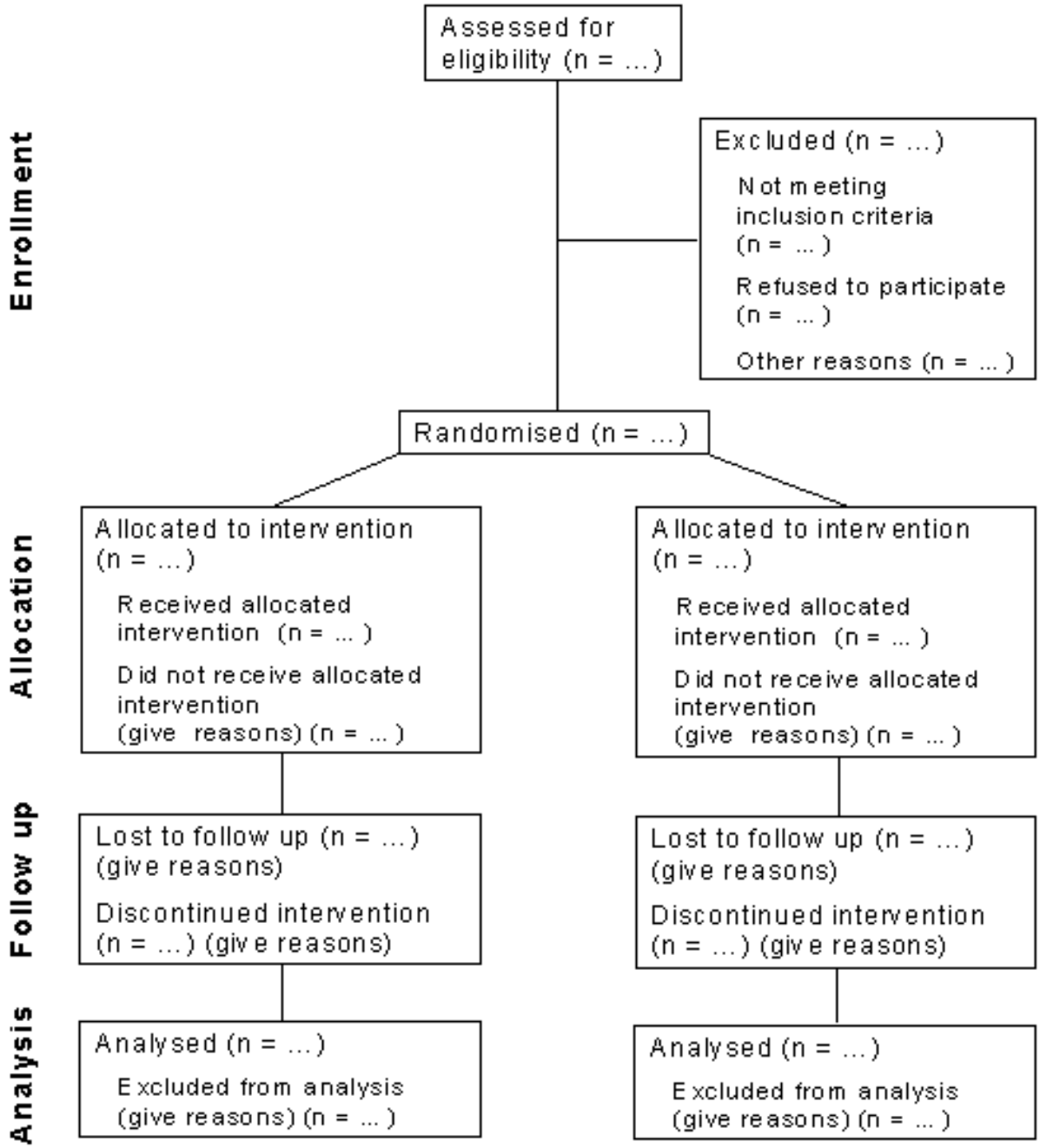


Figure 1
Flow Diagram of the progress through the phases of a randomized trial (i.e., enrollment, intervention allocation, follow-up, and data analysis)

Discussion

Specifically developed to provide guidance to authors about how to improve the quality of reporting of simple two-group parallel RCTs, CONSORT encourages transparency when reporting the methods and results so that reports of RCTs can be interpreted both readily and accurately. However, CONSORT does not address other facets of reporting that also require attention, such as scientific content and readability of RCT reports. Some authors in their enthusiasm to use CONSORT have modified the checklist, [24]. We recommend against such modifications because they may be based on a different process than the one used by the CONSORT group.

The use of CONSORT seems to reduce (if not eliminate) inadequate reporting of RCTs [14,15]. Potentially, the use of CONSORT should have a positive influence on how RCTs are conducted. Granting agencies have noted this potential relationship and in at least in one case, [25] have encouraged grantees to consider in their application how they have dealt with the CONSORT items.

The evidence-based approach used to develop CONSORT has also been used to develop standards for reporting meta-analyses of randomized trials, [26], meta-analyses of observational studies, [27] and diagnostic studies (personal communication - Jeroen Lijmer). Health economists have also started to develop reporting standards, [28] to help improve the quality of their reports, [29]. The intent of all of these initiatives is to improve the quality of reporting of biomedical research, [30] and by doing so to bring about more effective health care.

The revised CONSORT statement will replace the original one in those journals and groups that already support it. Journals that do not yet support CONSORT may do so by registering on the CONSORT Internet site, [13]. In order to convey to authors the importance of improved quality in the reporting of RCTs, we encourage supporting journals to reference the revised CONSORT statement and the CONSORT Internet address, [13] in their "Instructions to Contributors". As the journals publishing the revised CONSORT statement have waived copyright protection, CONSORT is now widely accessible to the biomedical community. The CONSORT checklist and flow diagram can also be accessed at the CONSORT Internet site, [13].

A lack of clarification of the meaning and rationale for each checklist item in the original CONSORT statement has been remedied with the development of the CONSORT explanation and elaboration document, [19] which can also be found on the CONSORT Internet site, [13]. This document includes reporting the evidence on which the checklist items are based, including the references, which annotated the checklist items in the previous version. We encourage journals to also include reference to this document also in their Instructions to Contributors.

Emphasizing the evolving nature of CONSORT, the CONSORT group invites readers to comment on the updated checklist and flow diagram through the CONSORT Internet site, [13]. Comments and suggestions will be collated and considered at the next meeting of the group in 2001.

Footnote

The revised CONSORT statement is also published in JAMA (Moher D, Schulz KF, Altman DG, for the CONSORT group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. JAMA 2001; **285**:1987-1991), The Lancet (Moher D, Schulz KF, Altman DG, for the CONSORT group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials Lancet; 2001; **357**:1191-1194), and Annals of Internal Medicine (Moher D, Schulz KF, Altman DG, for the CONSORT group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. Annals of Internal Medicine; 2001; **134**:657-662). Authors can use any one of these references, as well as the reference in BMC Medical Research Methodology, when citing CONSORT.

Acknowledgements

The effort to improve the reporting of randomized trials, from its beginnings with the Standards of Reporting Trials (SORT) group to the current activities of the Consolidated Standards of Reporting Trials (CONSORT) group, have involved a large number of people around the globe. We wish to thank Leah Lepage for keeping everybody all lined up and moving in the same direction.

Financial support to convene meetings of the CONSORT group was provided in part by Abbott Laboratories, American College of Physicians, GlaxoWellcome, The Lancet, Merck, the Canadian Institutes for Health Research, National Library of Medicine, and TAP Pharmaceuticals.

References

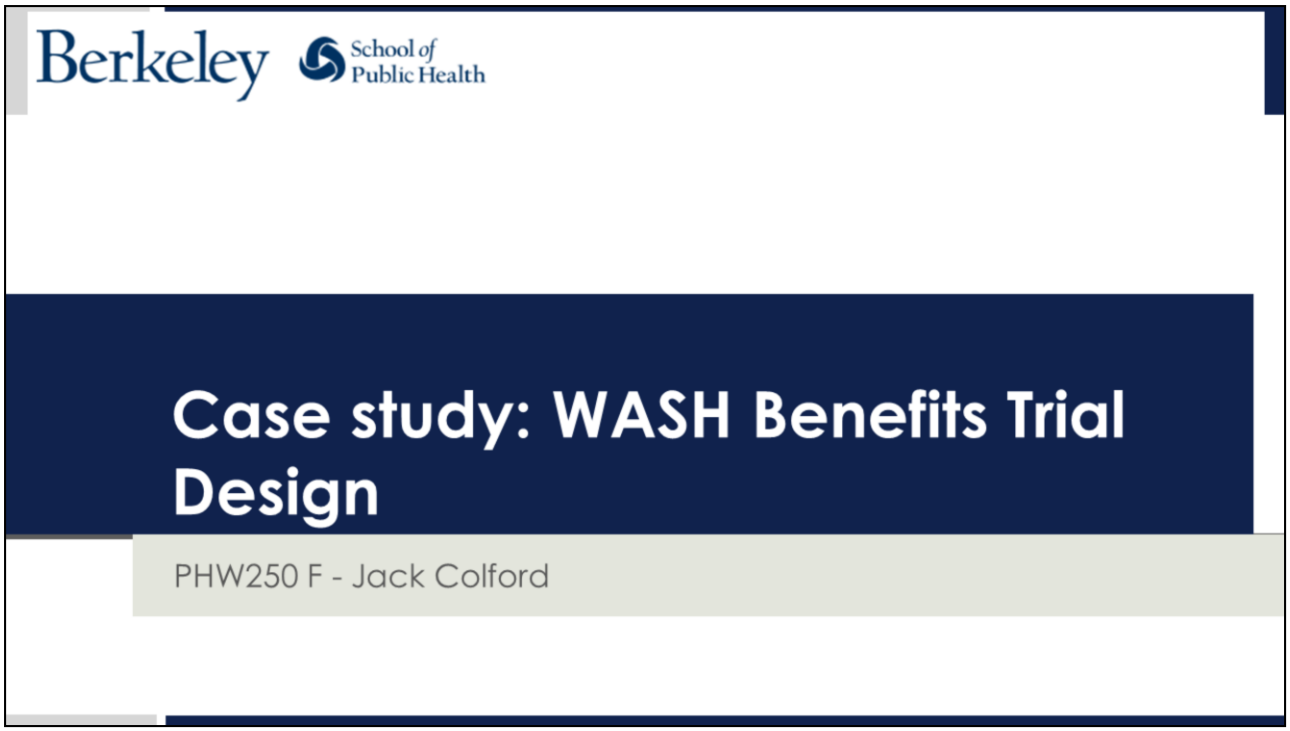
1. Schulz KF, Chalmers I, Hayes RJ, Altman DG: **Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials.** *JAMA* 1995, **273**:408-412
2. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P: **Does the quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses?** *Lancet* 1998, **352**:609-613
3. Jadad AR, Boyle M, Cunningham C, Kim M, Schachar R: **Treatment of Attention-Deficit/Hyperactivity Disorder.** *Evidence Report/ Technology Assessment No.11 (Prepared by McMaster University under Contract No. 290-97-0017),* 2000,
4. Thornley B, Adams CE: **Content and quality of 2000 controlled trials in schizophrenia over 50 years** *BMJ* 1998, **317**:1181-1184
5. Hotopf M, Lewis G, Normand C: **Putting trials on trial- the costs and consequences of small trials in depression: a systematic**

review of methodology. *J Epidemiol Community Health* 1997, **51**:354-358
6. Dickinson K, Bunn F, Wentz R, Edwards P, Roberts I: **Size and quality of randomised controlled trials in head injury: review of published studies.** *BMJ* 2000, **320**:1308-1311
7. Begg CB, Cho MK, Eastwood S, Horton R, Moher D, Olkin I, Rennie D, Schulz KF, Simel DL, Stroup DF: **Improving the quality of reporting of randomized controlled trials: the CONSORT statement.** *JAMA* 1996, **276**:637-639
8. Freemantle N, Mason JM, Haines A, Eccles MP: **CONSORT: an important step toward evidence-based health care.** *Ann Intern Med* 1997, **126**:81-83
9. Altman DG: **Better reporting of randomized controlled trials: the CONSORT statement.** *BMJ* 1996, **313**:570-571
10. Schulz KF: **The quest for unbiased research: randomized clinical trials and the CONSORT reporting guidelines.** *Ann Neurol* 1997, **41**:569-573
11. Huston P, Hoey J: **CMAJ endorses the CONSORT statement.** *CMAJ* 1996, **155**:1277-1279
12. Davidoff F: **News from the International Committee of Medical Journal Editors.** *Ann Intern Med* 2000, **133**:229-231
13. 2000, [http://www.consort-statement.org]
14. Moher D, Jones A, Lepage L, : **Use of the CONSORT statement and quality of reports of randomized trials: a comparative before and after evaluation?** *JAMA* 2001, **285**:1992-1995
15. Egger M, Jüni P, Bartlett C, : **The value of patient flow charts in reports of randomized controlled trials: bibliographic study.** *JAMA* 2001, **285**:1996-1999
16. Meinert CL: **Beyond CONSORT: Need for improved reporting standards for clinical trials.** *JAMA* 1998, **279**:1487-1489
17. Chalmers I: **Current Controlled Trials: an opportunity to help improve the quality of clinical research.** *Curr Control Trials Cardiovasc Med* 2000, **1**:3-8
18. Bailer JC III, Mosteller F: **Guidelines for statistical reporting in articles for medical journals: amplifications and explanations.** *Ann Intern Med* 1988, **108**:266-273
19. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gøtzsche PC, Lang T, : **The revised CONSORT statement for reporting randomized trials: explanation and elaboration.** *Annals of Internal Medicine.* 2001, **134**:663-694
20. Elbourne DR, Campbell MK: **Extending the CONSORT statement to cluster randomised trials: for discussion.** *Stats Med* 2001, **20**:489-496
21. Hollis S, Campbell F: **What is meant by intention-to-treat analysis? Survey of published randomized controlled trials.** *BMJ* 1999, **319**:670-674
22. Ruiz-Canela M, Martinez-Gonzalez MA, de Irala-Estevez J: **Intention-to-treat analysis is related to methodological quality.** *BMJ* 2000, **320**:1007-
23. Lee YJ, Ellenberg JH, Hirtz DG, Nelson KB: **Analysis of clinical trials by treatment actually is it really an option?** *Stat Med.* 1991, **10**:1595-1605
24. Bentzen SM: **Towards evidence based radiation oncology: improving the design, analysis, and reporting of clinical outcome studies in radiotherapy.** *Radiother Oncol.* 1998, **46**:5-18
25. O'Toole LB: **MRC uses checklist similar to CONSORTs.** *BMJ* 1997, **314**:1127-
26. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF, : **Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement.** *Lancet* 1999, **354**:1896-1900
27. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB: **Meta-analysis of observational studies in epidemiology: a proposal for reporting.** *JAMA.* 2000, **283**:2008-2012
28. Siegel JE, Weinstein MC, Russell LB, Gold MR: **Recommendations for reporting cost-effectiveness analysis.** *JAMA* 1996, **276**:1339-1341
29. Neumann PJ, Stone PW, Chapman RH, Sandberg EA, Bell CM: **The Quality of Reporting in Published Cost-Utility Analyses, 1976-1997.** *Ann Intern Med* 2000, **132**:964-972
30. Altman DG: **The scandal of poor medical research.** *BMJ* 1994, **308**:283-284

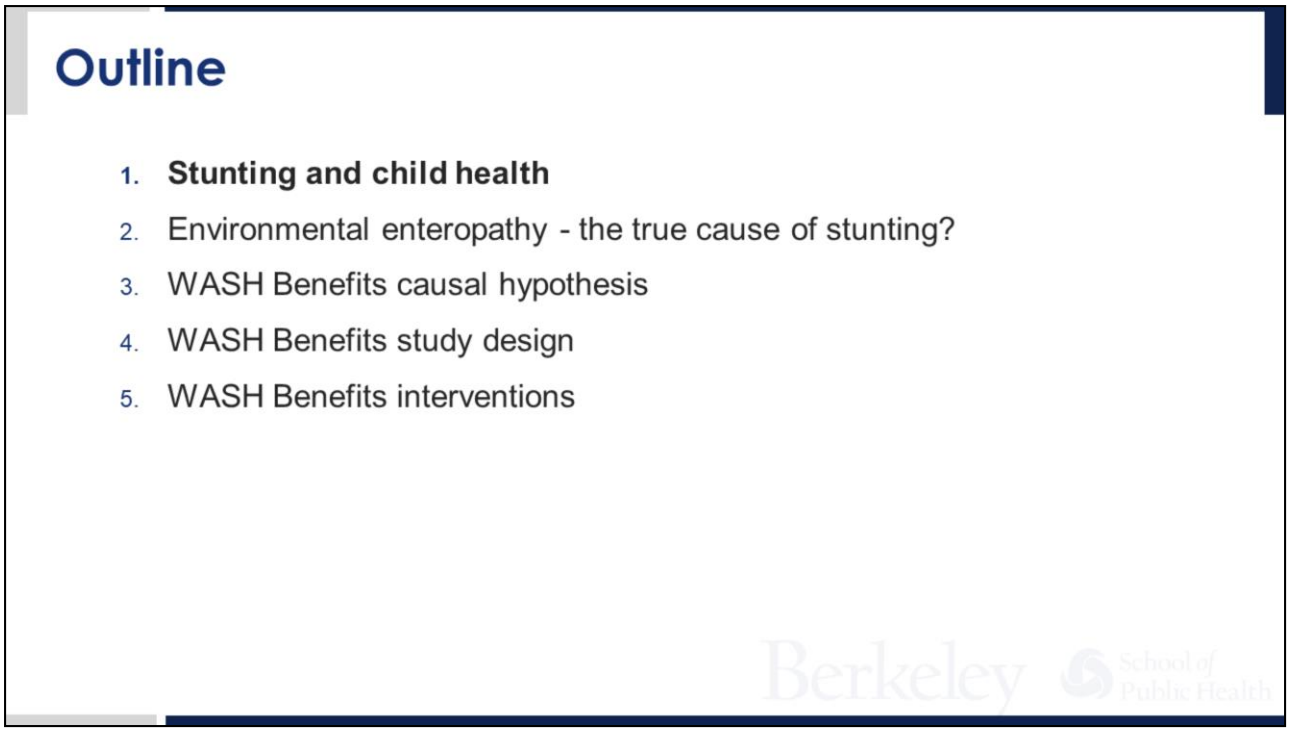
Correspondence

Correspondence should be addressed to: Leah Lepage, Thomas C Chalmers Center for Systematic Reviews, Children's Hospital of Eastern Ontario research Institute, Room R235; 401 Smyth Road, Ottawa, Ontario, KIH 8LI, Canada. Email: llepage@ottawa.ca

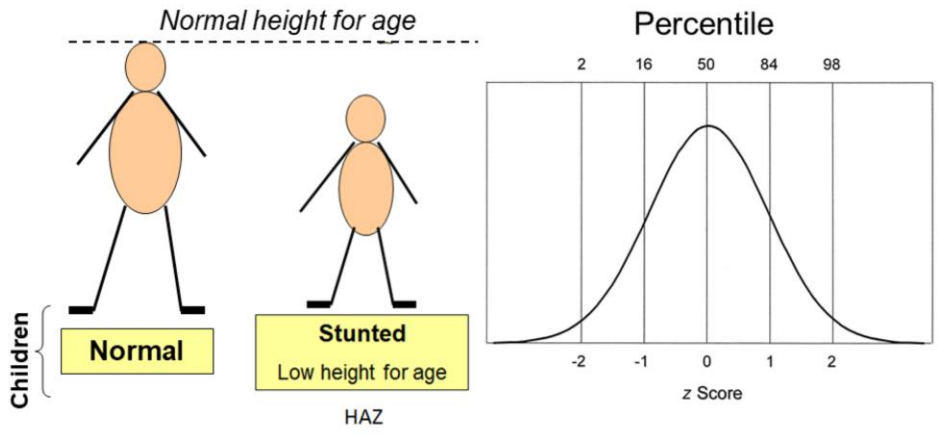
Lecture: WASH Benefits Design



JACK COLFORD: We are now going to discuss an applied example of a randomized-controlled trial. This is a study done by my group, and Jade and I were very involved in this for the past eight years or so, so we know it well and we'd like to share it with you. In this first part of this two lecture sequence, I'm going to discuss some of the background and science behind this study so the trial results will make sense. The study is called WASH benefits, and WASH stands for water, sanitation, and hygiene. And this was a study designed to look at the various benefits that might occur from interventions on the WASH aspects of life.



The way I've organized this is I'll first talk about stunting in child health. Stunting is basically poor linear growth. Then I'll talking about a concept called environmental enteropathy, which some speculate might be the true cause of stunting. We'll talk about the causal hypothesis and WASH benefits, what the study design was for WASH benefits, and then what the WASH benefits interventions were. And then in the next lecture, I'll talk about some of the results.

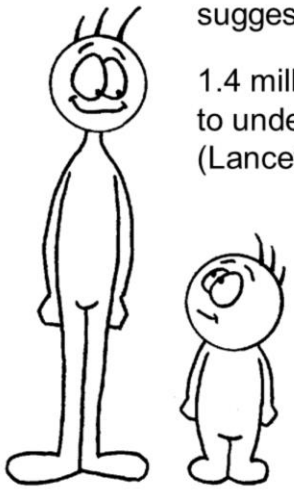


So when you look at the growth of children, a normal child attains a normal height for age, according to different standards that are set but usually WHO or World Health Organization type guidelines. But a stunted child doesn't attain the proper linear growth, so it has a low height for age, and this is measured by something called the HAZ score, or the height for age z-score. And the height for a z-score comes from a normal distribution that is drawn, essentially, for all the different heights that children might obtain. So on average, children would have a z-score of 0, a child who is very, very tall would have a positive z-score, up above 1 or 2, and a child who was very, very short for their age would have a score in the negative range, down below minus 1 or minus 2. This is a standardized score for all the heights of children in a given country.

Why worry about stunting?

>2.5% prevalence of short stature in a community, suggests chronic under-nutrition

1.4 million child deaths annually attributable to undernutrition.
(Lancet 2012; 380: 2224–60)



http://printablecolouringpages.co.uk

Guatemala trial follow-up (*Am J Clin Nutr* 2013;98:1170–8.)

1 SD increase in height at 2 years:

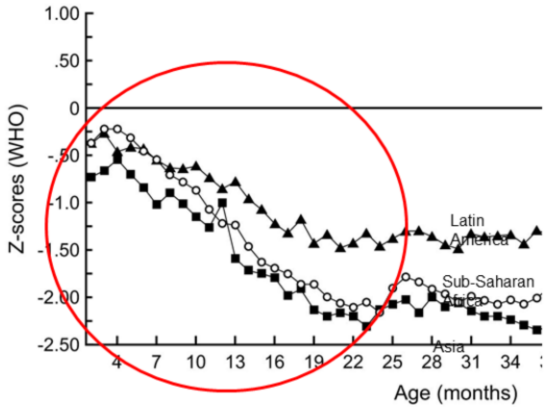
- 0.78 more years in school
- 21% higher adult income

Malnourished children face:

- cognitive impairment
- less success in school
- decreased wages

Why is it worth worrying about stunting? Well, greater than 2.5% prevalence of short stature stunting in a community suggests chronic under-nutrition, and there are 1.4 million child deaths annually attributed to under-nutrition. In a long-term trial in Guatemala, a 1 standard deviation increase in height two years was associated with a 0.78 additional years of schooling, a 21% higher adult income. Whereas malnourished children face cognitive impairments and less success in school, decreased wages, so number of penalties in the sense that come from lack of adequate growth.

Critical period for growth faltering



The first 1000 days

- Maternal nutrition
- Early child nutrition
- Key area for
  - Research
  - Interventions

Slide from Christine

Adapted from Victora CG, Pediatrics March; 125(3):e473-480

If you look at the ages of different children in populations, what we can see in children who falter, that is don't attain their normal height, this faltering occurs in the first 24 months for approximately first two years of life, in multiple different studies on different continents, what we observe is that when this growth faltering occurs, it's almost impossible to recover from it. So therefore, many different interventions are targeted early growth interventions in the first two years of life to try to prevent this faltering from occurring. And this first 1,000 days of growth is considered to be particularly important for maternal nutrition, early child nutrition, and it's a key area for research and interventions to try to target children as early as possible.

Photo: Mubina Agboatwalla

If children are malnourished

- Feed them more
  - But more calories are insufficient
  - need nutrient dense food
- Supplement with nutrient dense foods
  - only correct 15-30% of growth faltering
    (Dewey K. *Matern Child Nutr* 2008, 4 Suppl 1: 24--85)

- 118 Kcal
- 9.6 gm fat
- 2.6 gm protein
- ≥100% RDA of 12 vitamins
- 9 minerals

So if children are nourished, we can feed them more, but more calories are insufficient. They need a nutrient dense type food. Much work has been done on trying to supplement their diet with nutrient-dense foods, but doing so only corrects about 15% to 30% of the growth faltering that we see. Now there are different formulations to try to do it. The one we're going to use in this study called WASH benefits that I'm talking about is a formulation called nutra butter, and you see the description here on the slide of its various sub-components. And this is added to food, it's not a food replacement. It's a supplement to the food that young children are eating.

Outline

1. Stunting and child health
2. **Environmental enteropathy - the true cause of stunting?**
3. WASH Benefits causal hypothesis
4. WASH Benefits study design
5. WASH Benefits interventions

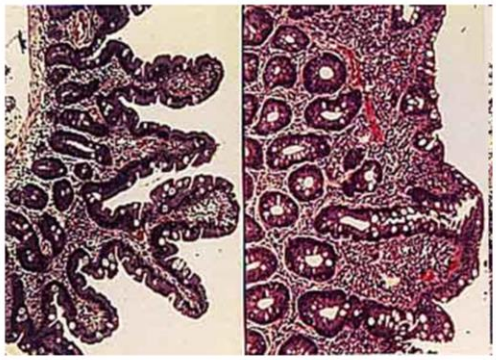
Berkeley School of Public Health

All right, now I'd like to talk about environmental enteropathy.

Pathology

Environmental Enteropathy
Environmental Enteric Dysfunction

- Change in intestinal villa architecture
  - Flattened; Reduced villous height
  - Increased
    - crypt depth
    - mitosis per crypt
- Inflammatory cell infiltration
  - Increased intraepithelial lymphocyte count
  - Mucosal T-cell activation
    - CD3+ CD69+
    - CD3+HLA-DR+

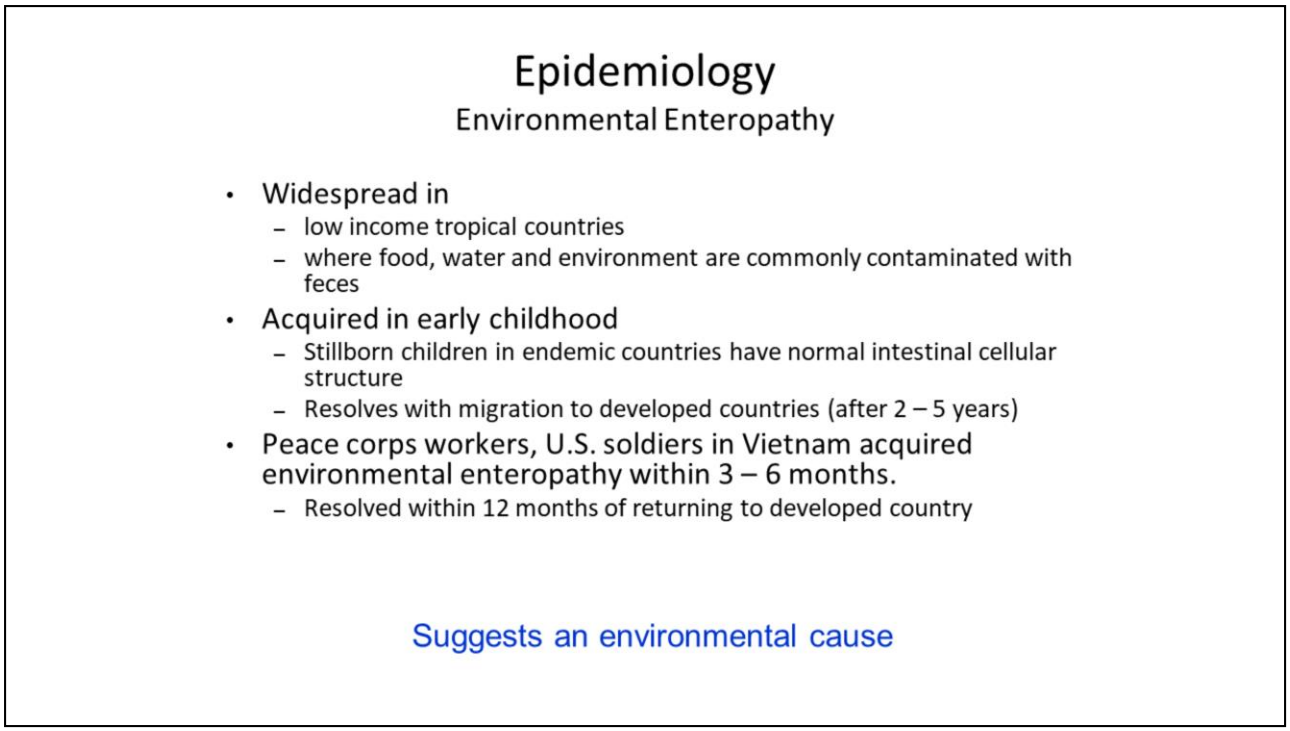


Normal Environmental Enteropathy

http://www.bio.davidson.edu/courses/Immunology/Students/spring2006/Mohr/Villi%20Atrophy.jpg

Veitch AM, *Euro J Gastro Hepatology* 2001, 13:1175-1181

In children in challenging situations in the developing world, we often see a change in their intestinal gut structure, the intestinal villi. They become flattened with a reduced height and an increased crypt death and lots more mitosis per crypt, so a number of microscopic changes that are visible in children who have this condition would need to be environmental enteropathy. And we also see inflammatory cell infiltration with increased intraepithelial lymphocyte counts and activation of particular immune cells. So lots of processes are going on in the guts of children who are believed to have environmental enteropathy.



The condition environmental enteropathy is believed to be widespread in low-income tropical countries where food, water, and environment are commonly contaminated with feces. It's acquired early in childhood, and stillborn children in endemic countries have normal intestinal cellular structure, so this seems to be something that happens after birth, environmental enteropathy is. And it resolves with migration to developed countries after the age of two to five years. So a child with these gut intestinal changes who then moves to developed country can see a reversal of these changes.

Also adults, like Peace Corps workers and US soldiers in Vietnam, who acquired changes consistent with environmental enteropathy after three to six months in the countries they were serving in, showed a reversal of this process 12 months after returning home. This suggest there's some environmental cause that's present in the developing country and not present back in the home country. We'll come back to that in a bit.

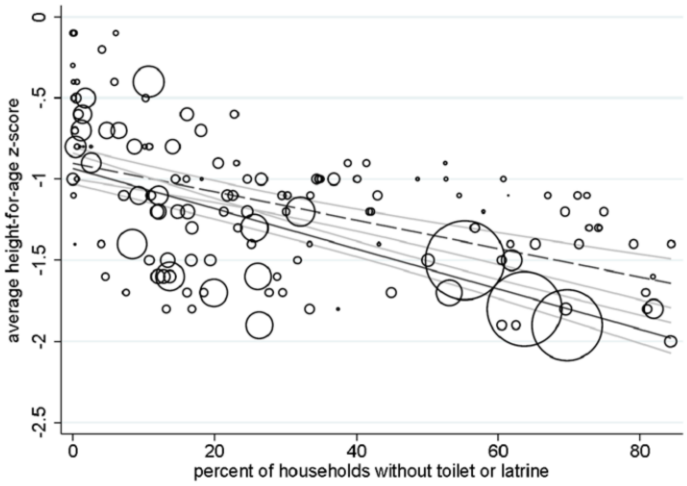
Outline

1. Stunting and child health
2. Environmental enteropathy - the true cause of stunting?
3. **WASH Benefits causal hypothesis**
4. WASH Benefits study design
5. WASH Benefits interventions

So let's talk about the causal hypotheses that the WASH benefits trial was attempting to study.

Child height versus open defecation

150 DHS assessments



- R^2 = 54%
- Minimal change in coefficient when adjusted for
  - GDP
  - Maternal
    - Height
    - Literacy
  - Water accessibility
  - Food availability
  - Breast feeding rates
  - Polity and autocracy scores

Spears D. How much international variation in child height in sanitation explain? Working paper www.riceinstitute.org

So if you plot the level of countries, the percent of households without toilets or latrines, you see that as the percentage on the x-axis gets larger of lack of toilets, lack of latrine services, the average height for HAZ score-- the HAZ that we talked about earlier-- pretty convincingly decreases, with a reasonably good correlation of about 54%. There is minimal change in the co-efficient to this association, even when you adjust for gross domestic product or maternal factors, such as height, literacy, or adjusting for water accessibility, or food availability, or breastfeeding rates, or policy and autocracy, or government scores. None of these things, when adjusted, seemed to lessen or much change this relationship between sanitation and height for age z-score.

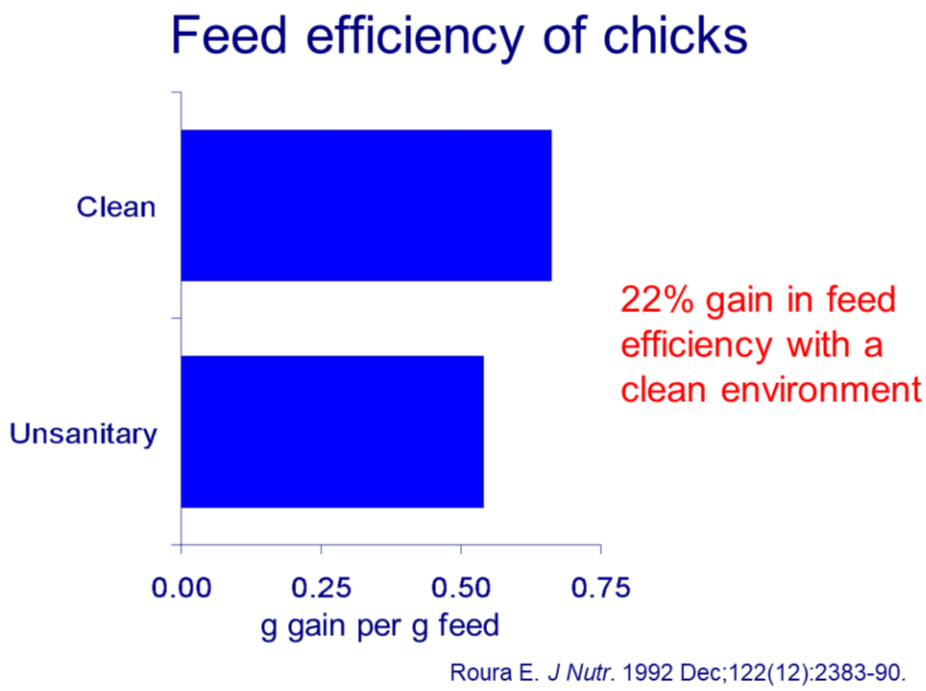
Do farm animals grow better in a clean environment?

- Randomized trial of chickens
- Outcome: Feed efficiency
  - g weight gain per g feed
- Unsanitary vs. clean cages
  - Unsanitary
    - Multiple cycles of chicks raised in the same cages
    - Feces, dust and dander allowed to accumulate
  - Clean
    - Cages steam cleaned between cycles
    - Bedding changed 3 times per week

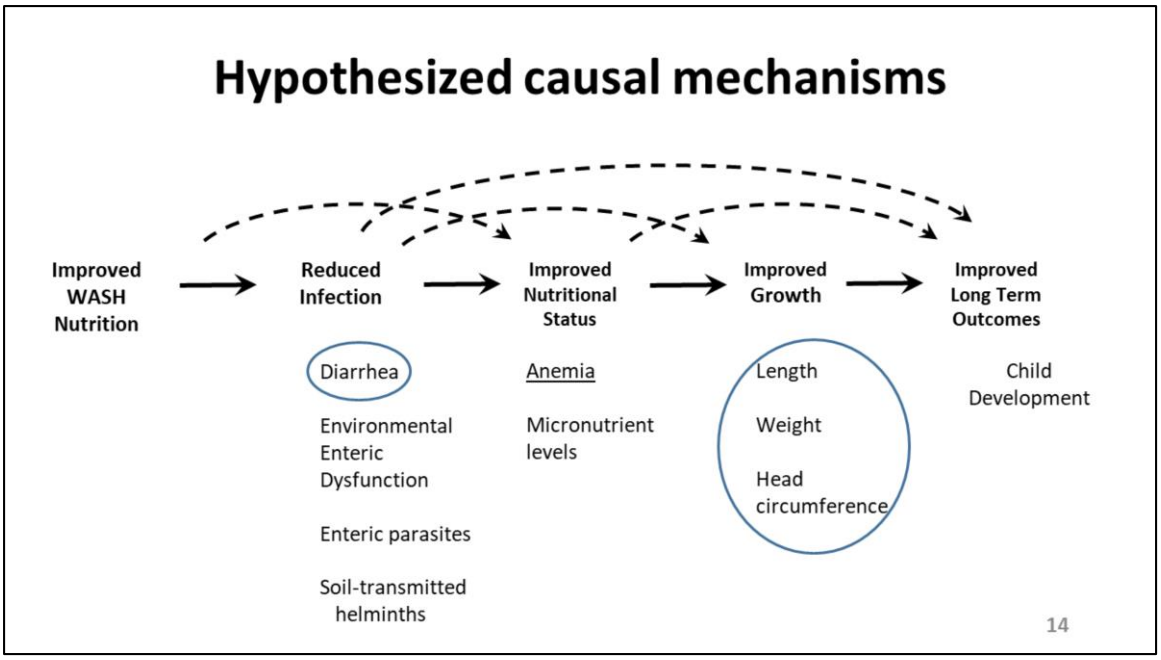
www.farmsanctuary.org

Roura E. *J Nutr*. 1992 Dec;122(12):2383-90.

So another observation related to this line of thinking is that looking at how animals grow in a clean versus dirty environment. In a randomized trial of chickens, where the outcome was feed efficiency, or grams of weight gain per grams of feed, when chickens were growing in unsanitary versus very clean cages, in the unsanitary cages multiple cycles of chicks raised in the same cages were observed, where feces dust and dander were allowed to accumulate. And in the clean cages, the cages were steam cleaned between cycles and the bedding was changed multiple times per week.



In the clean cages, there was a 22% gain in the feed efficiency in the clean environment, suggesting the cleanliness of the environment really mattered.

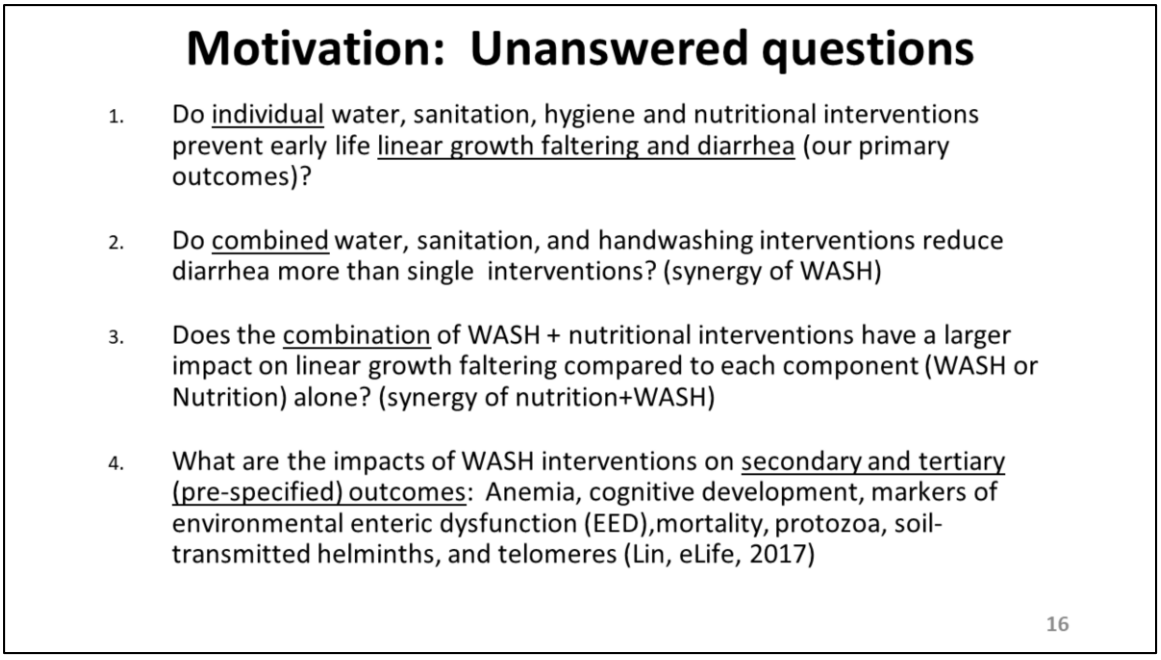


So in WASH benefits, we were studying many different potential outcomes resulting from water sanitation hygiene, or WASH interventions along with nutrition interventions, and this diagram is meant to show kind of a flow of the impact we hypothesized might occur. With improved WASH nutrition we would expect to see reduced infection, and reduced infection would be specifically measured by the things that you see here like diarrhea, environmental enteric dysfunction, enteric parasites, or soil transmitted helminths.

The reduced infection might itself lead to improved nutritional status, which would be measured by an improvement in anemia and micro nutrient levels in the children. And this might in turn lead to improved growth, which could be measured with specific outcomes like length, weight, and head circumference. And finally, all of these might lead to improved long-term outcomes, like child development.

Now you see these arrows at the top. These potential effects kind of skip over some of the intermediate steps. So improved WASH and nutrition might itself directly lead to improved nutritional status, it might not be required to go through the reduced infection stage and so forth. So this is just kind of an outline of the ideas that the trial was trying to test through its interventions, and you'll see what those interventions are specifically in a bit.

So let's talk about the study design of WASH benefits.



The study was a very large study, about a $30 million study funded by the Gates Foundation, that was setting out to answer a number of specific questions. Let me just read these and talk a little bit about them.

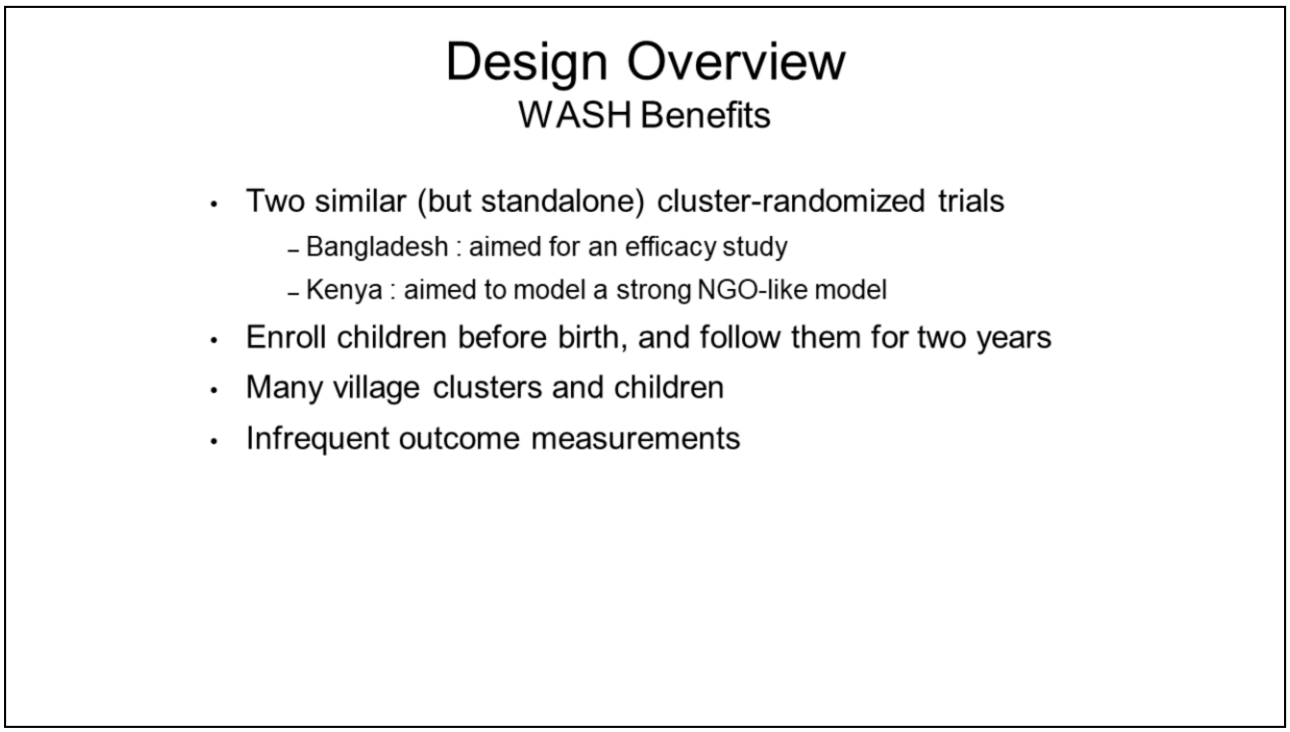
So the first is, do individual water sanitation, hygiene, and nutritional interventions prevent early life linear growth faltering and diarrhea? These were primary outcomes. So when I say individual, and I've underlined that there, I mean a water intervention alone, or a sanitation intervention alone, or hygiene intervention alone, or nutritional intervention alone. So those are four separate arms of this study, as you'll see.

Next question was, does a combination of water sanitation and hygiene interventions reduce diarrhea more than single interventions? So this question is asking about the synergy of water sanitation and hygiene together. Notice that nutrition is not part of this question.

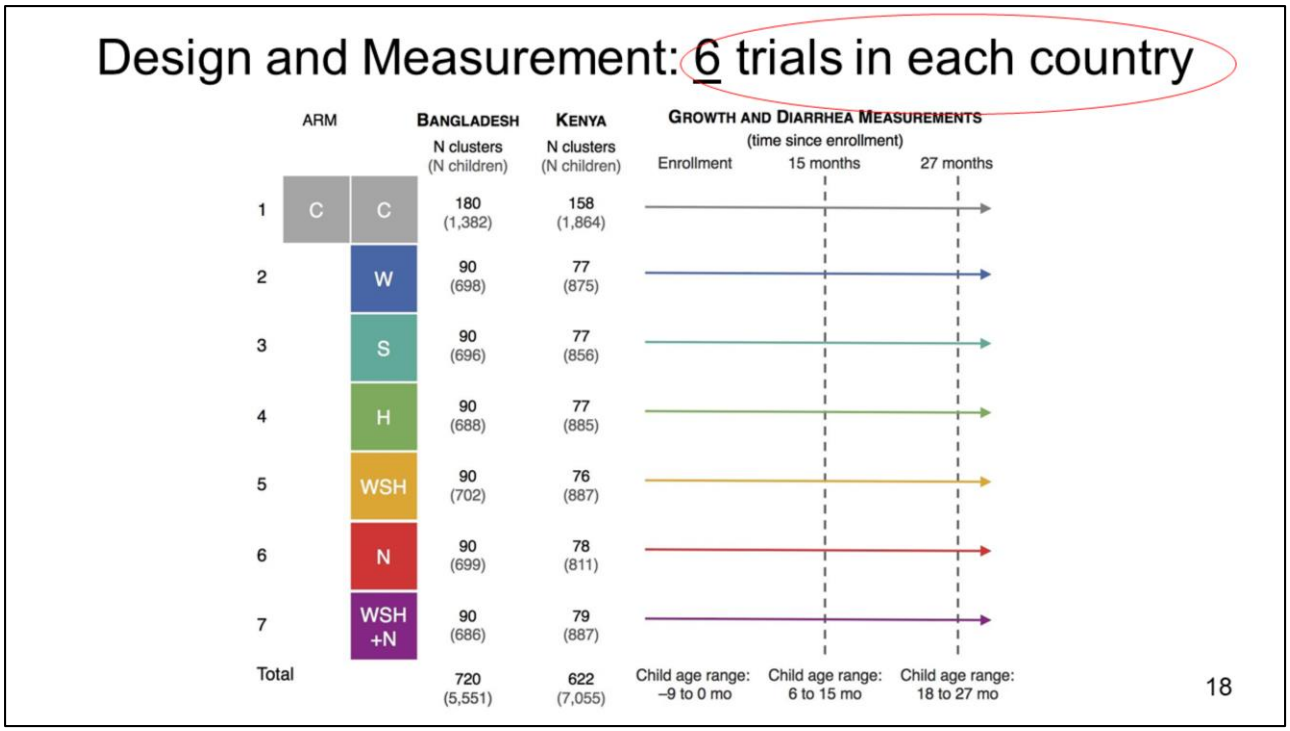
Next, the question is does the combination of all of the wash interventions, not individually but all together in a package, plus nutritional interventions, have a

larger impact on linear growth faltering compared to each component. The two components here are the WASH and the nutrition. Each of them alone, or the comparison of nutrition plus wash, to WASH alone or nutrition alone.

And finally, what are the impacts of WASH interventions on secondary and tertiary or pre-specified outcomes, like anemia, cognitive development, markers of environmental enteric dysfunction, as we discussed, mortality or death, protozoa, soil-transmitted helminth, and telomeres.

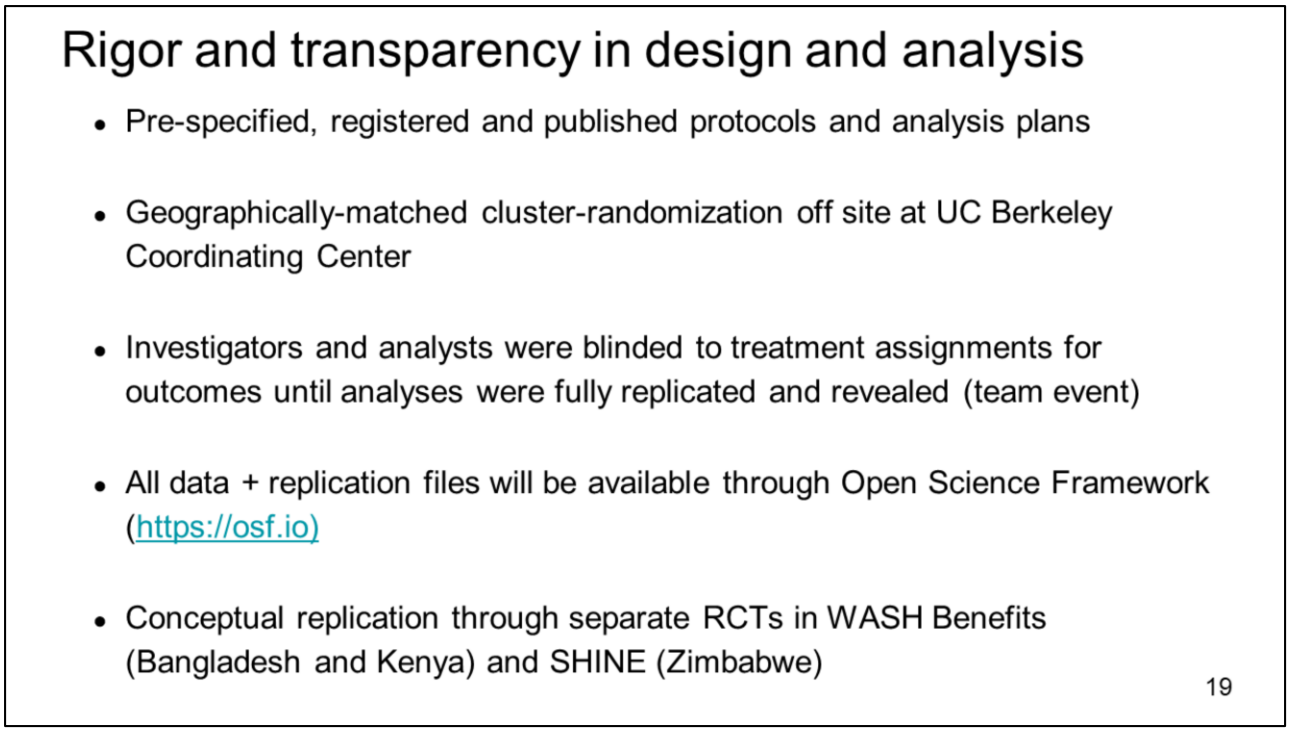


The basic design was these were two similar but standalone separate, cluster-randomized trials. One done in Bangladesh to essential model an efficacy study or a kind of best possible case study, and in Kenya it was designed to model what a strong NGO-like model would look like. Like a realistic rollout of these interventions in a big population. So we enrolled children before birth and followed them for two years. There were many village clusters and many children, and we had infrequent outcome measurement on purpose. So we measured children early, middle, and end of the two-year study, but we weren't measuring them constantly week-to-week week or month-to-month.



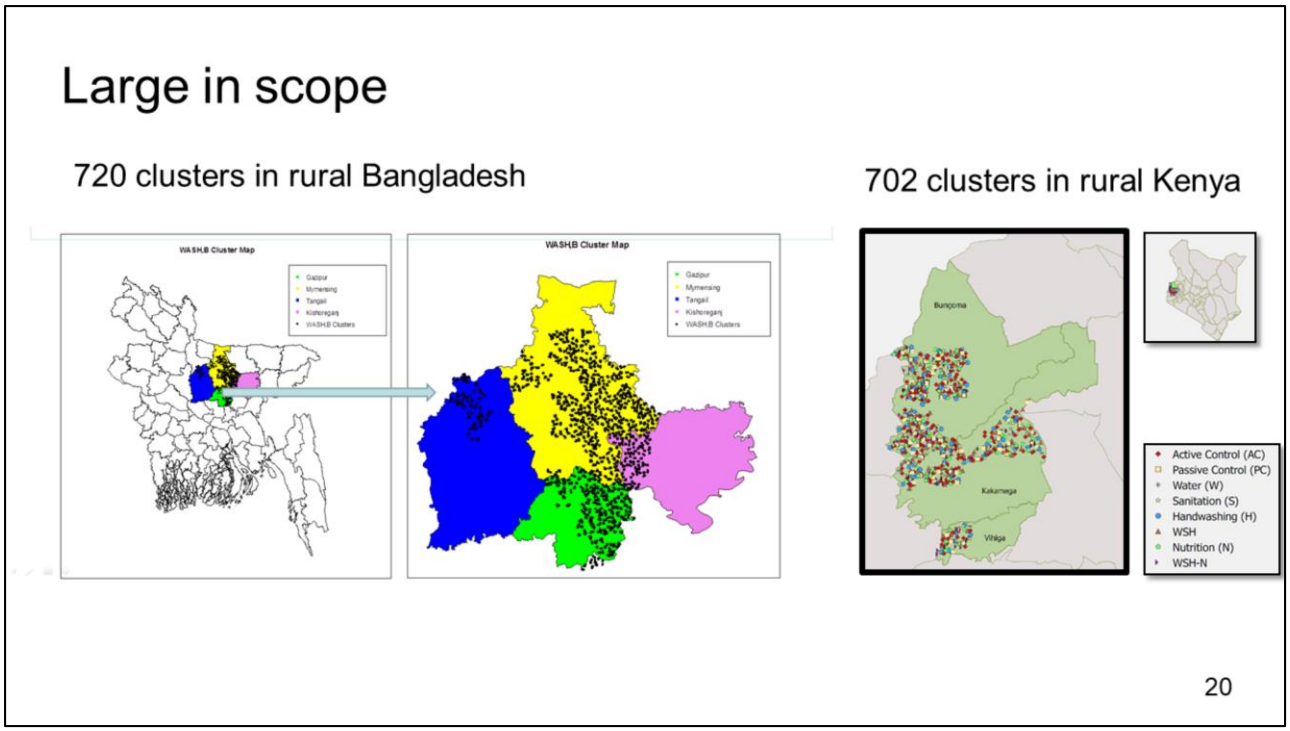
Here's a picture of the design of the study. So each row in this picture is an arm of the study. So the first arm is the control arm, and this arm has two boxes because it's meant to represent that it was twice the size of all the other arms, and that's because the control arm is going to be compared against all of the other individual arms. You see the second arm was the water only arm, third arm sanitation only, fourth arm hygiene only, fifth arm the wash arm, combination of water sanitation and hygiene. The sixth arm was the nutrition arm. The seventh arm was the WASH combination plus nutrition, so sort of a super combination of all the interventions.

And you see on the chart when growth and diarrhea was measured. It was measured at about one year, roughly 15 months, and about two years, roughly 27 months into the study. And I should mention that the way children were enrolled in the study was to enroll their pregnant mothers. So the women were enrolled when pregnant, and so the kids were not yet born. So there were no measurements of growth and diarrhea at the beginning the study. Those were done, as I said, at the midpoint and end of the study.



So one important feature of this study that was all of these steps were taken to try to reduce bias with the following. We pre-specified, registered, and published all the protocols and analysis plans. We geographically matched and cluster-randomized off site at the UC Berkeley coordinating center. Our investigators and analysts were blinded to treatment assignments for outcomes until the analyses were fully replicated or revealed at a big team event.

All the data and replication files will be available through the Open Science Framework, with the link there. And then conceptually, those studies were replicated because we get one of these in Bangladesh, one in Kenya, and then a completely different team did the same sort of study but designed it separately from us in Zimbabwe, and these were all funded by the Gates Foundation.



So the studies were large in scope, with 720 clusters in rural Bangladesh and 702 clusters in rural Kenya. So that's a point to notice that these were all rural studies. These are not about urban interventions.

Community Promoters

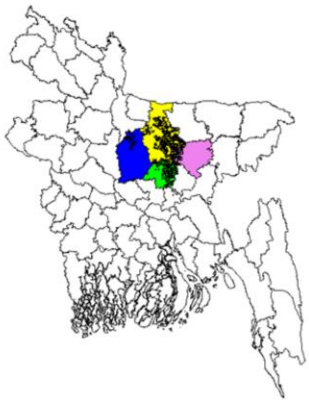
- Merit based hiring
- Trained by supervisors
  - 5 day initial session
  - Monthly 6 hour meetings
    - Grouped by interventions
    - Develop promoter's problem solving skills
    - Built *espirit d'corps*
- Payments via mobile phones

Community promoters were hired, trained, and rigorously screened and prepared to go out to the different communities to teach these interventions.

So there was merit-based hiring for these promoters. They were trained by supervisors with a five day initial session. There were monthly six hour meetings, grouped by the different intervention arms, that were designed to develop the promoters problem solving skills and build an esprit d'corps, and payments to the promoters were made via mobile phones.

Participant enrollment

- Canvassed study area seeking women in their 1st or 2nd trimester of pregnancy.
- Mapped the location of pregnant women
- Identified cluster of 8 pregnant women
  - who could be reached by a single health promoter on foot
  - Separated from nearest cluster by a 1 kilometer buffer zone
- After 8 clusters identified
  - Cluster ID numbers assigned
  - Off site statistician randomly assigned each cluster to one of 6 interventions; with 2 clusters assigned to control

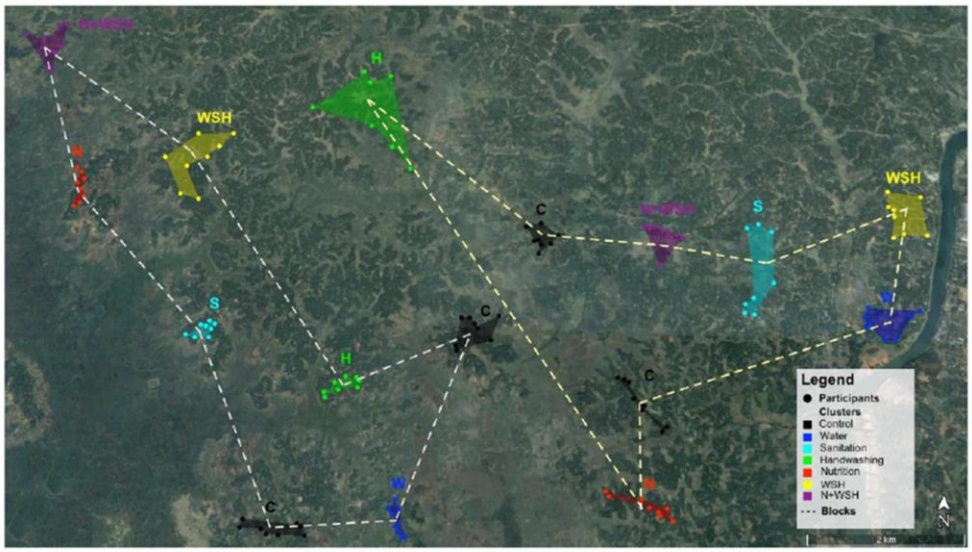


Thanks to Sania

To enroll participants, we canvassed the study areas seeking women in their first or second trimester of pregnancy. We mapped the location of the pregnant women. We identified a cluster of eight pregnant women in each area who could be reached by a single health promoter on foot who were separated from the nearest cluster by a 1 kilometer buffer zone.

After eight clusters were identified in each area, cluster ID numbers were assigned, and then an off site statistician randomly assigned each cluster to one of the different interventions, with two clusters assigned to control for the double-sided control arm I mentioned.

Geographically & temporally matched clusters



The clusters were geographically and temporally matched in these areas. So for instance, in this map you see the two different clusters here with the different arms of the study.

Outline

1. Stunting and child health
2. Environmental enteropathy - the true cause of stunting?
3. WASH Benefits causal hypothesis
4. WASH Benefits study design
5. **WASH Benefits interventions**

And finally, let's talk about what the specific interventions are in WASH benefits.

Water Quality Interventions

Chlorine tablets to treat drinking water (NaDCC)

+

Safe Storage

So there were different water quality interventions to improve the water quality, and these were clearing tablets used to treat the drinking water, plus a safe storage vessel. Here you're seeing a picture from Bangladesh.

Sanitation Interventions

Improved latrine

Child potty

Sanitary scoop

The sanitation interventions include improvements to the latrines, the use of child potties, and sanitary scoops to help families remove the fecal material from around their houses, generally from animals.

Handwashing Interventions

Handwashing station + soapy water bottles

Hand-washing interventions were provided with hand-washing stations and soapy water bottles.

Nutrition Interventions

Nutritional Promotion

- Exclusive breastfeeding through 6 months
- Continued breastfeeding through 24 months
- Diverse nutrient dense weaning foods

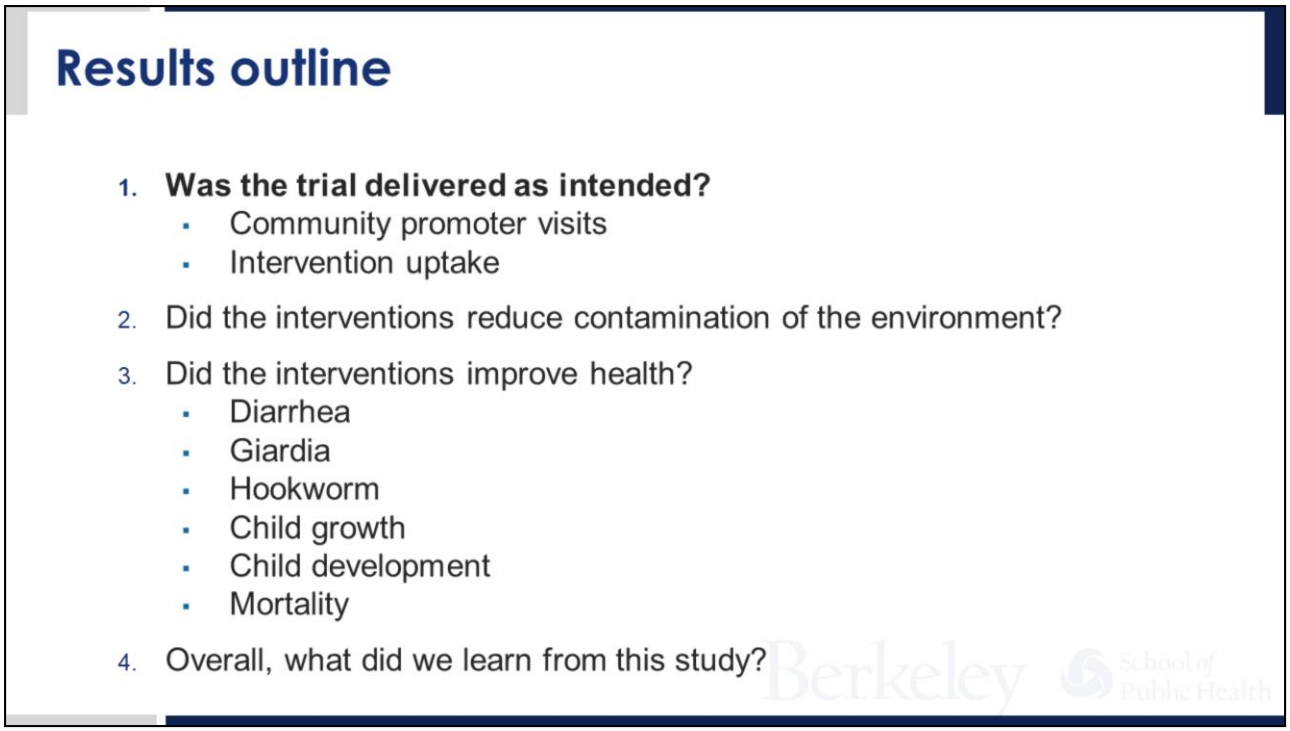
+

Daily lipid based nutrient supplement

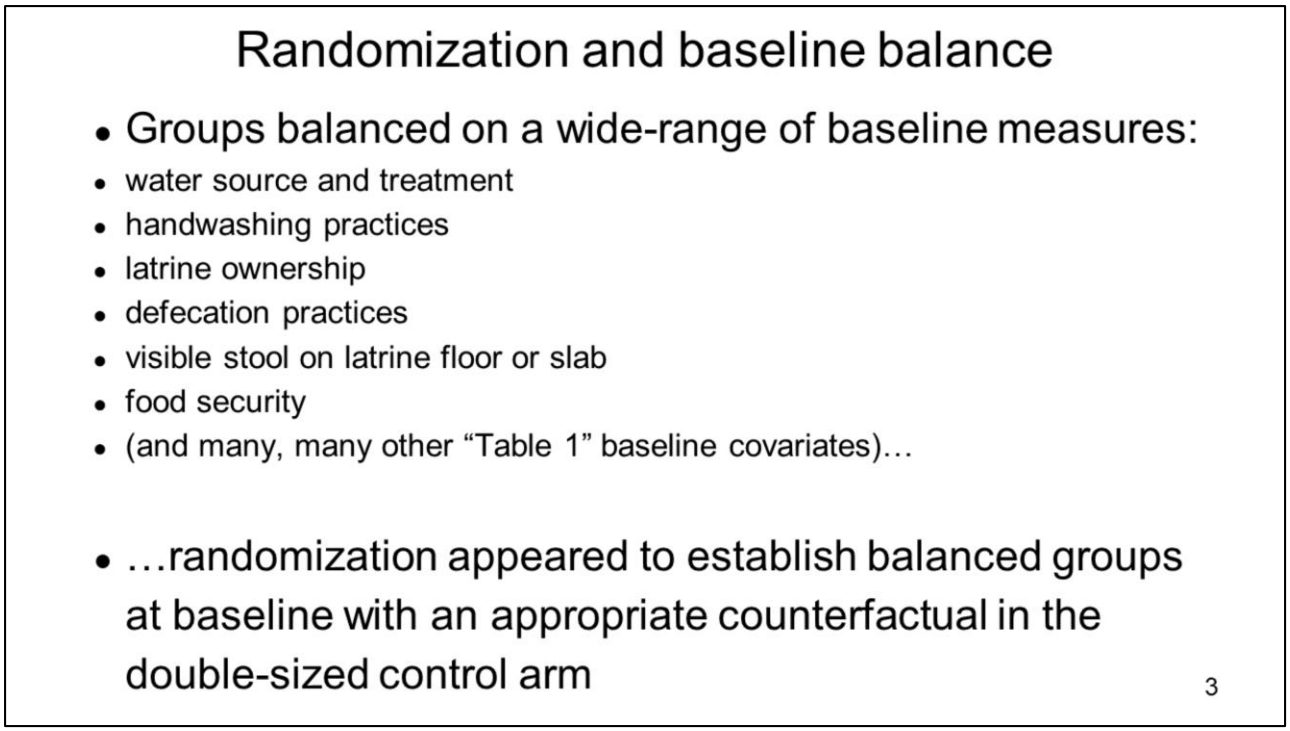
- 6 – 24 months
- 10-gm sachet twice daily
  - 118 Kcal
  - 9.6 gm fat
  - 2.6 gm protein
  - ≥100% RDA of 12 vitamins
  - 9 minerals

And the nutrition interventions included both nutritional promotion, with exclusive breastfeeding up through age six months, which was the national recommendation, continued breastfeeding through 24 months, diverse nutrient-dense weaning foods, and then the supplement that I discussed earlier was given from ages six to 24 months in a 10 gram sachet twice daily, with the components that you've see before.

JACK COLFORD: Let's now talk about the results from the WASH Benefits trial, the randomized controlled trial of water sanitation, hygiene, and nutrition interventions in Bangladesh and Kenya.

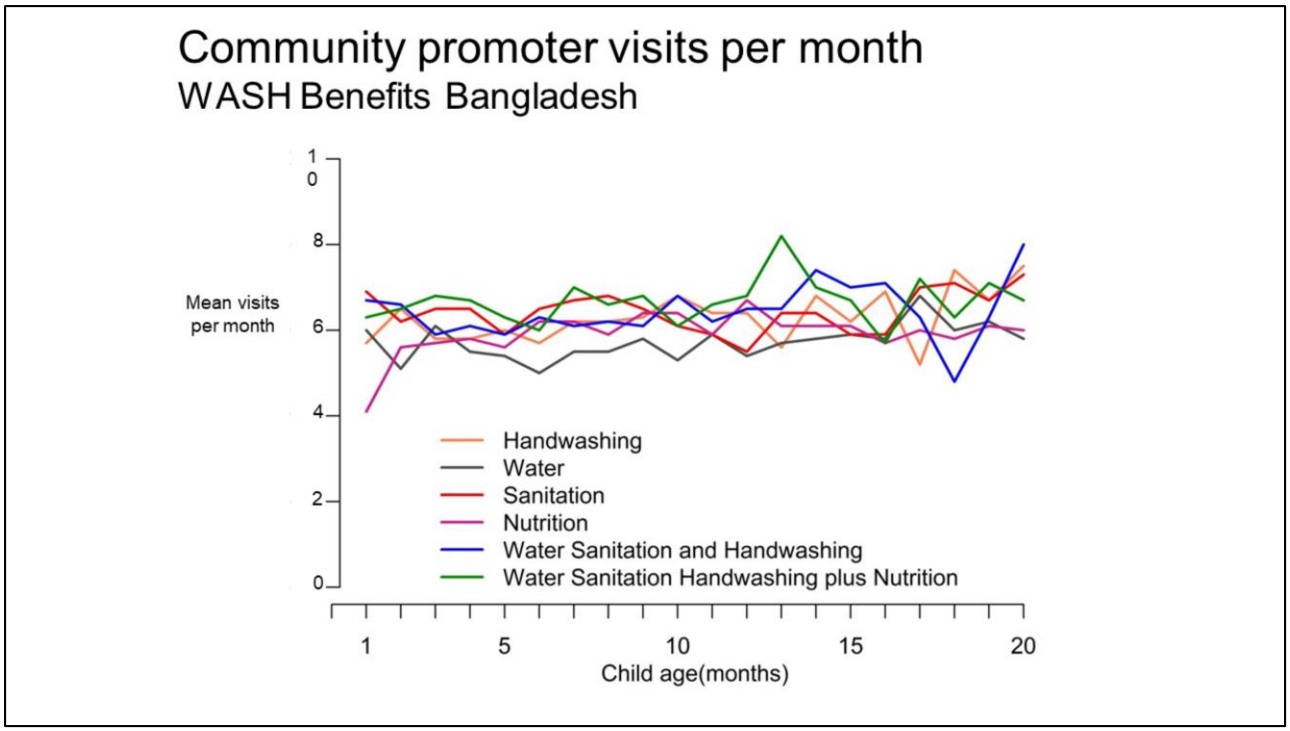


The way we'll organize this is to first talk about whether the trial was delivered as intended. Then I'll talk about whether the intervention reduced contamination of the environment. And then next, whether the intervention's improved health and all the different ways we talked about measuring health. And overall, what did we learn from this study?

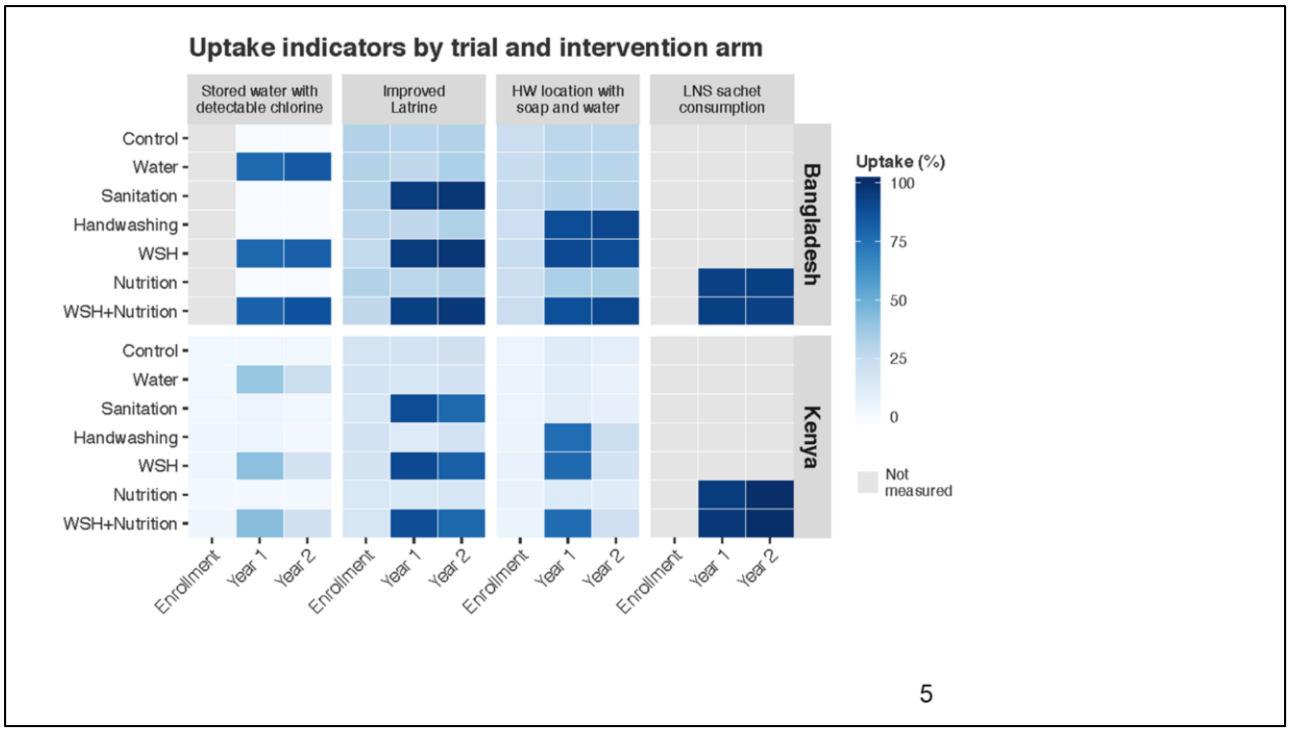


Let's look first at randomization and baseline balance. The groups were balanced well on a wide range of baseline measures, such as water source and treatment, handwashing practices, latrine ownership, defecation practices, visible stool on the latrine floor or slab, and food security, and many, many other Table 1 baseline covariates if you look in the papers.

And this is a goal in a trial, is to see if your randomization led to groups that are serving as balance to each other, so that any difference you see between them should just be attributable to the intervention and not to baseline differences. Our randomization appeared to establish well-balanced groups at baseline with an appropriate counterfactual in the double-size control arm.



- This is a graph showing the community promoter visits in Bangladesh and WASH Benefits Bangladesh trial. You can see there were quite a number of promoter visits each month in all the different groups. And this surprised us that the promoters went so often. They weren't instructed to go. They didn't need to go this frequently, but chose to. They really became quite enthusiastic about the trial.

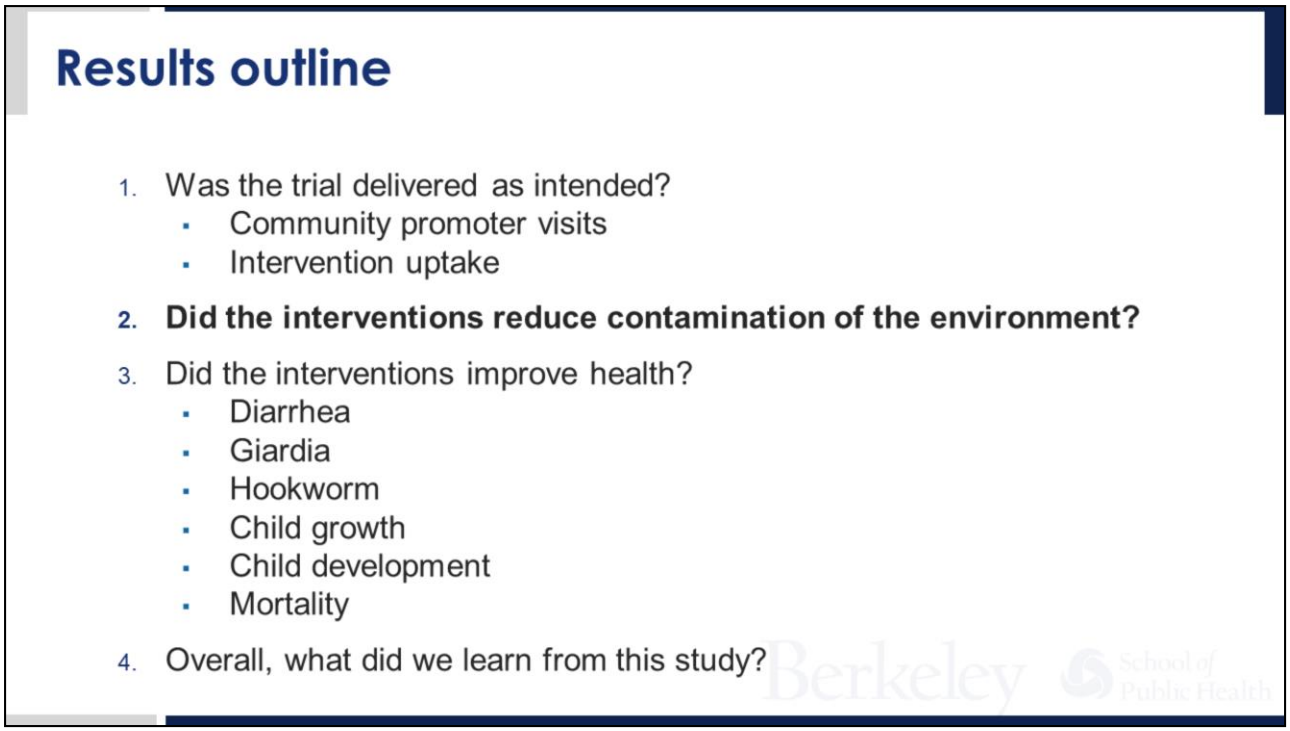


This heat map is showing the uptake of the different interventions and the different trial arms. The way to read this for the two countries, Bangladesh is on the top, Kenya's on the bottom. Let's just focus on Bangladesh for the moment.

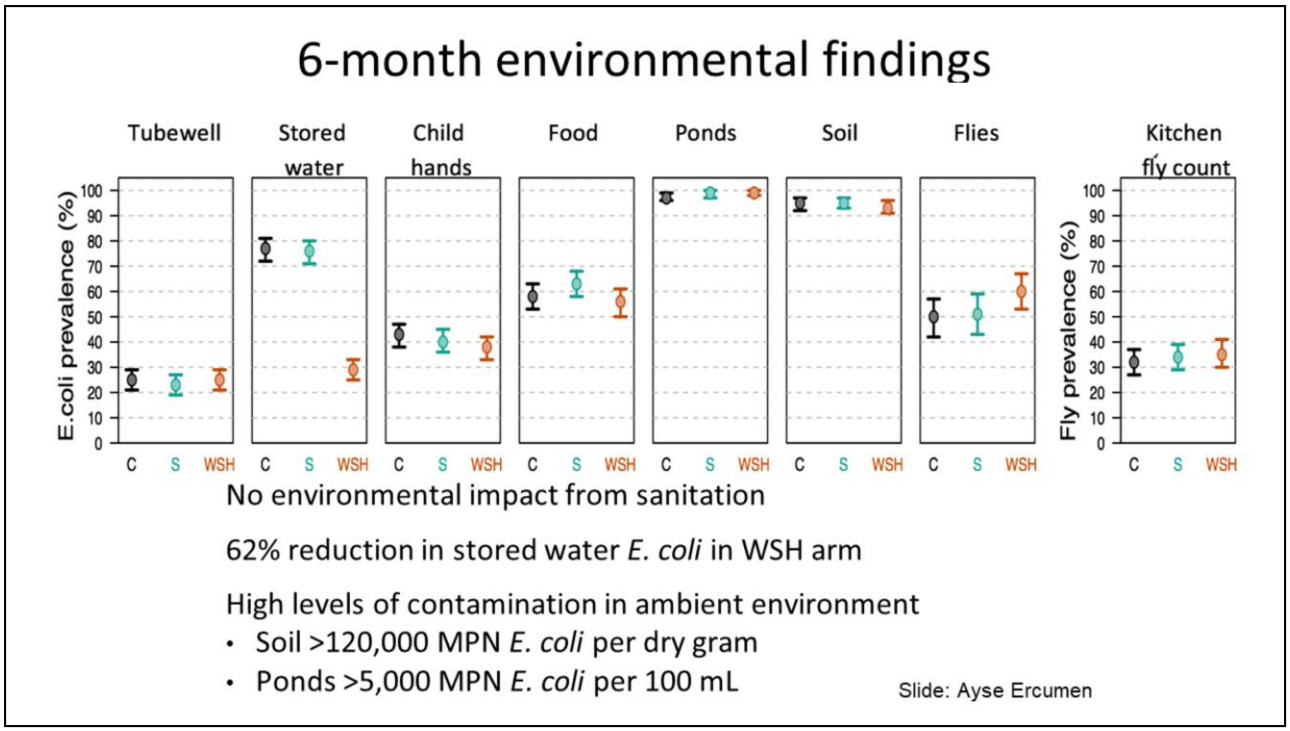
The darker the blueness of the box, the stronger the uptake percentage. And you see the scale over on the far right. In the upper half of the heat map, what we're seeing is the different arms of the trial. Control, water, sanitation, handwashing, et cetera, and showing how well the interventions were taken up an enrollment year 1 and year 2. Let's just follow one of these.

In the water, sanitation, and hygiene arm-- that is the fifth row of the heat map on the top-- you see that enrollment, there was no detectable chlorine into the stored water, and that's to be expected. But by year 1 and year 2, there was quite high uptake.

And similarly with the improved latrine, as was expected. The handwashing location was soap and water. At enrollment, there was some soap and water, but much more at year 1 and year 2. And finally, Lipid Nutrient Supplement, the LNS sachet consumption was very, very low or not measured at enrollment year 1 and year 2 because it just wasn't being distributed in the water, sanitation, and hygiene arm.



Did the interventions reduce contamination of the environment? Because if these interventions work, they should have improved the environment. That was our hypothesis.

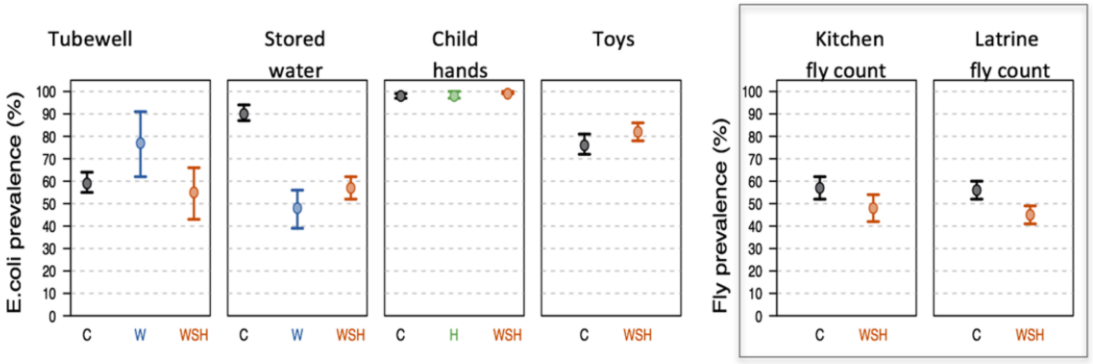


In this series of figures, we're seeing the environmental findings from different measurements within the environment, whether they be tube wells or water that was stored, on the child hands, and so forth. Let's just look at how to read some of these.

For example, look at the stored water. In the control and sanitation arms, the first two results there in the second box, we see that there was a very high prevalence of E. coli in the stored water, as would be expected. But in the water, sanitation, hygiene arm, when we measured it there was a great reduction in E. coli in the environment. But if you look at child hands, we didn't seem to have much change from control to sanitation or to water, sanitation, and hygiene from our interventions.

In the WASH arm, the combination WSH arm, there was a 62% reduction in the stored water E. coli. What we're seeing here is very high levels of contamination in the ambient environment, with greater than 120,000 E. coli per dry gram in the soil and greater than 5,000 E. coli per 100 cc in the ponds. These are very contaminated areas. No surprise that people are ingesting and exposed to these organisms with such high levels of contamination.

One-year environmental findings



37-47% reduced in stored water *E. coli* in water and WSH

16-19% reduced flies near latrine and kitchen in WSH

Slide: Ayse Ercumen

At one year, we see slightly different findings. Let's just look at the stored water again. Here we still see good reduction in the E. coli found in water in the arm that had water, sanitation, and hygiene intervention. Some reduction in flies near the latrine and the kitchen. We seem to have some impact on cleaning up the environment a bit.

Two-year environmental findings

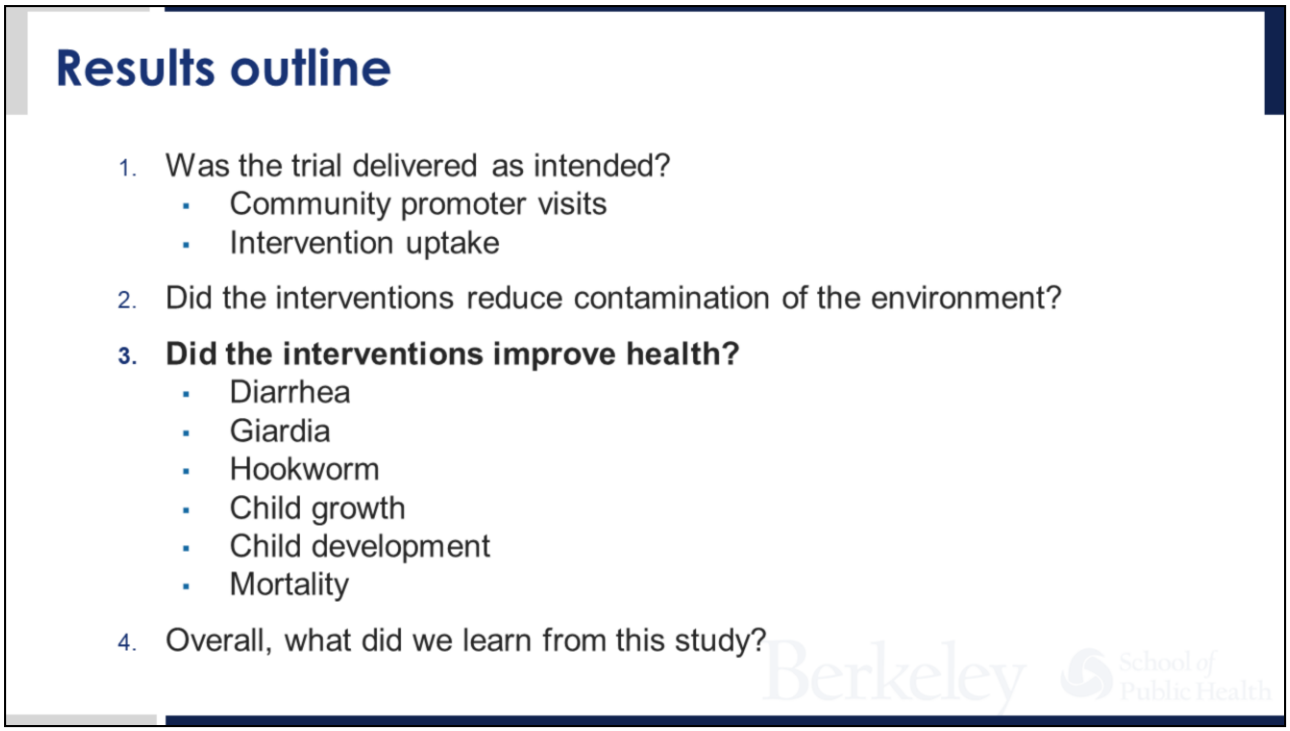
35-49% reduced stored water *E. coli* in water and WSH

30-32% reduced food *E. coli* in water and handwashing

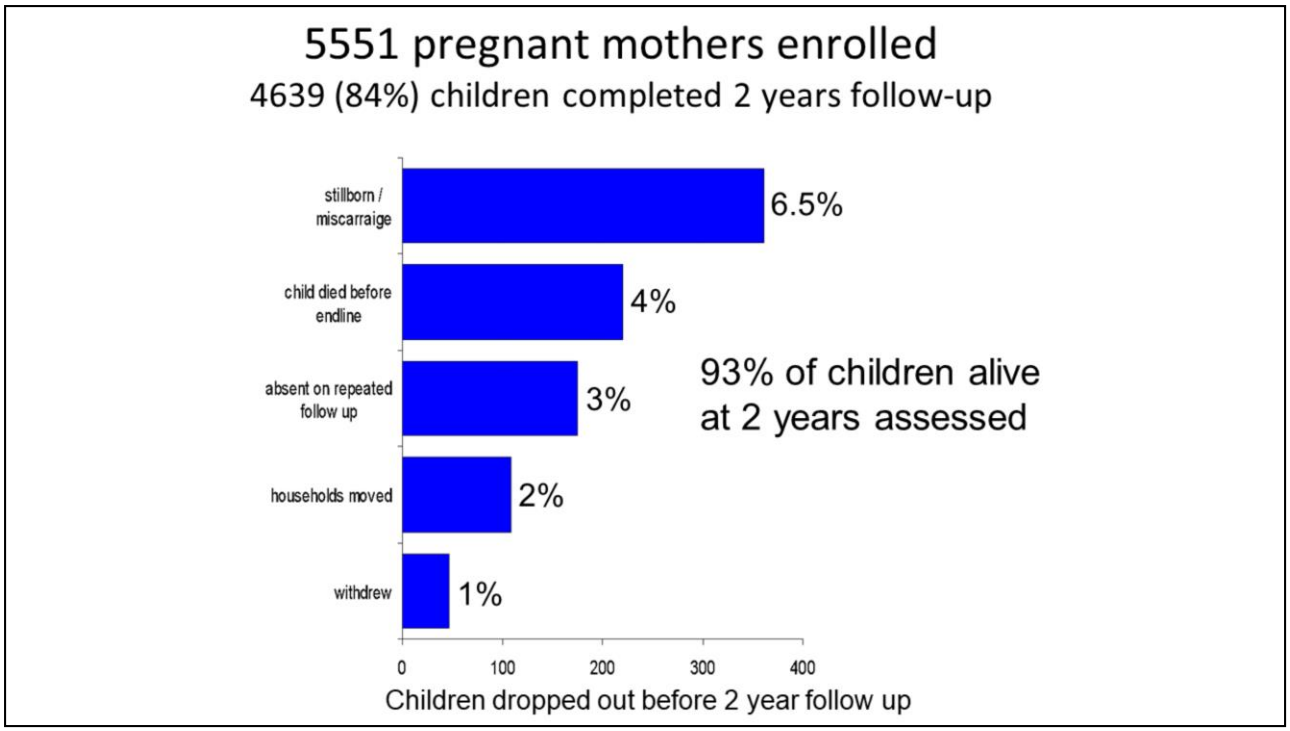
11% borderline reduced food *E. coli* in WSH

Slide: Ayse Ercumen

And at two years, we see, again, some reduced stored water E. coli in the water in WSH arms, and then some reduced food E. coli in the water and handwashing arms, and a borderline reduction in food E. coli in the combination arm.

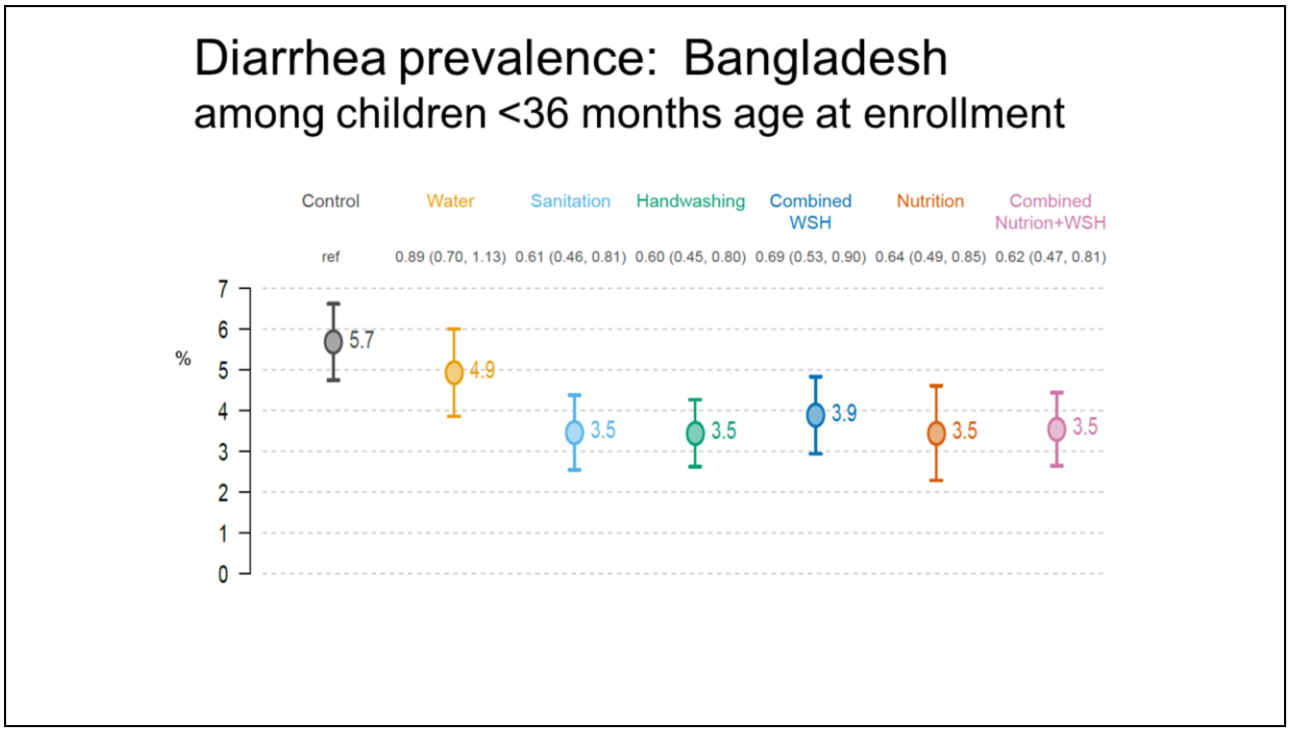


Key question. Did these interventions improve health? And we're going to look at several different outcomes.



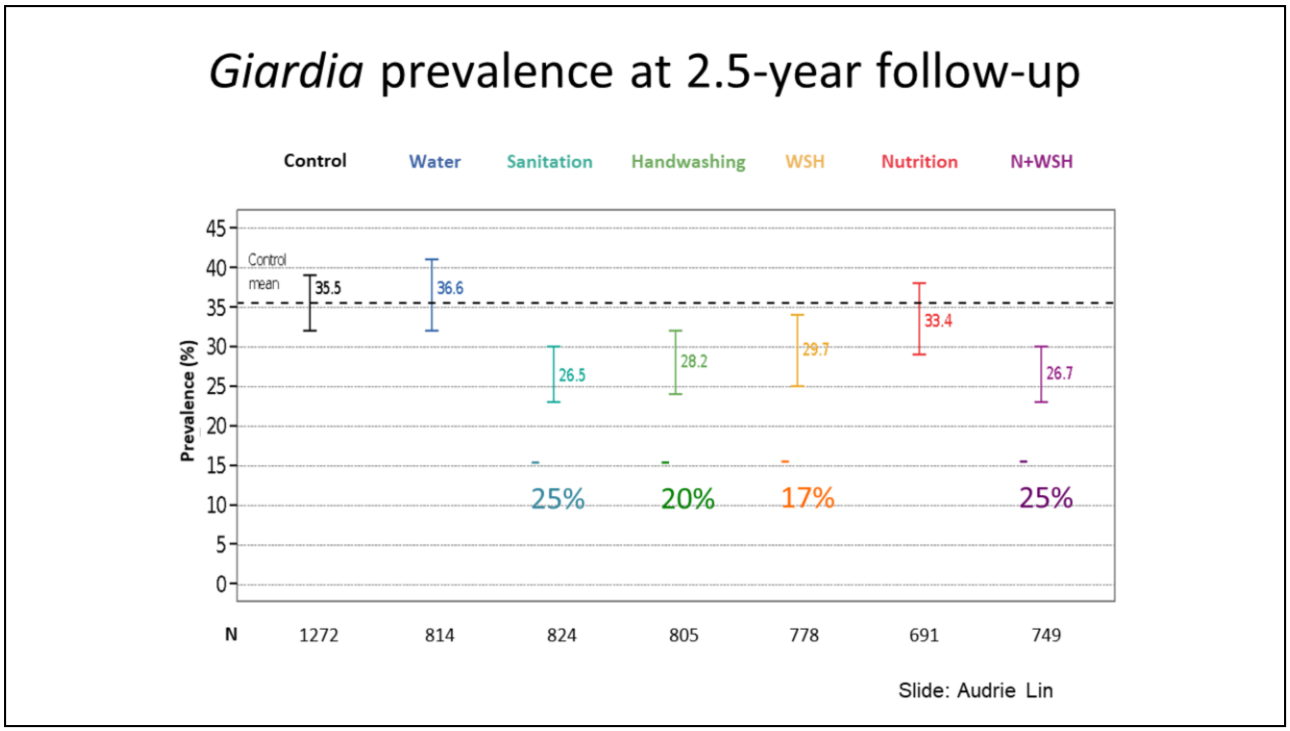
Looking at the 5,551 pregnant mothers who were enrolled, we had a very good completion rate at two years where we had information after two years on 84% of the children. You see what happened to a small number of children. There were stillborn, miscarriages, or children who died before the end line, children who were absent on repeated follow-up, households that moved or withdrew from the study.

But we lost very few children, which means that we had confidence in our results, because we had such good completion and follow-up of these study participants. 93% of the children were still alive at two years when they were assessed.



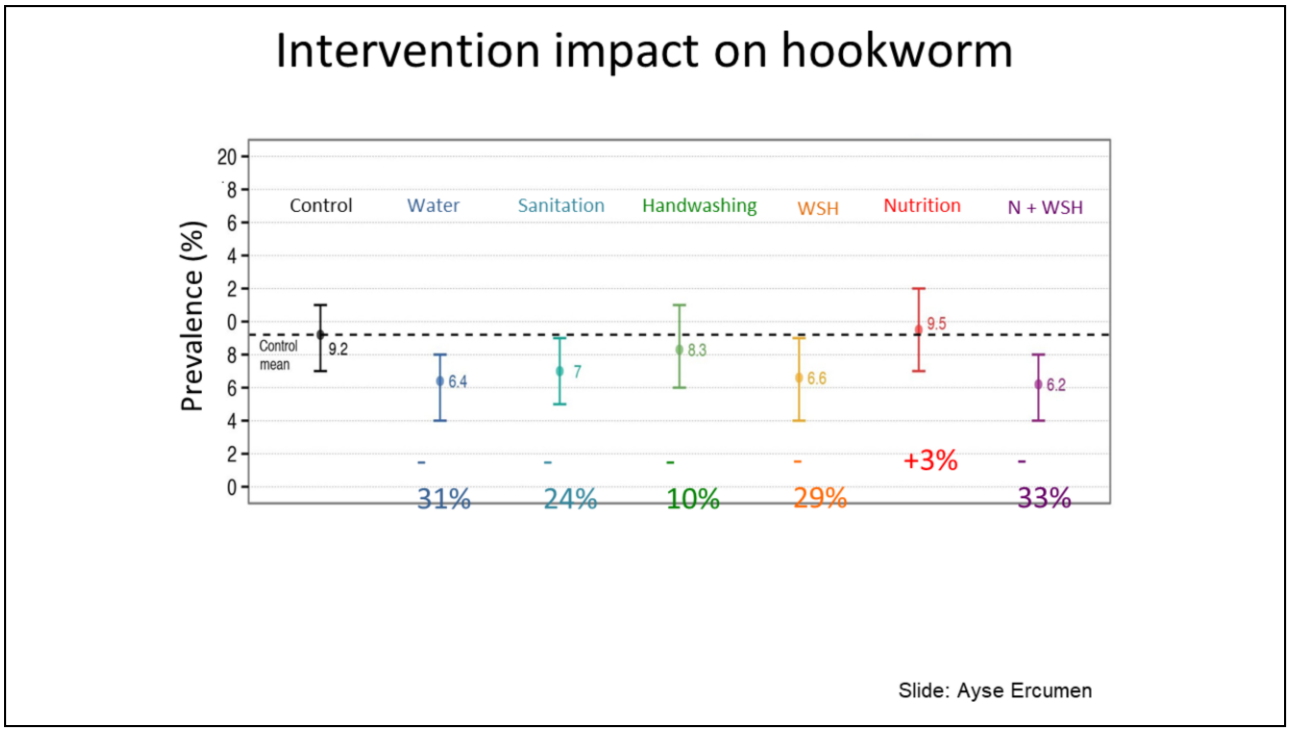
Let's look at what happened to diarrhea. The way to read this graph is the very left-most column is the control arm. So the prevalence of diarrhea was 5.7% in the control arm. And you see the measurements in all the other arms of the study, and the confidence interval showing the relative risk reduction. Let's just interpret a couple of these.

Let's look at the combined water, sanitation, and hygiene arm, the fifth column over. The number given is 0.69. That means the prevalence of diarrhea in the combined WASH arm was 69% of what it was in the control arm. That's a reduction from 5.7 to 3.9. This was statistically significant, because the confidence interval from 0.53 to 0.90 doesn't cross the null value of one.

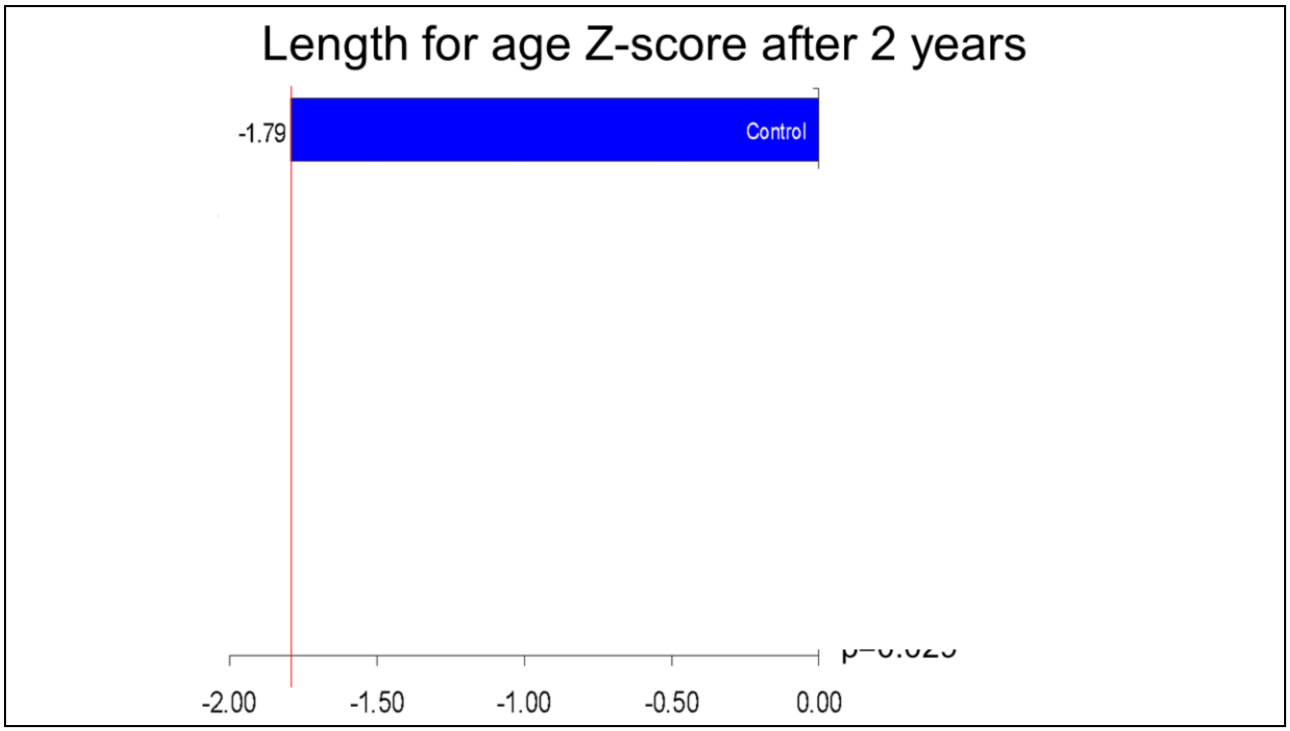


How about Giardia prevalence? Giardia is an organism in the gut that might be improved by these environmental interventions. And these results are very similar to the diarrhea interventions. In the sanitation arm, we saw a 25% reduction in Giardia. The handwashing arm, a 20% reduction. The WASH arm, a 17% reduction. And in the nutrition plus WASH arm, a 25% reduction.

We did not see a reduction in the nutrition arm from control. We went from 35.5 in controlled to 33.4 in nutrition. One wouldn't expect to see a reduction here, because nutrition shouldn't have any impact on Giardia, we wouldn't think, directly. However, we were a little surprised that in the water arm, there was no impact on Giardia. You see that in the second column.



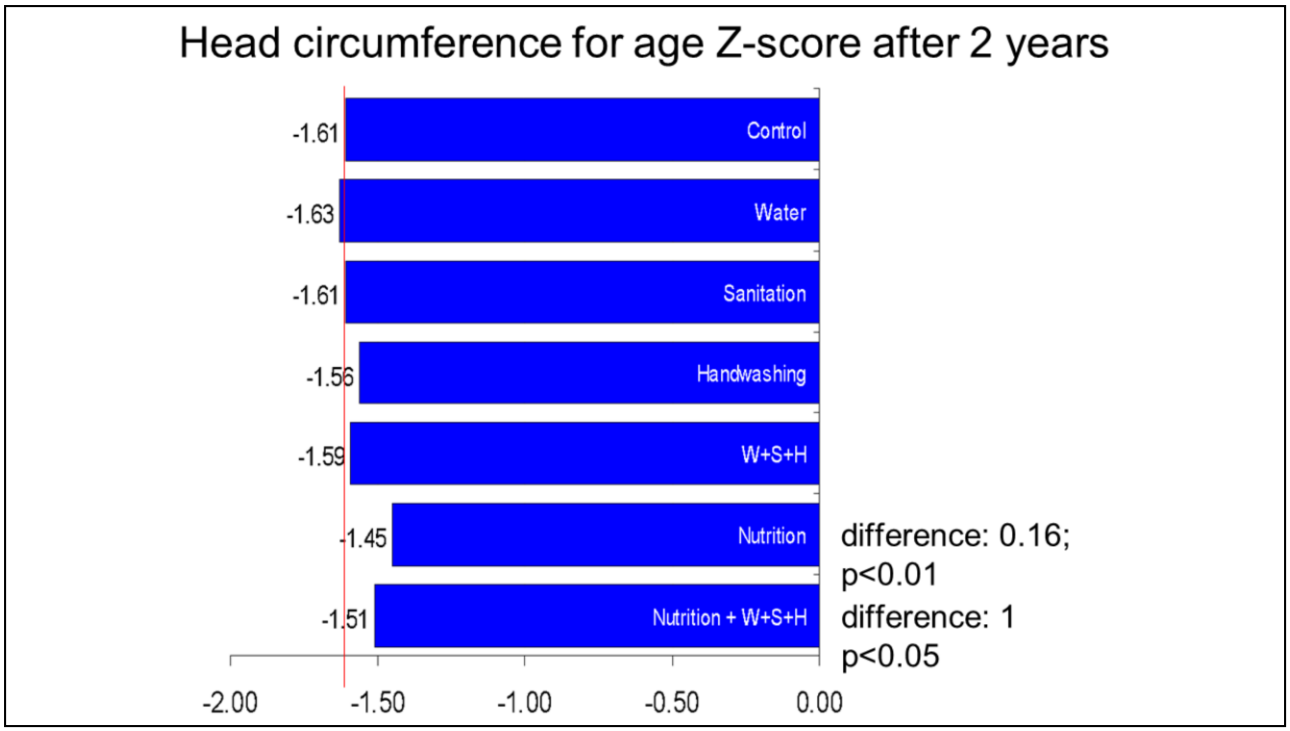
- We also looked at the impact on various soil transmitted helminths, including hookworm. And you can see the impacts of various different arms on hookworm prevalence in this slide.



Now a key primary outcome in the study was the impact of the length for age z-score after two years. In the control group, after two years, the children were stunted. That is, they had a negative 1.79 HAZ score. Let's look at what happened to the different arms with respect to z-score after two years of these interventions.

The water sanitation, handwashing, and combined arms showed really no change compared to the control arms. There appeared to be no impact in those arms. However, in the nutrition arm, there was a statistically significant reduction. And this is a big reduction. Down to 1.53 z-score.

And in the nutrition plus wash arm, there was also a statistically significant reduction. Clearly, it appears that nutrition is making a difference, as we would expect, in child growth. But the water, sanitation, and hygiene interventions are not seeming to have any impact on their own, and this was quite a big surprise, and an important finding for the field.



Similarly for head circumference, we found statistically significant reductions only in the arms with nutrition and nutrition plus water, sanitation, and hygiene. In other words, the two arms that had the nutrition intervention showed a reduction. None of the other arms did.

Potential explanations of lack of impact of WASH interventions on growth

1. ~~Low uptake of interventions~~
2. Environmental fecal contamination is not a major contributor to growth faltering in Bangladesh
3. Environmental fecal contamination does contribute to growth faltering, but WASH Benefits Bangladesh interventions did not reduce environmental fecal contamination enough

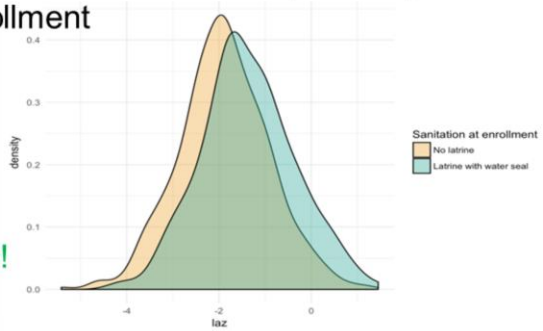
The potential explanations of the lack of impact of our WASH interventions on growth might be lower uptake of interventions than ideal, or environmental fecal contamination might not be the major contributor to growth faltering in Bangladesh. Or the environmental fecal contamination does contribute to growth faltering, but the WASH Benefits Bangladesh interventions didn't reduce the contamination enough.

Observational assessment in control group

Sanitation at enrollment	N children	(%)
No latrine	513	47
Latrine no water seal	391	35
Latrine with water seal	199	18

This is why we do trials!

Distribution of LAZ at 2 years, by sanitation at enrollment



Summary of the association between access to an improved latrine at enrollment and LAZ among the birth cohort at the final endpoint, Bangladesh

	Diff LAZ	Lower 95%	Upper 95%	P-value
Unadjusted	0.52	0.34	0.7	0.00
Adjusted, GLM	0.20	-0.01	0.4	0.06
Adjusted, double-robust TMLE	0.17	-0.06	0.4	0.15

A different analysis we did after the study was completed, and we think is quite interesting, we reanalyzed the children who were in the control group and treated them like an observational study. So looking at the children just in the control group and classifying them by their sanitation status, whether they had a latrine, a latrine with no water seal, or a latrine with a water seal, we see the number of children in each of those arms.

Then we analyzed again. This is just the control group intervention. These children didn't receive an intervention in the trial. They just are being classified by what their families were using.

And what we see is that at two years when you classify them by their sanitation status at enrollment, at time zero, the highest sanitation group, latrine with water seal, had a significant increase in their length for age z-score or their height for age z-score at two years. In this observational analysis, it looks like sanitation is very important, but we know that in the randomized trial that sanitation did not have an impact.

Here's the summary with numerical values of what happened to the change in length for age z-score when we adjusted and compared the children by their baseline sanitation status. And again, this is a post-hoc analysis, but it's interesting, because what it suggests is when one does an observational analysis, one is led to believe that sanitation is effective.

But in the trial, it was not effective. Clearly, we think the results of the trial are more credible, because in the observational analysis, there may be residual confounding. This is exactly why we do trials.

Child Development

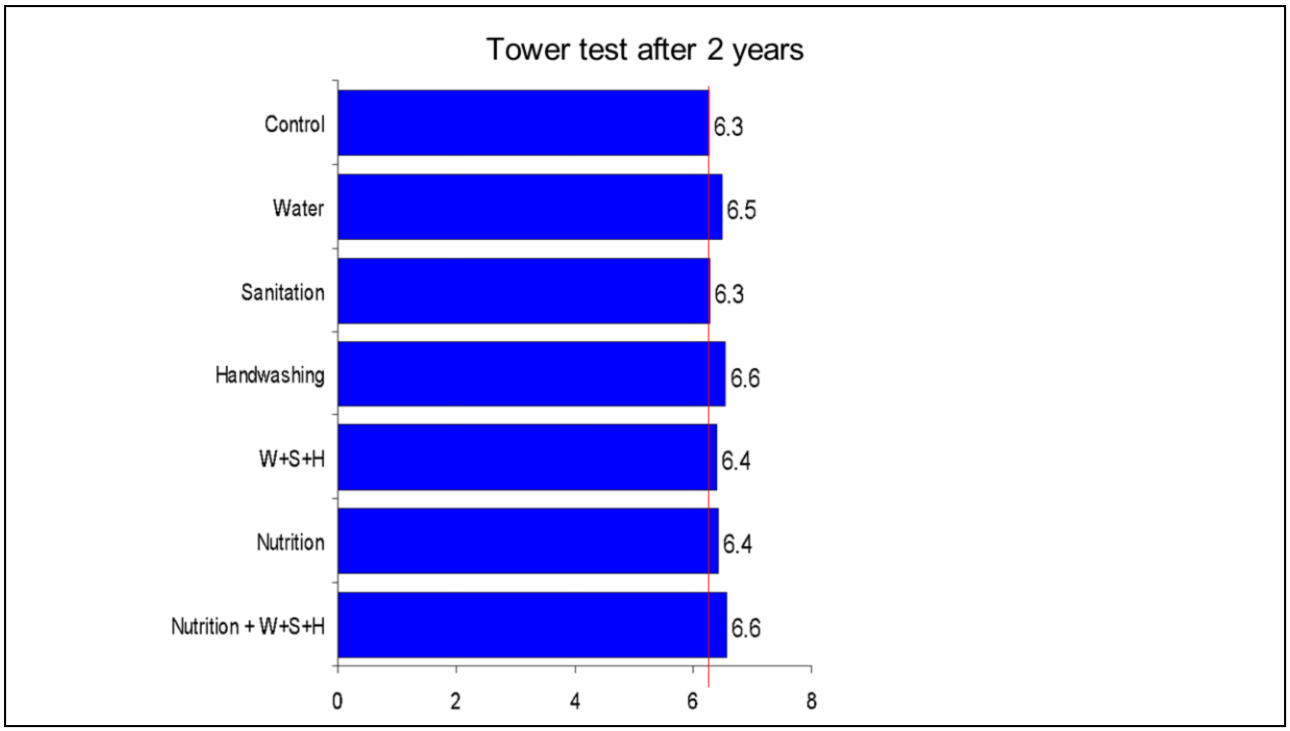
Tower Test

executive function / inhibitory control

- Tester has red blocks; child has yellow blocks
- Tester explains to child:
  - First I will place a block
  - Then you place a block
  - Then I place a block
  - Then you place a block
  - You have to wait until it is your turn
- Tester sums the number of times out of 8 that the child waits her turn and places a block

We also measured some child development tests in the children to look at executive function and inhibitory control. And in these sorts of tests, things like this happen. The child tester or examiner has red blocks. The child has yellow blocks.

The tester explains to the child, first I will place a block. Then you place a block, then I place a block, then you place a block. You have to wait until it's your turn. The tester sums up the number of times out of eight that the child waits her turn and places the block.

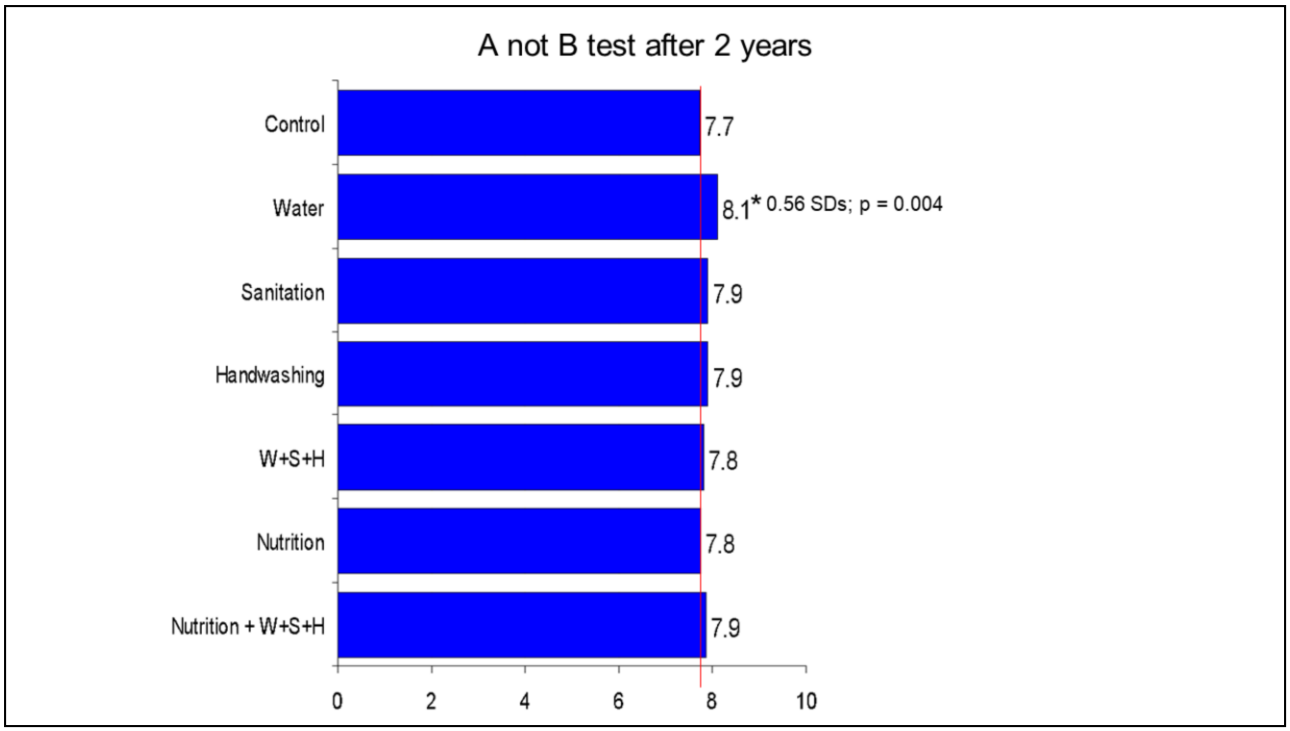


After two years, looking at this particular test, we don't see any differences or improvements in the accuracy of the children's block placement by this tower test.

A not B test
executive function / working memory

- The tester presents a child with a board with two shallow wells
  - Tester places a treat in one well and covers both wells with an opaque cup.
  - Tester distracts child for 5 seconds with a song
  - The tester prompts the child to find the treat
- Sum the number of times (out of 10) the child looked in the correct place for the treat

We also did something called an A not B test in which a tester presents a child with a board with two shallow wells. The tester places a treat in one well, covers both wells with a dark or opaque cup. The tester then distracts the child for five seconds with a song. The tester prompts the child to find the treat, and then counts up the number of times out of 10 attempts that the child looked in the correct place for the treat.



Once again, we don't see any difference between control group and the intervention groups with one exception. That's the water arm. So this is just one arm of the trial here that seemed to do better on the A not B test. We're not quite sure what to make of that.

Child Development

- Fieldworkers read each item to parent
- Record responses as
  - Yes
  - Sometimes
  - Not yet
- Some observational items
- Scores adjusted for
  - Child sex, child age, mother age, parents education, number of household members, number of household rooms, household roof, floor, wall materials, availability of electricity, type of fuel for cooking, household asset

GROSS MOTOR

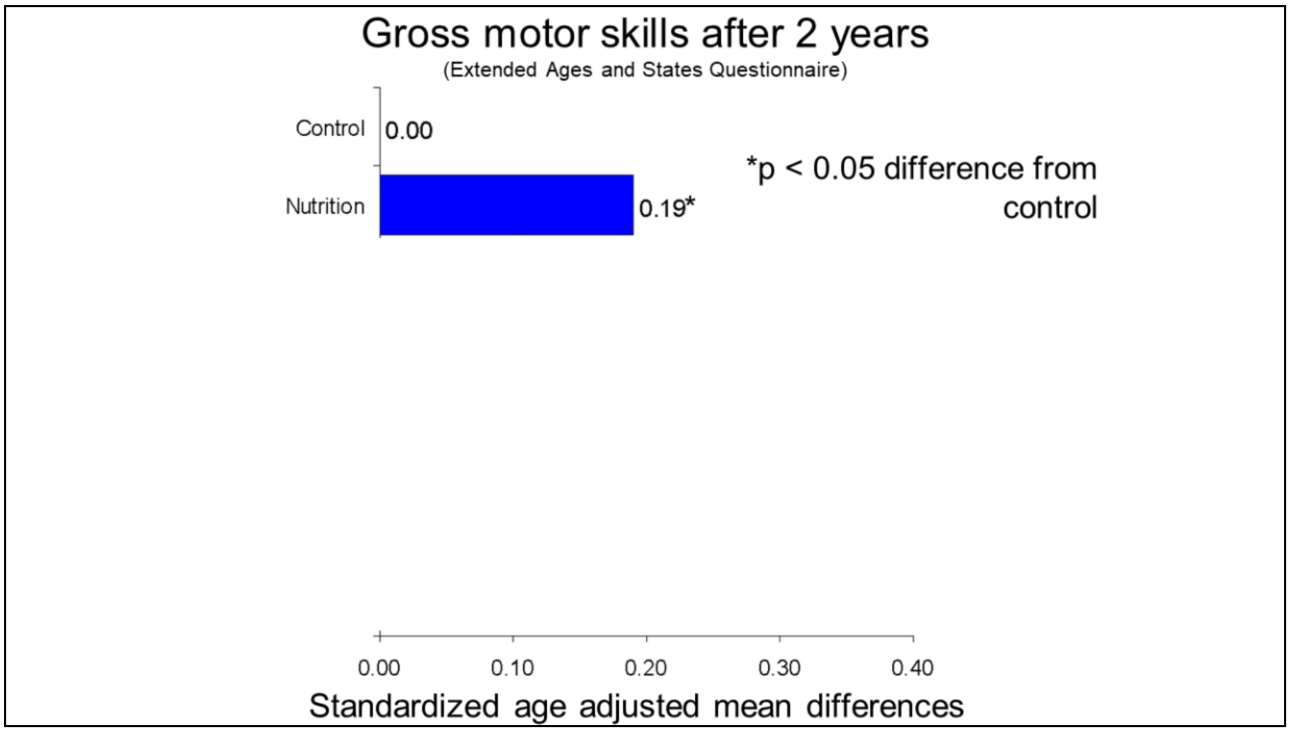
Does your child jump with both feet leaving the floor at the same time?

PERSONAL-SOCIAL

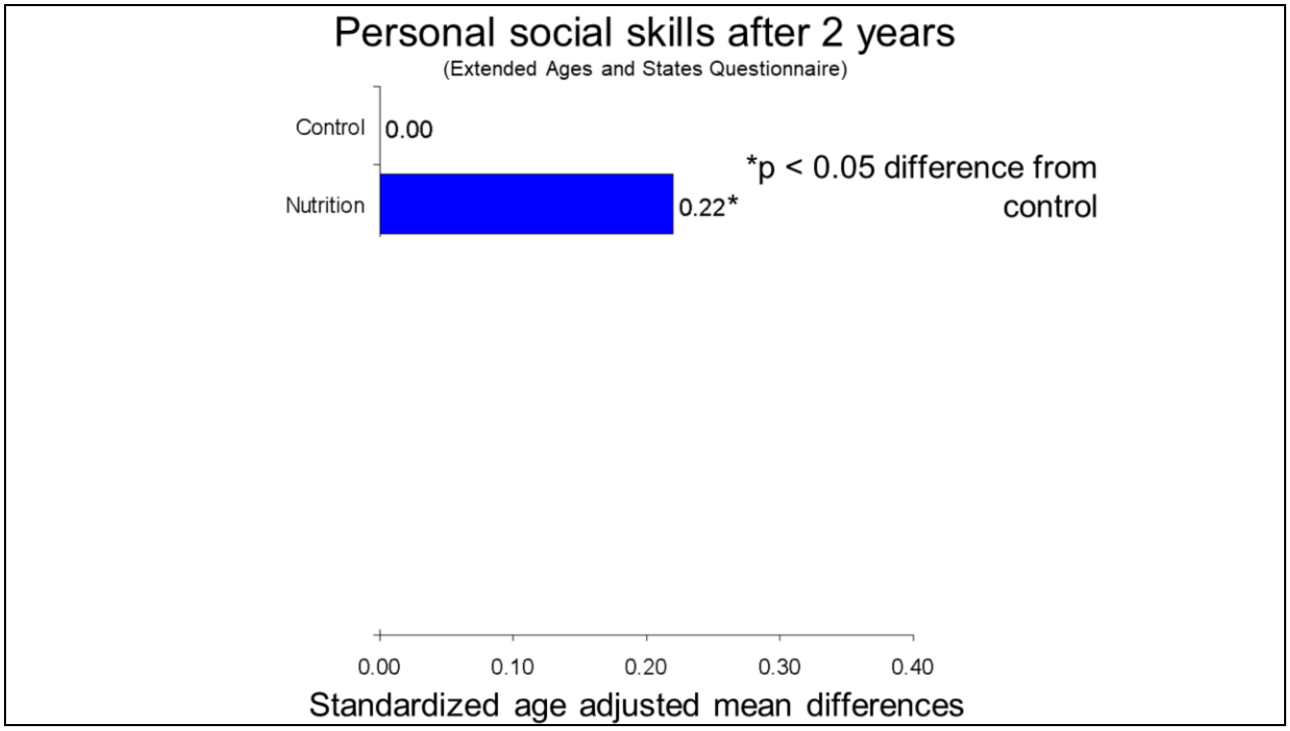
Does your child copy the activities you do, such as wipe up a spill, sweep, shave, or comb hair?

And finally, we looked at a child development measure called the extended ages and stages questionnaire where the field workers read each item to a parent, and then recorded responses as yes, sometimes, or not yet. Some required observation. And then we adjusted these scores for a number of factors that you see on the slide.

For example, a gross motor test question would be, does your child jump with both feet leaving the floor at the same time? Or does your child copy the activities you do, such as wipe up a spill, sweep, shave, or comb hair?



Let's look at the gross motor skills. In the control group, it [INAUDIBLE] at zero. In the nutrition group, we saw an increase in the score, 0.19. It was statistically significant. There were also increases seen in the WASH combination group, in the nutrition plus WASH group.



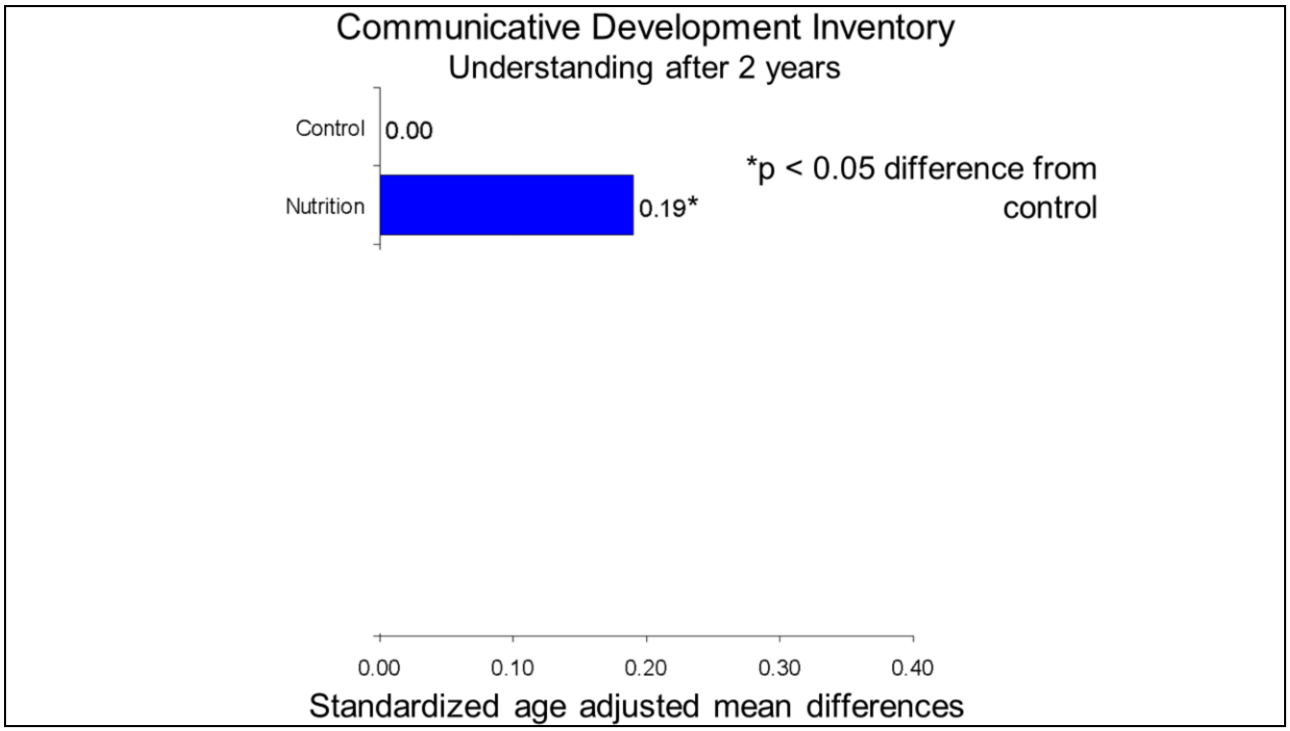
Additionally, we looked at personal and social skills after two years from the extended ages and stages questionnaire. And again in the nutrition group, there was a statistically significant improvement. And in several of the other groups. In fact, all of the other groups as well. On several measures, there seemed to be some improvements in cognition after the interventions.

MacArthur Bates Communicative Development Inventories
Bangladesh adapted short form

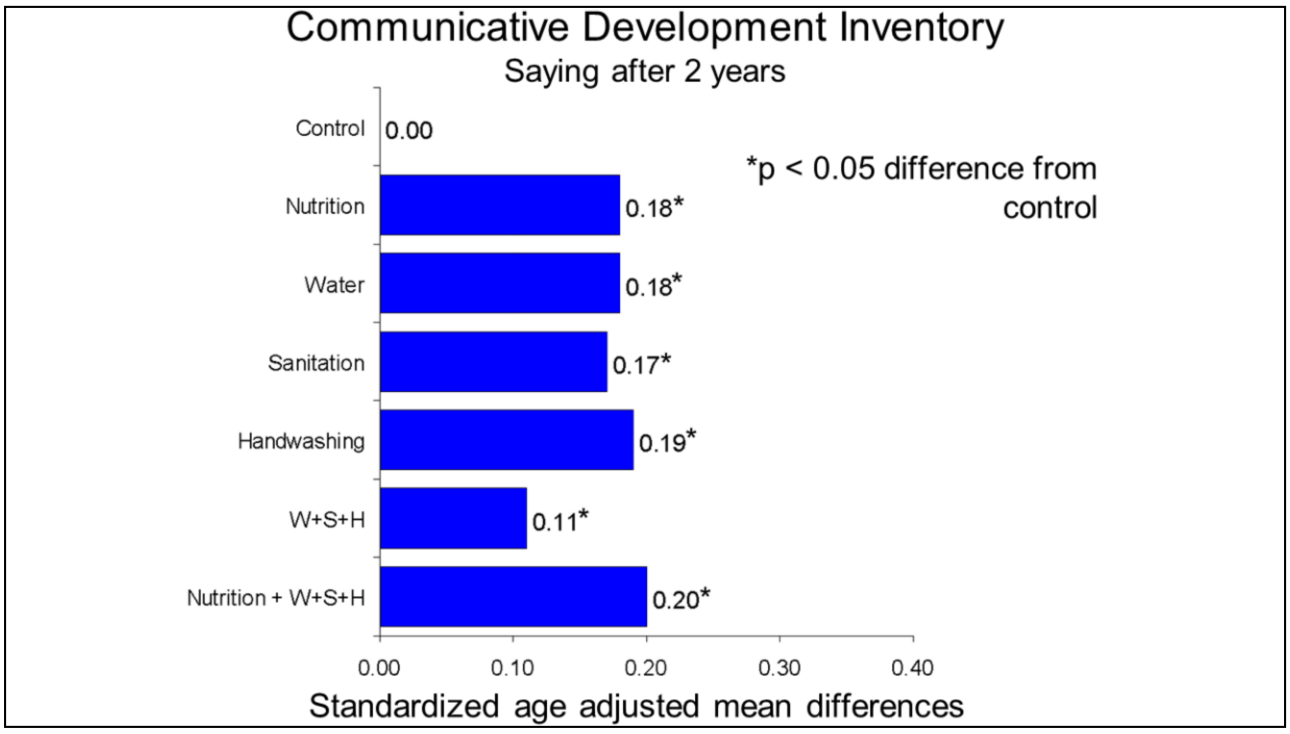
- Structured parental interview
- List of words
  - Does the child:
    - Understand?
    - Understand and say?
  - # of words summed
- Valid, reliable, normed, translated
- Adjusted difference for:
  - Child sex, child age, mother age, parents education, number of household members, number of household rooms, household roof, floor, wall materials, availability of electricity, type of fuel for cooking, household asset

	UNDERSTANDS	UNDERSTANDS AND SAYS
choo choo	○	○
meow	○	○
ouch	○	○
uh oh	○	○
bird	○	○
dog	○	○
duck	○	○
kitty	○	○

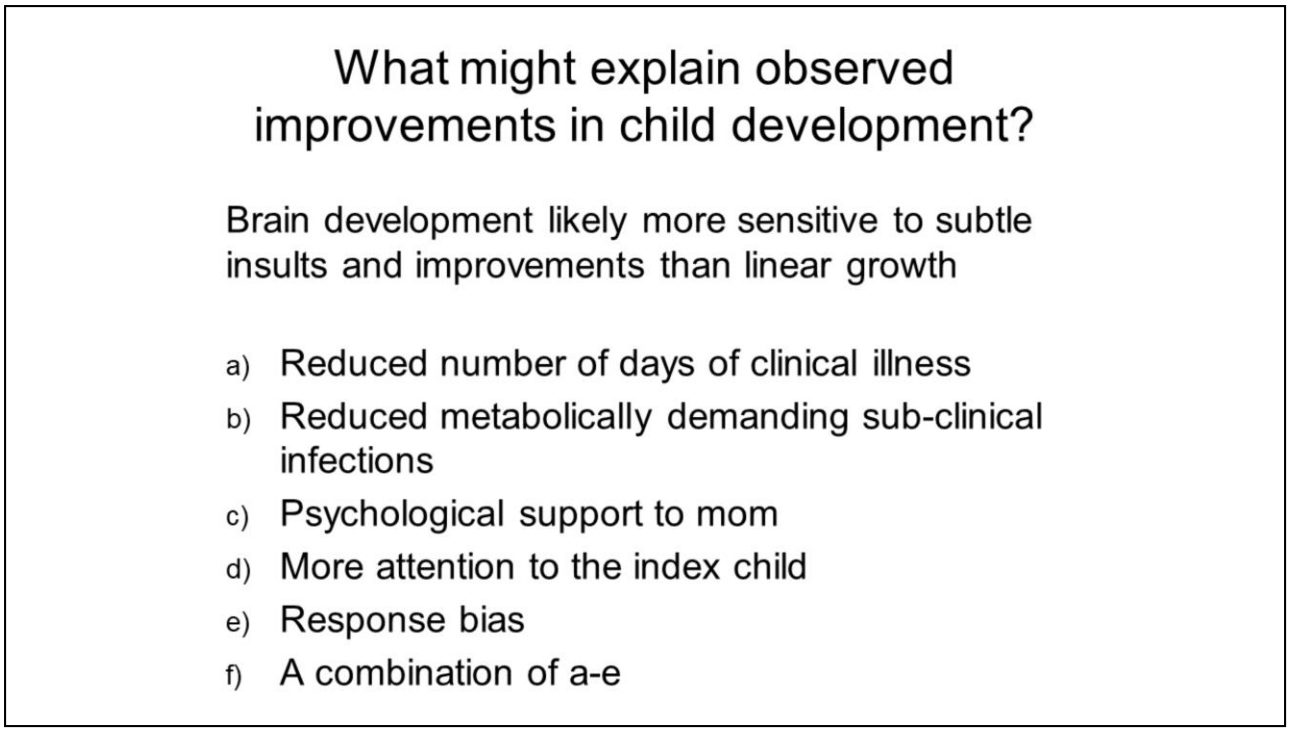
Another test that was done was the MacArthur Bates Communicative Development Inventory, or CDI. This was a structured parental interview with a list of words that the child either understood or doesn't understand and can say. And then we sum up the number of words and adjust the differences for all these different factors between the groups.



Here are the results of this Communicative Development Inventory with a statistically significant finding in the nutrition arm, as well as all the other arms.

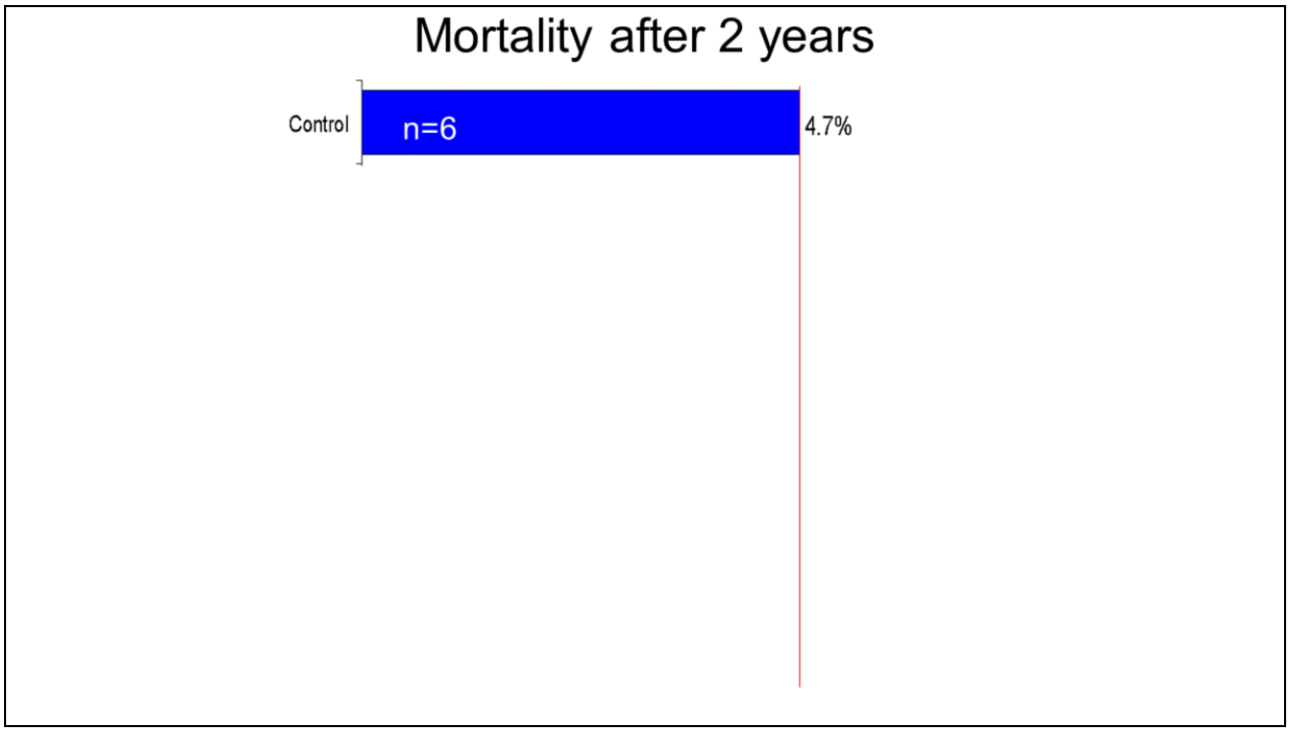


And for saying words after age two years, also significant improvements in the intervention arms.



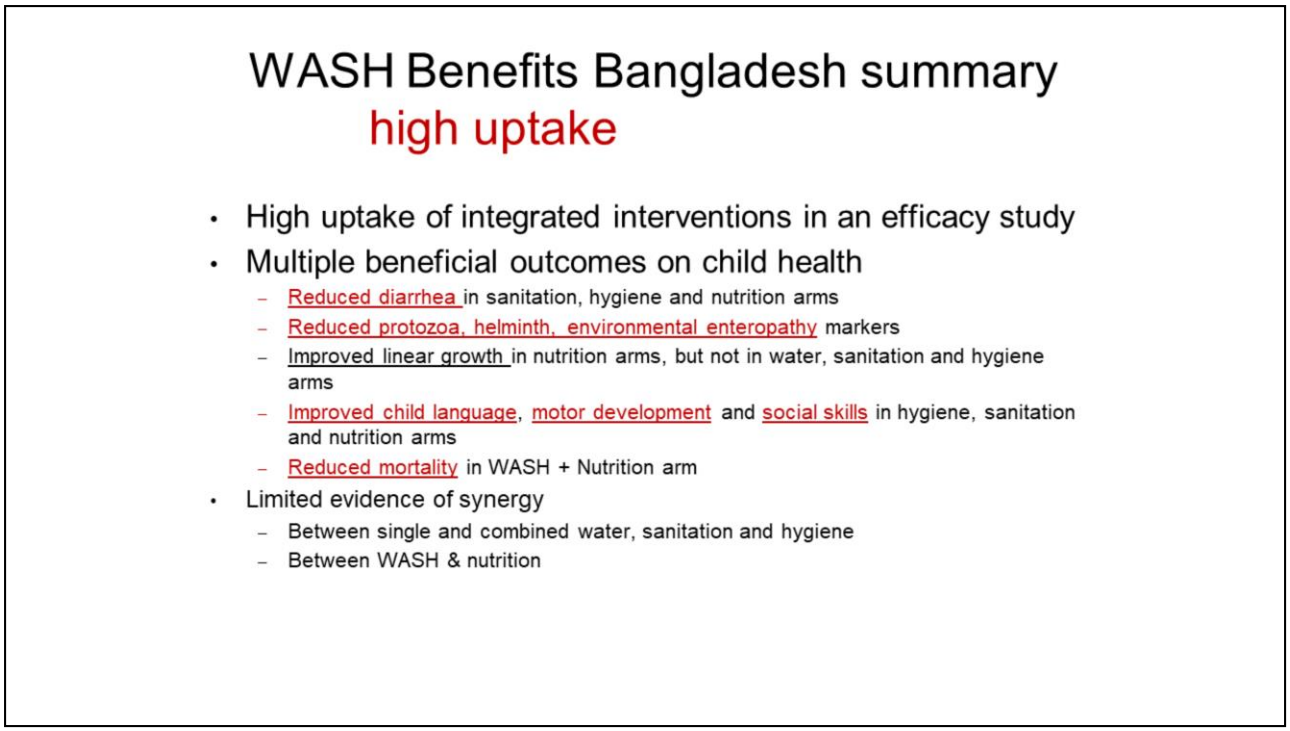
We asked ourselves what might explain the observed improvements in child development. Brain development is likely more sensitive to subtle insults and improvements than linear growth. There was a reduced number of days of clinical illness in these children who were in the intervention groups, which could lead to a reduction metabolically in demanding-- subclinical infections place metabolic demand. So those demands were reduced in the treated groups.

There was psychological support to the mom from the frequent visits of the promoters. There was more attention to the index child. There might be response bias and a bias towards giving the right answer, or some combination of all of these.



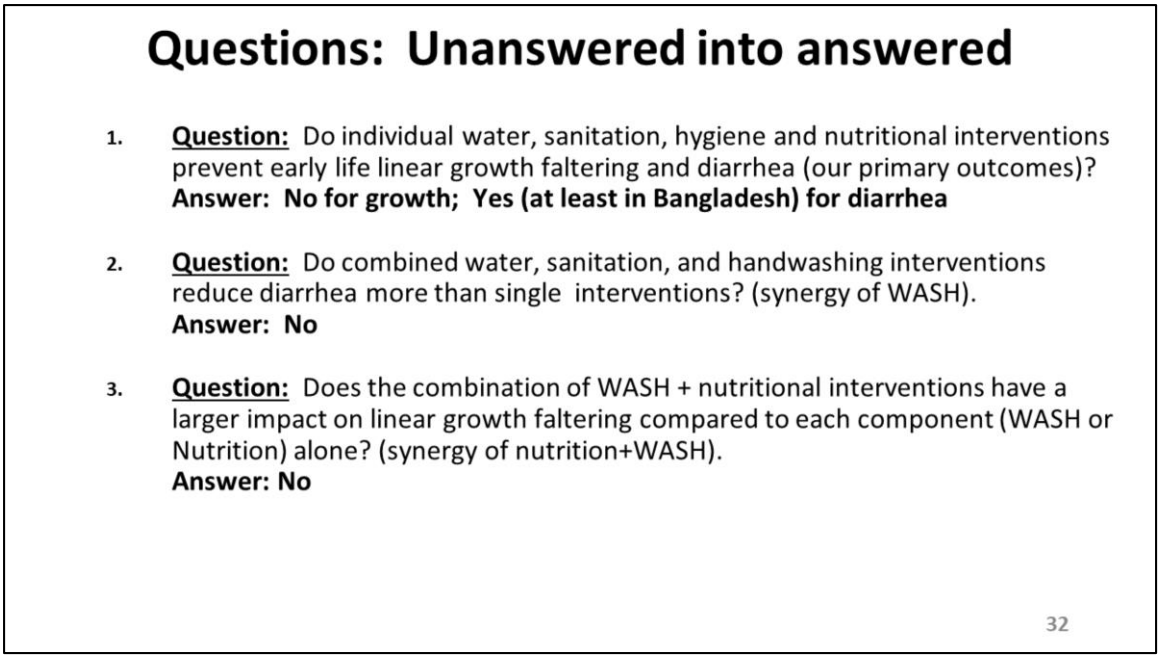
Finally, we looked at mortality after two years. So there were six deaths in the control group and a slight reduction in the other group. In the nutrition group, there was a risk ratio of 0.81. And these were very small numbers, of course, so it's hard to make much of how to interpret these. It looks as if there might be an important reduction going on.

And there was a statistically significant reduction from six to one in the nutrition plus WASH plus sanitation plus hygiene sort of super arm. And again, these numbers are so small, it's hard to know whether these would be borne out in a much larger study, but they're intriguing.



Overall, what did we learn from the study? So there was high uptake in the integrated interventions in this study. There were multiple beneficial outcomes on child health, such as reduced diarrhea in the sanitation, hygiene, and nutrition arms; reduced protozoa, helminth, environmental, and neuropathy markers; improved linear growth and nutrition arms, but not in the water, sanitation, and hygiene arm. There was improved child language, motor development, social skills in the hygiene, sanitation, nutrition arms, and reduced mortality in the WASH plus nutrition arm.

There's limited evidence of synergy here. That is, when we look at single interventions compared to the combined water, sanitation, and hygiene intervention, we did not see improvement. And we didn't see improvement in the WASH plus nutrition arm compared to the nutrition arm alone. All of these lead us to ask the question, did high uptake lead to high impact here?



Let's summarize our original questions. First, do individual water, sanitation, hygiene, and nutritional interventions prevent early life linear growth faltering and diarrhea, which is our primary outcome? The answer here was not for growth, but at least in Bangladesh, they did seem to approve the outcome diarrhea.

Do combined water, sanitation, and handwashing interventions reduce diarrhea more than single interventions? That is, is there a synergy from the combination of WASH and individual arms? No, we did not find evidence of this.

And finally, does the combination of WASH plus nutritional interventions have a larger impact on linear growth faltering compared to each component, which is WASH or nutrition alone? That is, is there a synergy between nutrition and WASH? And the answer here was also no.

A few conclusions

- WASH interventions do not provide a simple approach to correcting linear growth deficits.
- In some circumstances, trials are essential for unconfounded inference
- Child development
  - Not simply proxied by linear growth
  - Arguably more important than linear growth
  - A variety of drivers
- Team Bangladesh, my broader WASH Benefits colleagues and the Gates Foundation are awesome

Some conclusions. The WASH interventions don't provide a simple approach to correcting linear growth deficits. In some circumstances, trials are essential for unconfounded inference. The child development was not simply proxied by linear growth. It's arguably more important than linear growth, and there seemed to be a variety of drivers. And I just want to thank my team partners in Bangladesh and my broader WASH benefits colleagues and the Gates Foundation for the amazing support work in this project.

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

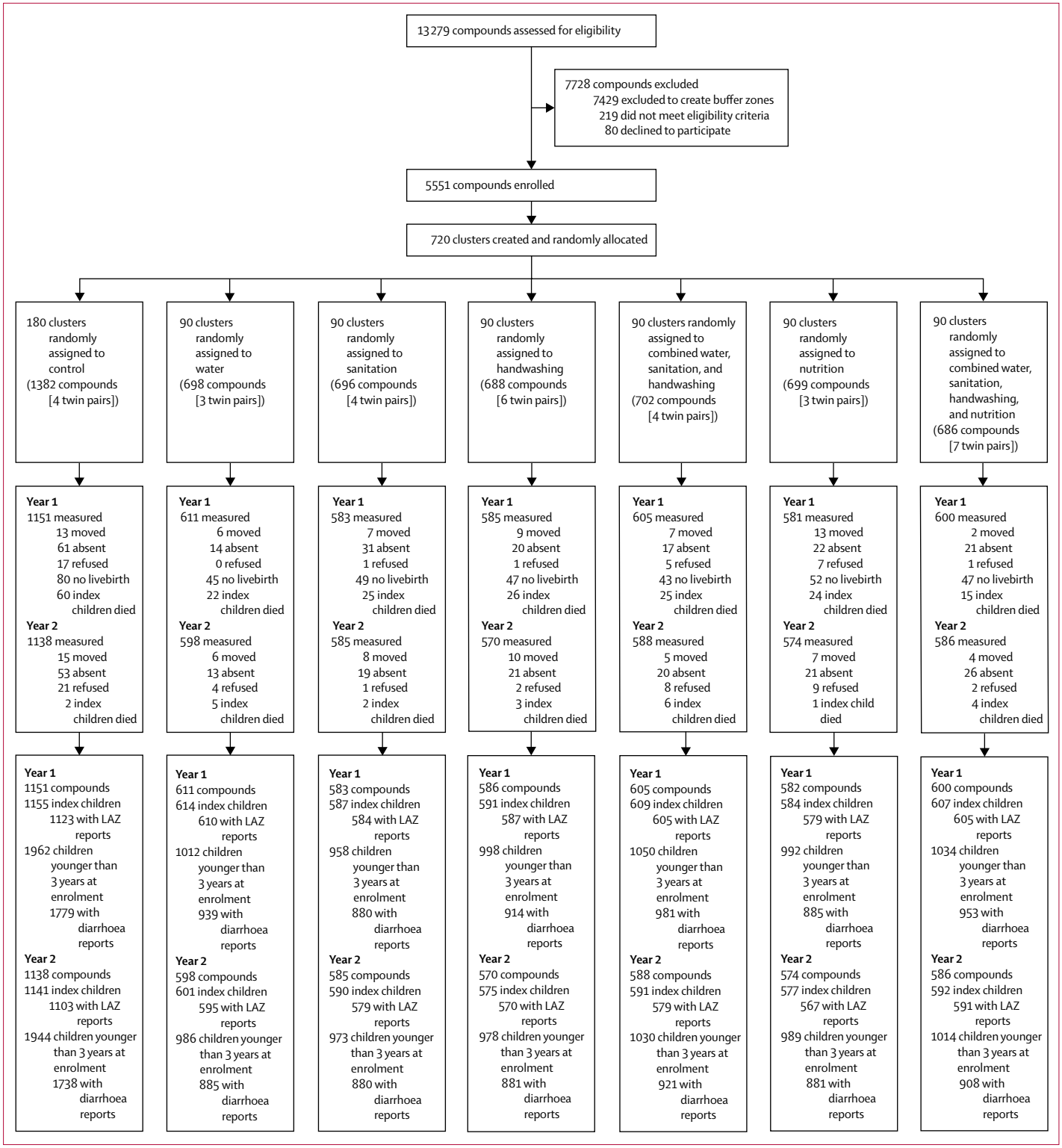


Figure 1: **Trial profile and analysis populations for primary outcomes**
LAZ=length-for-age *Z* scores.

	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.

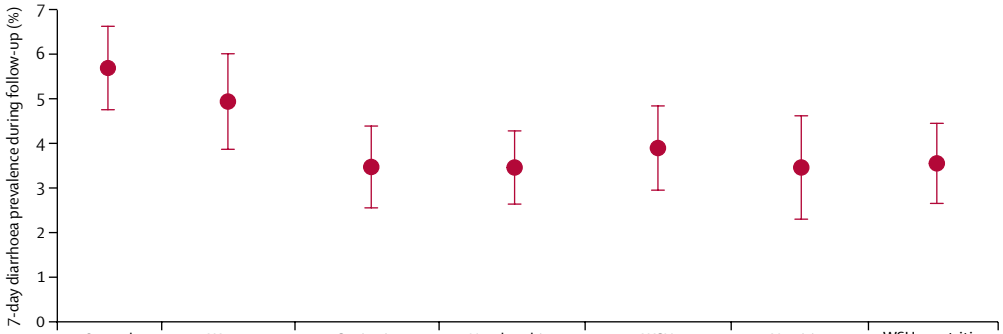
The trial is registered at ClinicalTrials.gov, number NCT01590095. The International Centre for Diarrhoeal Disease Research, Bangladesh convened a data and safety monitoring board and oversaw the study.

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

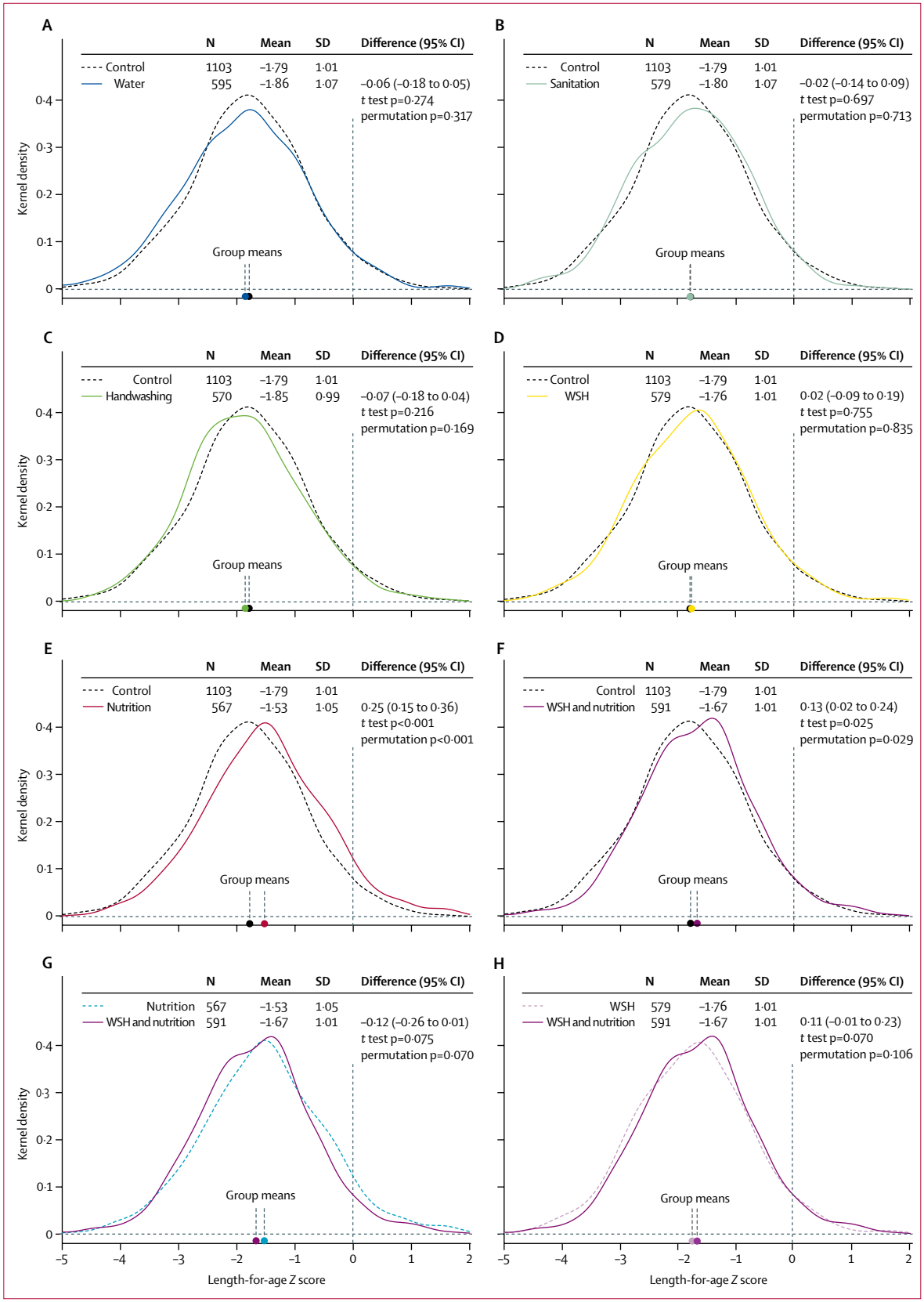


Figure 3: **Intervention effects on length-for-age Z scores in 4584 children after 2 years of intervention**

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 proportion of the reduced diarrhoea was because of

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

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

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

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

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

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

Contributors

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

Declaration of interests

We declare no competing interests.

Acknowledgments

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Clair Null, Christine P Stewart, Amy J Pickering, Holly N Dentz, Benjamin F Arnold, Charles D Arnold, Jade Benjamin-Chung, Thomas Clasen, Kathryn G Dewey, Lia C H Fernald, Alan E Hubbard, Patricia Kariger, Audrie Lin, Stephen P Luby, Andrew Mertens, Sammy M Njenga, Geoffrey Nyambane, Pavani K Ram, John M Colford Jr

Summary

Background Poor nutrition and exposure to faecal contamination are associated with diarrhoea and growth faltering, both of which have long-term consequences for child health. We aimed to assess whether water, sanitation, handwashing, and nutrition interventions reduced diarrhoea or growth faltering.

Methods The WASH Benefits cluster-randomised trial enrolled pregnant women from villages in rural Kenya and evaluated outcomes at 1 year and 2 years of follow-up. Geographically-adjacent clusters were block-randomised to active control (household visits to measure mid-upper-arm circumference), passive control (data collection only), or compound-level interventions including household visits to promote target behaviours: drinking chlorinated water (water); safe sanitation consisting of disposing faeces in an improved latrine (sanitation); handwashing with soap (handwashing); combined water, sanitation, and handwashing; counselling on appropriate maternal, infant, and young child feeding plus small-quantity lipid-based nutrient supplements from 6–24 months (nutrition); and combined water, sanitation, handwashing, and nutrition. Primary outcomes were caregiver-reported diarrhoea in the past 7 days and length-for-age *Z* score at year 2 in index children born to the enrolled pregnant women. Masking was not possible for data collection, but analyses were masked. Analysis was by intention to treat. This trial is registered with ClinicalTrials.gov, number NCT01704105.

Findings Between Nov 27, 2012, and May 21, 2014, 8246 women in 702 clusters were enrolled and randomly assigned an intervention or control group. 1919 women were assigned to the active control group; 938 to passive control; 904 to water; 892 to sanitation; 917 to handwashing; 912 to combined water, sanitation, and handwashing; 843 to nutrition; and 921 to combined water, sanitation, handwashing, and nutrition. Data on diarrhoea at year 1 or year 2 were available for 6494 children and data on length-for-age *Z* score in year 2 were available for 6583 children (86% of living children were measured at year 2). Adherence indicators for sanitation, handwashing, and nutrition were more than 70% at year 1, handwashing fell to less than 25% at year 2, and for water was less than 45% at year 1 and less than 25% at year 2; combined groups were comparable to single groups. None of the interventions reduced diarrhoea prevalence compared with the active control. Compared with active control (length-for-age *Z* score –1·54) children in nutrition and combined water, sanitation, handwashing, and nutrition were taller by year 2 (mean difference 0·13 [95% CI 0·01–0·25] in the nutrition group; 0·16 [0·05–0·27] in the combined water, sanitation, handwashing, and nutrition group). The individual water, sanitation, and handwashing groups, and combined water, sanitation, and handwashing group had no effect on linear growth.

Interpretation Behaviour change messaging combined with technologically simple interventions such as water treatment, household sanitation upgrades from unimproved to improved latrines, and handwashing stations did not reduce childhood diarrhoea or improve growth, even when adherence was at least as high as has been achieved by other programmes. Counselling and supplementation in the nutrition group and combined water, sanitation, handwashing, and nutrition interventions led to small growth benefits, but there was no advantage to integrating water, sanitation, and handwashing with nutrition. The interventions might have been more efficacious with higher adherence or in an environment with lower baseline sanitation coverage, especially in this context of high diarrhoea prevalence.

Funding Bill & Melinda Gates Foundation, United States Agency for International Development.

Lancet Glob Health 2018; 6: e316–29

Published Online January 29, 2018 http://dx.doi.org/10.1016/S2214-109X(18)30005-6

See Comment page e236

See Articles page e302

Innovations for Poverty Action, Kakamega, Kenya (C Null PhD, H N Dentz MPH, G Nyambane MA); **Center for International Policy Research and Evaluation, Mathematica Policy Research, Washington, DC, USA** (C Null); **Rollins School of Public Health, Emory University, Atlanta, GA, USA** (C Null, Prof T Clasen PhD); **Department of Nutrition, University of California, Davis, CA, USA** (C P Stewart PhD, H N Dentz, C D Arnold MS, Prof K Dewey PhD); **Department of Civil and Environmental Engineering** (A J Pickering PhD), **and Department of Infectious Diseases and Geographic Medicine** (S P Luby MD), **Stanford University, Stanford, CA, USA; Department of Civil and Environmental Engineering, Tufts University, Medford, MA, USA** (A J Pickering); **Division of Epidemiology** (B F Arnold PhD, J Benjamin-Chung PhD, A Lin PhD, A Mertens MS, Prof J M Colford Jr MD), **Division of Community Health Sciences** (Prof L C H Fernald PhD, P Kariger PhD), **and Division of Biostatistics** (A E Hubbard PhD) **School of Public Health, University of California, Berkeley, CA, USA; Eastern and Southern Africa Centre of International Parasite Control, Kenya Medical Research Institute, Nairobi, Kenya** (S M Njenga PhD); **Department of Epidemiology and Environmental Health, School of Public Health and Health**

Introduction

An estimated 156 million children worldwide suffer from stunting (linear growth faltering) and are unlikely to reach their full potential as adults.[1] Linear growth faltering is the most apparent sign of chronic undernutrition and is the physical manifestation of combined physiological and developmental insults. Early-life stunting leads to poor cognitive development in childhood, reduced

Professions, University at Buffalo, Buffalo, NY, USA
(P K Ram MD)

Correspondence to:
Dr Clair Null, Center for International Policy Research and Evaluation, Mathematica Policy Research, Washington, DC 20002, USA
cnull@mathematica-mpr.com

Research in context

Evidence before this study

Malnutrition and enteric infection are thought to act together to impair child health and survival, yet there is limited evidence of low cost, scalable interventions effective at breaking this cycle. A 2008 meta-analysis by Dewey and Adu-Afarwuah found that interventions offering nutrient supplementation or counselling on complementary feeding could result in modest improvements to child growth. Another meta-analysis by Waddington and Snilsveit in 2009 showed that water treatment or handwashing could prevent diarrhoea, but there had not been any randomised trials of the effect of sanitation on diarrhoea. During this study, five other randomised trials of the effects of sanitation on diarrhoea and growth were published, but three were limited by low adherence. Whether combining water, sanitation, handwashing, or nutrition interventions could result in added benefits for health and growth was not known.

Added value of this study

This trial is one of the first to provide experimental evidence on whether individual and combined water, sanitation, or handwashing interventions improve growth; combined water, sanitation, and handwashing interventions are more effective at reducing diarrhoea and growth faltering than any intervention alone; and nutrition counselling and supplementation are more effective when combined with improved water, sanitation, and handwashing. This is the first rigorous evaluation of upgrading from unimproved to improved latrines in sub-Saharan Africa. None of the interventions reduced diarrhoea, and only the interventions that included nutrition counselling and nutrient supplementation improved growth.

Implications of all the available evidence

Our results on growth effects are consistent with those from previous research on the combination of nutrition counselling and nutrient supplementation, finding modest effects on linear growth. It is possible that more intensive promotion and higher adherence would have resulted in larger effects, especially in this context of high diarrhoea prevalence, but few programmes are likely to be able to afford sustaining a more ambitious behaviour change programme than was included in this trial. In a context where most households already had an unimproved sanitation facility, provision of technologically simple interventions including chlorination for household treatment of drinking water, improved pit latrines, and handwashing stations—standard for most WASH programmes in rural areas of low-income countries—might not be sufficient to improve growth. By contrast with previous studies, this trial provided evidence that technologically simple water, sanitation, and handwashing interventions with adherence rates at least as high as most programmes achieve might not reduce childhood diarrhoea in all situations.

economic productivity in adulthood, and increased risk of morbidity and mortality.[2,3] Because nutrient supplementation and counselling interventions for maternal, infant, and young child feeding have been only marginally successful at preventing growth faltering, exposure to faecal contamination in the environment has recently been hypothesised to lead to environmental enteric dysfunction, which features chronic immune stimulation and impaired nutrient absorption, thereby constraining a growth response to improved nutrition.[4] In addition to the detrimental effects on growth and development, undernutrition was estimated to cause 45% of all child deaths in 2011, and it has long been recognised that undernutrition is an important determinant of susceptibility to infectious disease.[5,6] Diarrhoea is the second leading cause of death in children aged 1–59 months, contributing to almost 500 000 deaths in children younger than 5 years in 2015.[7] Frequent diarrhoea is also associated with linear growth faltering.[8] If there is a pathway independent of symptomatic diarrhoea linking environmental contamination to growth faltering, the benefits of improving water safety, sanitation, and handwashing could be underestimated because studies have generally focused on diarrhoea. It is unclear whether combined water, sanitation, handwashing, and nutritional interventions reduce diarrhoea or improve growth more than single interventions.

See Online for appendix

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[9] in Bangladesh evaluated the same objectives.

Methods

Study design

The Kenya WASH Benefits study was a cluster-randomised trial done in rural villages in Bungoma, Kakamega, and Vihiga counties in Kenya's western region (appendix p 11). We used a cluster design to facilitate the logistics of the behaviour change component of the interventions and minimise contamination between intervention and comparison households. We hypothesised that the interventions would improve the health of the index child in each household. We optimised the trial design to measure group-level differences in primary outcomes by including a large number of clusters, each comprising relatively few children (12 on average) with infrequent measurement. Each measurement round lasted roughly 1 year and was balanced across treatment

groups and geography to minimise seasonal or geographic confounding when comparing outcomes across groups.

With active and passive control groups and six intervention groups (water; sanitation; handwashing; combined water, sanitation, and handwashing; nutrition; and combined water, sanitation, handwashing, and nutrition), the design enabled 11 comparisons of each intervention group with the active control; combined water, sanitation, and handwashing with each intervention alone; and combined water, sanitation, handwashing, and nutrition with nutrition alone, and combined water, sanitation, and handwashing. A double-sized active control group was used to increase power because there were six separate intervention comparisons against control.[10] Households in the active control and all intervention groups were visited by community-based health promoters monthly to measure the child's mid-upper arm circumference. Health promoters did not visit households in passive control clusters. Measurement of outcomes, as well as water, sanitation, handwashing, and nutrition characteristics were measured in the passive control group at the same times as in other groups. The study design and rationale have been published previously.[10]

The study protocol was approved by the Committee for the Protection of Human Subjects at the University of California, Berkeley (protocol number 2011-09-3654), the institutional review board at Stanford University (IRB-23310), and the scientific and ethics review unit at the Kenya Medical Research Institute (protocol number SSC-2271). Under direction of the study investigators, Innovations for Poverty Action (IPA) was responsible for intervention delivery and data collection.

Participants

Villages were eligible for selection into the study if they were rural, most of the population relied on communal water sources and had unimproved sanitation facilities, and there were no other ongoing water, sanitation, handwashing, or nutrition programmes. Participants were identified through a complete census of eligible villages. Within selected villages, women were eligible to participate if they reported that they were in their second or third trimester of pregnancy, planned to continue to live at their current residence for the next 2 years, and could speak Kiswahili, Luhya, or English well enough to respond to an interviewer administered survey. IPA staff formed clusters from one to three neighbouring villages to have six or more pregnant women per cluster after the enrolment survey. Outcomes were assessed in the children born from these pregnancies (index children), including twins. Although the study area is one of the areas with the highest HIV prevalence in Kenya, according to the 2012 Kenya AIDS Indicator Survey, the prevalence in women aged 15–64 years in the study area was below 6% (that survey did not include testing of children). Because there would not have been sufficient sample size to allow for subgroup analysis by HIV status, no attempt was made to identify participants who were HIV positive. Participants gave written informed consent before enrolment.

Randomisation and masking

Clusters were randomly allocated to treatment using a random number generator with reproducible seed at the University of California, Berkeley. Groups of nine geographically-adjacent clusters were block-randomised into a double-sized active control; passive control; water; sanitation; handwashing; water, sanitation, and handwashing; nutrition; or water, sanitation, handwashing, and nutrition. Allocation by cluster identification number was communicated directly to the field team; investigators remained blinded to treatment assignments. Blinding of participants was not possible. Participants were informed of their treatment assignment after baseline data collection and might have known the treatment assignment of nearby villages. The health promoters and staff who delivered the interventions were not involved in data collection, but the data collection team could have inferred treatment status if they saw intervention materials in study communities.

Procedures

The interventions were designed to maximise adherence to behaviours that could protect children from exposure to pathogens in their environment and improve diet quality. Formative research in the study area concluded that the health benefits of target behaviours were already well understood, but this knowledge was not sufficient to lead to action. As such, the behaviour change strategy and intervention materials were selected to create enabling environments, build supportive social norms, and target emotional drivers of decision making. The messages and delivery modes for the behaviour change strategy drew from existing information, education, and communication materials from organisations such as WHO, the Kenyan Government, UNICEF, and the Alive and Thrive network, and extensive previous qualitative work on the drivers of handwashing behaviours. Monthly visit modules were developed and pilot-tested to provide behavioural recommendations to mothers and other caregivers using key thematic constructs of convenience, nurturing care, and aspiration. We did a pilot randomised controlled trial[11] to test the feasibility and acceptability of all the interventions and to collect data that allowed us to optimise the ratio of community-based promoters to study participants. To identify and correct systematic problems with adherence, staff confirmed that intervention materials were delivered to all study participants at the outset of the trial, and collected monitoring data on availability of intervention materials and recommended behaviours during unannounced visits to a random sample of at least 20% of participants in intervention groups 2, 6, 10, and 19 months after the interventions began.

Community-based promoters for intervention and active control groups were nominated by study mothers and other mothers of children younger than 3 years in the community. A second promoter was added if there were more than ten participants (single groups) or more than eight participants (combined groups) in the cluster, giving a total of 1031 promoters. Promoters attended 2 days (active control), 6 days (single groups), or 7 days (combined groups) of initial training led by study staff on how to measure mid-upper-arm circumference, communication skills, intervention-specific behaviour change messages and intervention materials, and the information they were expected to report to IPA. Refresher trainings were done 6, 12, and 18 months after the initial training. At 2, 4, 9, 15, and 21 months, study staff met with promoters in their clusters to observe visits and offer supportive supervision. Study staff called promoters monthly to collect information on their activities, intervention adherence in the households they visited, referrals to health centres, and births or deaths of study children. Promoters received a branded T-shirt, a mobile phone, job aids and intervention materials, and compensation of approximately US$15 per month for the first 6 months when they had more intensive engagement with the study participants, and $9 per month thereafter (the prevailing daily wage for unskilled labour in the study area is $1–2). Promoters were instructed to visit all participants in their cluster monthly and measure the child's arm circumference or the pregnant mother's abdomen.

In intervention groups, promoters engaged study participants and other compound members through interactive activities such as guided discussions using visual aids, song, and storytelling; resupplied consumable intervention materials; encouraged consistent practice of targeted behaviours; and helped troubleshoot barriers to adherence, including problems with intervention hardware and behavioural barriers. Promoters were provided with detailed plans for every visit, including key messages, scripts for discussing visual aids, and instructions for activities that emphasised the learning objectives. Visits lasted about 10 min in the active control group and 45–60 min in intervention groups during the first year when the key messages were conveyed. In the second year, promoters reinforced messages to maintain habits. All groups used messages on themes of nurture, aspiration, and self-efficacy, particularly in the context of a new birth. Interventions used convenience and social norms to encourage target behaviours.

For **intervention-specific training materials** see https://osf.io/fs23x

In the three intervention groups that included water, promoters advocated treatment of drinking water with sodium hypochlorite. Chlorine dispensers for convenient water treatment at the point of collection were installed at an average of five communal water sources in the cluster and refilled as needed. Every 6 months, households in study compounds were given a 1 L bottle of chlorine for point-of-use water treatment in case households collected rainwater or used a source without a dispenser. Promoters used chlorine test strips during their regular visits to determine if the household was using chlorine, and negative results stimulated conversation about addressing barriers to chlorination.

In the three intervention groups that included sanitation, promoters advocated using latrines for defecation and safe disposal of children's and animals' faeces into a latrine. Existing unimproved latrines in study households were upgraded to improved latrines by installing a plastic slab, which also had a tight-fitting lid over the hole. New latrines were constructed for study households that did not have a latrine or whose latrine was unlikely to last for 2 years. All households in study compounds received a sani-scoop with a paddle as a dedicated faeces-removal tool. Finally, all households with children younger than 3 years in study compounds received plastic potties to facilitate toilet training and transfer of child faeces to the latrine.

In the three intervention groups that included handwashing, promoters advocated handwashing with soap before handling food and after defecation (including assisting a child). Study compounds were given two permanent, water-frugal handwashing stations intended to be installed near the food preparation area and the latrine. Handwashing stations were constructed of painted metal, with two foot-pedal-operated jerry-cans that dispensed a light flow of rinse water and soapy water. Promoters added chunks of bar soap to the soapy water container quarterly.

In the two intervention groups that included nutrition, a set of ten age-targeted modules were developed to enable promoters to advocate for best practices in maternal, infant, and young child feeding: recommendations for dietary diversity during pregnancy and lactation, early initiation of breastfeeding, exclusive breastfeeding until 6 months, introduction of appropriate and diverse complementary foods at 6 months, and continued breastfeeding through 24 months. Facilitators and barriers to behaviour change were elicited using formative research and health promoter guides were developed to address common barriers and questions. Study mothers with children between 6–24 months were provided with two 10 g sachets per day of a small quantity of lipid-based nutrient supplement (LNS; Nutriset; Malauny, France) that could be mixed into the child's food. LNS provided 118 kcal per day and 12 essential vitamins and ten minerals. Promoters explained that LNS was not to replace breastfeeding or complementary foods.

Promoters and intervention materials were introduced at community meetings roughly 6 weeks after enrolment. All interventions were delivered within 3 months of enrolment (appendix p 1). LNS was introduced to each child when they turned 6 months old. All handwashing stations and latrines were inspected within a month of construction, and a subset of households was periodically visited to observe group-specific indicators of intervention adherence. These data alerted study investigators to any

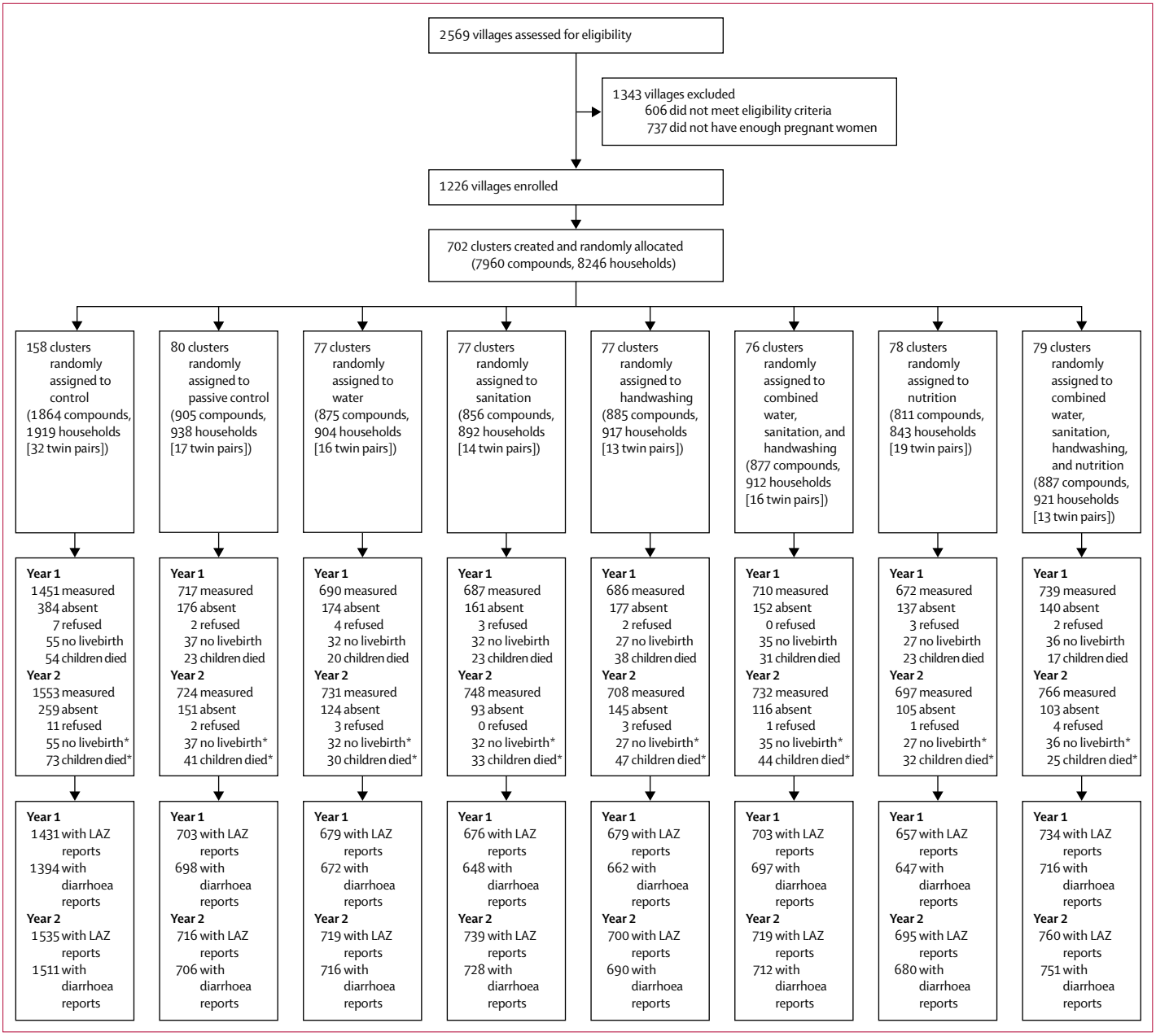


Figure 1: **Trial profile and analysis populations for primary outcomes**
LAZ=length-for-age *Z* scores. *Stillbirth and child death counts are cumulative.

issues with intervention implementation so they could be addressed consistently across all clusters and groups.

The enrolment survey included baseline demographics; assets; water, sanitation, and handwashing infrastructure; and target behaviours. Follow-up at 1 year and 2 years after intervention delivery consisted of an unannounced visit to study compounds to observe objective indicators of target behaviours (in all groups other than the passive control) and, on the following day, growth and health outcome measurements at a central location in the cluster (eg, a church or school).

Children identified as possibly malnourished (mid-upper-arm circumference <11·5 cm), either by the promoter during routine visits or by study staff during follow-up measurements, were referred to health facilities for treatment.

Outcomes

Adherence to the interventions was assessed using objective, observable indicators where possible (appendix pp 2, 3). 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. All child deaths reported by the health promoters were confirmed by a staff nurse who visited households. All outcomes were prespecified. Primary outcomes were caregiver-reported diarrhoea 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 and tertiary outcomes reported in this paper are 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; prevalence of stunting (length-for-age *Z* score less than –2), severe stunting (length-for-age *Z* score less than –3), wasting (weight-for-length *Z* score less than –2), and underweight (weight-for-age *Z* score less than –2); and all-cause mortality. We excluded children from *Z*-score analyses if their measurements were outside biologically plausible ranges following WHO recommendations. More details on exclusion criteria, measurement protocols, and outcome definitions are in the appendix (p 1).

Statistical analyses

Sample size calculations for the two primary outcomes were based on a minimum detectable effect of 0·15 in length-for-age *Z* score (intracluster correlation of 0·02 in our pilot study) and a relative risk of diarrhoea of 0·7 or smaller (assuming a 7-day prevalence of 12% in the active control group based on a pilot study to inform this trial) for a comparison of any intervention with the double-sized control group, assuming 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.[10,11] Sample size calculations indicated 80 clusters per group, each with ten children.

For more on the **updates to the analysis plan** see https://osf.io/7urqa/

Two biostatisticians, blinded to treatment assignment, independently replicated the analyses following the prespecified analysis plan with minor updates.[10] We analysed participants according to their randomised assignment (intention to treat), regardless of adherence to the intervention, using the active control group as the comparator. We used paired *t* tests for unadjusted length-for-age *Z* score comparisons and the Mantel-Haenszel prevalence ratio and difference for unadjusted diarrhoea and stunting comparisons, with randomisation block defining matched pairs or stratification. In secondary analyses, we estimated prevalence ratios and differences, adjusting for baseline covariates using targeted maximum likelihood estimation.[12] 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, which tested whether primary outcomes were the same in control households with more versus fewer households receiving interventions within a 2 km radius. In an analysis that was not prespecified, we tested for intervention effects on diarrhoea using only year 1 data.

The trial is registered at ClinicalTrials.gov, number NCT01704105. IPA convened a data and safety monitoring board.

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 the data in the study and had final responsibility for the decision to submit for publication.

Results

2569 villages were assessed for eligibility, of which 606 were excluded on the basis of village-level characteristics (primarily not meeting the study's rural criteria). 1226 villages were grouped into 702 clusters that had six or more pregnant women (figure 1). Between Nov 27, 2012, and May 21, 2014, 8246 pregnant women were enrolled in the study. 281 women did not have a livebirth and 140 women delivered twins. After at least three attempts to measure each child, 6659 (86%) of 7780 surviving children were measured at year 2, with diarrhoea reports for 6494 children and length-for-age *Z* score measures for 6583 children. Children were aged 2–18 months (median 12 months) at 1-year follow-up (January, 2014, to June, 2015) and aged 16–31 months (median 25 months) at 2-year follow-up (February, 2015, to July, 2016), but 11 184 (87%) of 12 841 children were in the target age ranges of 9–15 months at year 1 and 21–27 months at year 2 (appendix p 12).

Household characteristics were similar across groups at enrolment (table 1). Roughly three-quarters of participants collected drinking water from an improved source, but had to walk at least 10 min on average to the source. Over 80% of households owned a latrine, but less than 20% had access to an improved latrine. Less than 15% of households had soap available at a handwashing location. The prevalence of moderate-to-severe household hunger was 12% or lower.

Around 75% of households were visited by their promoter within the past month at year 1, but frequency of contact fell by year 2, with 40% or fewer households reporting a visit in the past month in each group (monitoring data suggest that most households were still visited at least every other month during the second year of the trial; see details in the appendix p 2, and table 2). Slightly less than half of households had detectable free chlorine in stored drinking water in the water group. Around 40% of drinking water samples tested in the water, sanitation, handwashing, and nutrition group had detectable free chlorine at year 1, which fell to around 20% by year 2. A high proportion of households (75%) had improved latrine access, which remained stable in year 1 and year 2 in households in the sanitation groups, increasing by more than 50% compared with the active control group. Reported safe disposal of children's faeces into a latrine fell by roughly half in all

	Active control (N=1919)	Passive control (N=938)	Water (N=904)	Sanitation (N=892)	Handwashing (N=917)	Water, sanitation, and handwashing (N=912)	Nutrition (N=843)	Water, sanitation, handwashing, and nutrition (N=921)
Maternal								
Age (years)	26 (6)	26 (7)	26 (6)	26 (7)	26 (6)	26 (6)	26 (6)	26 (6)
Completed at least primary education	916 (48%)	441 (47%)	447 (50%)	430 (48%)	402 (44%)	430 (47%)	409 (49%)	438 (48%)
Height (cm)	160 (6)	160 (7)	160 (6)	160 (6)	160 (6)	160 (6)	160 (7)	160 (7)
Study child is firstborn	490 (26%)	237 (25%)	205 (23%)	222 (25%)	208 (23%)	191 (21%)	206 (24%)	225 (25%)
Paternal								
Completed at least primary education	1098 (62%)	521 (60%)	532 (64%)	482 (58%)	500 (59%)	521 (61%)	491 (64%)	526 (62%)
Works in agriculture	749 (41%)	376 (43%)	378 (44%)	362 (43%)	363 (42%)	374 (43%)	343 (43%)	372 (43%)
Household								
Number of households per compound	2 (1)	2 (1)	2 (1)	2 (1)	2 (1)	2 (1)	2 (1)	2 (1)
Number of people per compound	8 (5)	8 (6)	8 (6)	8 (5)	8 (6)	8 (5)	8 (7)	8 (5)
Number of children <18 years in the household	3 (2)	3 (2)	3 (2)	3 (2)	3 (4)	3 (2)	3 (2)	3 (4)
Has electricity	122 (6%)	51 (5%)	60 (7%)	73 (8%)	67 (7%)	64 (7%)	58 (7%)	67 (7%)
Has a cement floor	107 (6%)	50 (5%)	71 (8%)	48 (5%)	41 (4%)	50 (5%)	48 (6%)	55 (6%)
Has an iron roof	1302 (68%)	600 (64%)	610 (68%)	587 (66%)	581 (63%)	574 (63%)	580 (69%)	615 (67%)
Owns a mobile phone	1526 (80%)	742 (79%)	705 (78%)	690 (77%)	722 (79%)	722 (79%)	685 (81%)	730 (79%)
Owns a motorcycle	185 (10%)	75 (8%)	81 (9%)	72 (8%)	91 (10%)	72 (8%)	81 (10%)	71 (8%)
Drinking water								
Primary drinking water source is improved*	1446 (76%)	699 (75%)	679 (75%)	675 (76%)	708 (78%)	624 (69%)	603 (72%)	697 (76%)
One-way walking time to primary water source (min)	11 (12)	12 (16)	12 (30)	10 (10)	11 (13)	11 (13)	11 (12)	11 (12)
Reported treating stored water	196 (13%)	92 (12%)	81 (11%)	94 (13%)	96 (13%)	97 (13%)	79 (12%)	106 (14%)
Sanitation								
Always or usually use primary toilet for defecation								
Men	1778 (95%)	867 (95%)	828 (94%)	810 (94%)	845 (95%)	851 (95%)	785 (95%)	854 (95%)
Women	1822 (96%)	898 (96%)	868 (96%)	840 (94%)	871 (96%)	877 (96%)	812 (96%)	872 (95%)
Daily defecating in the open								
Children aged 3 to <8 years	145 (12%)	87 (14%)	74 (13%)	68 (13%)	81 (14%)	75 (13%)	82 (15%)	75 (12%)
Children aged 0 to <3 years	789 (78%)	378 (77%)	376 (80%)	370 (75%)	358 (76%)	394 (77%)	363 (79%)	388 (78%)
Latrine								
Own any latrine	1561 (82%)	774 (83%)	750 (83%)	722 (81%)	756 (83%)	754 (83%)	701 (83%)	764 (83%)
Access to improved latrine	309 (17%)	153 (17%)	150 (18%)	131 (16%)	157 (19%)	153 (18%)	119 (15%)	143 (16%)
Human faeces observed in the compound	163 (9%)	79 (8%)	66 (7%)	72 (8%)	84 (9%)	73 (8%)	73 (9%)	87 (9%)
Handwashing location								
Has water within 2 m of handwashing location	487 (25%)	236 (25%)	242 (27%)	245 (28%)	245 (27%)	251 (28%)	228 (27%)	249 (27%)
Has soap within 2 m of handwashing location	164 (9%)	94 (10%)	91 (10%)	75 (8%)	83 (9%)	115 (13%)	90 (11%)	87 (9%)
Food security								
Prevalence of moderate-to-severe household hunger†	203 (11%)	113 (12%)	106 (12%)	91 (10%)	92 (10%)	101 (11%)	98 (12%)	104 (11%)

Data are n (%) or mean (SD). Percentages were calculated from smaller denominators than those shown at the top of the table for all variables because of missing values. *Defined by WHO UNICEF Joint Monitoring Program's definition for an improved water source. †Assessed by the Household Food Insecurity Access Scale.

Table 1: **Baseline characteristics by intervention group**

groups between year 1 and year 2, although the practice remained over twice as likely in the groups that included sanitation compared with other groups at year 1 and year 2. More than 75% of households in the intervention groups that included handwashing had water and soap available at a handwashing location at year 1, but this indicator also fell to about 20% by year 2. Adherence to LNS recommendations was high (≥95%) at year 1 and year 2, with children consuming a few more LNS sachets per month on average than would be expected at year 2. Across all indicators, adherence was comparable between the water, sanitation, and handwashing group and the water, sanitation, handwashing, and nutrition group compared with single intervention groups.

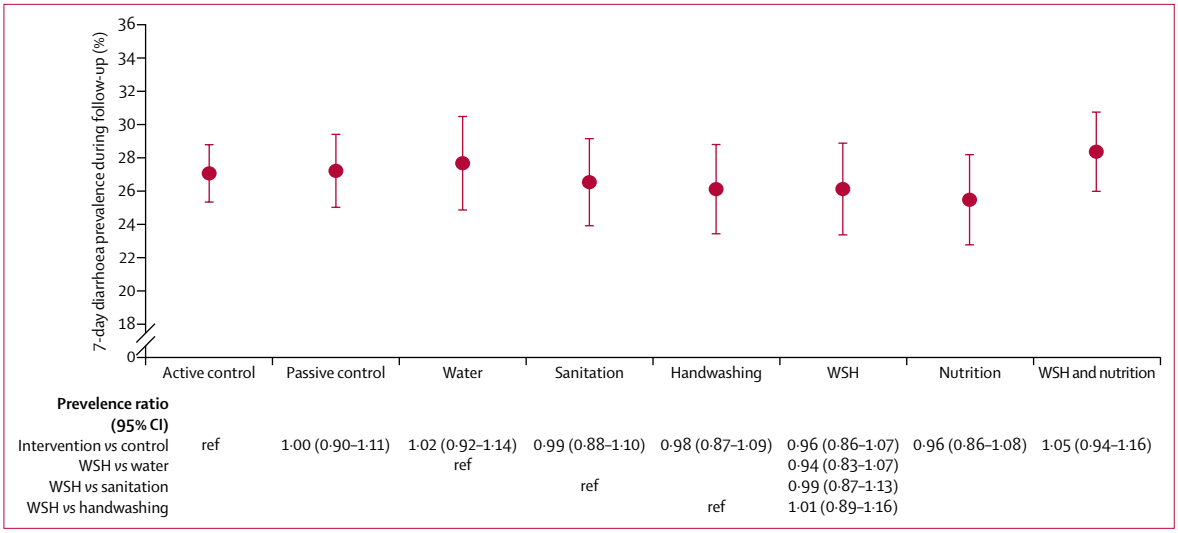
	Active Control (N=1919)	Passive Control (N=938)	Water (N=904)	Sanitation (N=892)	Handwashing (N=917)	Water, sanitation, and handwashing (N=912)	Nutrition (N=843)	Water, sanitation, handwashing, and nutrition (N=921)
Number of compounds assessed								
Enrolment	1913/1919 (100%)	936/938 (100%)	902/904 (100%)	890/892 (100%)	914/917 (100%)	912/912 (100%)	843/843 (100%)	918/921 (100%)
Year 1	1043/1919 (54%)	..	477/904 (53%)	473/892 (53%)	501/917 (55%)	536/912 (59%)	454/843(54%)	493/921 (54%)
Year 2	1458/1919 (76%)	..	696/904 (77%)	712/892 (80%)	690/917 (75%)	675/912 (74%)	650/843 (77%)	735/921 (100%)
Visited by promoter in past month								
Enrolment	..	..	..	..	..	..	..	..
Year 1	666/980 (68%)	..	338/445 (76%)	333/445 (75%)	333/480 (69%)	386/512 (75%)	344/433 (79%)	388/474 (82%)
Year 2	492/1412 (35%)	..	255/680 (37%)	278/692 (40%)	228/678 (34%)	241/649 (37%)	251/635 (40%)	259/710 (36%)
Stored drinking water has detectable free chlorine								
Enrolment	44/1529 (3%)	24/736 (3%)	20/720 (3%)	20/715 (3%)	30/743 (4%)	29/711 (4%)	14/661 (2%)	26/729 (4%)
Year 1	25/847 (3%)	..	151/385 (39%)	18/367 (5%)	20/417 (5%)	180/424 (42%)	9/392 (2%)	156/367 (43%)
Year 2	38/1365 (3%)	..	144/637 (23%)	17/641 (3%)	16/648 (2%)	112/598 (19%)	15/614 (2%)	128/652 (20%)
Access to improved latrine								
Enrolment	309/1788 (17%)	153/878 (17%)	150/844 (18%)	131/836 (16%)	157/847 (19%)	153/867 (18%)	119/794 (15%)	143/872 (16%)
Year 1	178/993 (18%)	..	74/461 (16%)	409/458 (89%)	65/486 (13%)	472/526 (90%)	63/424 (15%)	425/477 (89%)
Year 2	271/1381 (20%)	..	128/664 (19%)	534/683 (78%)	119/654 (18%)	529/644 (82%)	99/613 (16%)	561/706 (79%)
Child faeces safely disposed of								
Enrolment	114/721 (16%)	51/323 (16%)	53/310 (17%)	67/347 (19%)	54/319 (17%)	65/369 (18%)	33/310 (11%)	56/353 (16%)
Year 1	338/903 (37%)	..	158/424 (37%)	317/412 (77%)	157/431 (36%)	326/463 (70%)	155/391 (40%)	287/432 (66%)
Year 2	136/1320 (10%)	..	52/625 (8%)	240/643 (37%)	62/616 (10%)	205/597 (34%)	52/578 (9%)	219/657 (33%)
Handwashing location has water and soap								
Enrolment	96/1913 (5%)	58/936 (6%)	56/902 (6%)	42/890 (5%)	52/914 (6%)	64/912 (7%)	57/843 (7%)	53/918 (6%)
Year 1	124/1043 (12%)	..	53/477 (11%)	49/473 (10%)	381/501 (76%)	416/536 (78%)	61/454 (13%)	381/493 (77%)
Year 2	127/1458 (9%)	..	49/696 (7%)	57/712 (8%)	159/690 (23%)	130/675 (19%)	76/650 (12%)	152/735 (21%)
LNS sachets consumed (% expected)*								
Enrolment	..	..	..	..	..	..	..	..
Year 1	..	..	..	..	..	..	5264/5558 (95%)	5583/5838 (96%)
Year 2	..	..	..	..	..	..	3577/3136 (114%)	4028/3458 (116%)

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 reported proportion of 14 sachets consumed in the past week in index children aged 6–24 months.

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

Diarrhoea prevalence over the past 7 days (combining data from year 1 and year 2) was 27·1% in children in the active control group (figure 2, table 3). The intracluster correlation for diarrhoea was 0·012. Compared with the active control group, the diarrhoea prevalence ratios across all groups were not significantly different from one and differences were not significantly different from zero (figure 2, table 3). Diarrhoea prevalence was the same in the combined water, sanitation, and handwashing group and the individual water, sanitation, and handwashing groups. Although adherence to the water and handwashing interventions was higher in year 1 than in year 2, in an analysis that was not prespecified, diarrhoea prevalence was not significantly lower in any of the intervention groups at year 1 (appendix p 12). The high diarrhoea prevalence was fairly stable over 2 years of follow-up and there were no apparent seasonal trends (appendix p 13). Although we had prespecified a sensitivity analysis by age group of child at year 2, we did not complete this analysis because sample sizes in the age group strata were smaller than expected.

By year 2, when children were between 16 and 31 months old (median 25 months), mean length-for-age *Z* score in children in the active control group was –1·54 (SD 1·11; figure 3). The intracluster correlation for length-for-age *Z* score was 0·037. Compared with the active control group, only nutrition and combined water, sanitation, handwashing, and nutrition had higher length-for-age *Z* score (mean difference in score 0·13 [95% CI 0·01–0·25] for nutrition; 0·16 [0·05–0·27] for combined water, sanitation, handwashing, and nutrition; figure 3). Children in the combined water, sanitation, handwashing, and nutrition group were not significantly taller than children in the nutrition group (mean difference 0·04 [95% CI –0·11 to 0·19]; figure 3). Most length-for-age *Z* score gains in these two groups were already apparent by year 1 (0·11 [–0·01 to 0·22] for nutrition; 0·12 [0·01–0·22] for combined water, sanitation, handwashing, and nutrition; appendix p 14).



Prevelence ratio (95% CI)								
Intervention *vs* control	ref	1·00 (0·90–1·11)	1·02 (0·92–1·14)	0·99 (0·88–1·10)	0·98 (0·87–1·09)	0·96 (0·86–1·07)	0·96 (0·86–1·08)	1·05 (0·94–1·16)
WSH *vs* water			ref			0·94 (0·83–1·07)		
WSH *vs* sanitation				ref		0·99 (0·87–1·13)		
WSH *vs* handwashing					ref	1·01 (0·89–1·16)		

Figure 2: **Intervention effects on diarrhoea prevalence 1 and 2 years after intervention**
Data are mean (95% CI). ref=reference. WSH=water, sanitation, and handwashing.

Mean weight-for-age *Z* score at year 2 was higher in children in the nutrition and combined water, sanitation, handwashing, and nutrition groups than the mean of –0·72 (SD 1·01) in the active control group (table 4). Children in the active control group were close to WHO standards for weight-for-length *Z* score; however, weight-for-length *Z* score at year 2 was higher in the combined water, sanitation, handwashing, and nutrition group (table 4). There were no differences in mean head circumference for age *Z* score at year 2 between children in any of the intervention groups and those in the active control group. Results were similar at year 1, with the exception that differences in mean weight-for-length *Z* score between the active control and two groups with the nutrition intervention appear to have been numerically larger at year 1 (appendix p 15).

Compared with the active control group, a smaller proportion of children in the combined water, sanitation, handwashing, and nutrition group were stunted (too short for their age; –5·4 percentage points [95% CI –9·4 to –1·4]), severely stunted (–2·7 percentage points [–5·1 to –0·2]), or underweight (–3·0 percentage points [–5·4 to –0·6]; table 5); no other groups appeared to affect these outcomes. Notably, there were no significant differences between the combined water, sanitation, handwashing, and nutrition and nutrition groups for any growth outcomes. 1% of active control children were wasted and the proportions were similar across all groups.

Differences in growth outcomes between the active control and intervention groups were similar in magnitude and precision when estimated using adjusted models (appendix pp 16–19). We found no evidence of between-cluster spillover effects (appendix p 20).

The cumulative incidence of all-cause mortality was 3·9% in the active control and ranged from 5·3% in the handwashing group to 2·8% in the combined water, sanitation, handwashing, and nutrition group; none of the differences between intervention groups and the active control were statistically significant at α=0·05 (figure 1, appendix p 21).

	Mean* prevalence	Unadjusted† prevalence difference (95% CI)	Adjusted‡ prevalence difference (95% CI)
Intervention *vs* active control			
Active control	27·1%	..	..
Passive control	27·2%	–0·0 (–2·9 to 2·9)	–0·4 (–3·3 to 2·4)
Water	27·7%	0·7 (–2·3 to 3·6)	0·4 (–3·2 to 4·0)
Sanitation	26·5%	–0·3 (–3·3 to 2·6)	–0·3 (–3·2 to 2·6)
Handwashing	26·1%	–0·6 (–3·5 to 2·3)	–1·1 (–4·0 to 1·8)
Water, sanitation, and handwashing	26·1%	–1·2 (–4·1 to 1·7)	–1·1 (–4·3 to 2·0)
Nutrition	25·5%	–1·0 (–4·0 to 2·0)	–0·6 (–4·0 to 2·7)
Water, sanitation, handwashing, and nutrition	28·4%	1·2 (–1·7 to 4·1)	0·7 (–2·4 to 3·7)
Water, sanitation, and handwashing *vs* single groups			
Water, sanitation, and handwashing	26·1%	..	..
Water	27·7%	–1·6 (–5·1 to 1·9)	–2·1 (–6·0 to 1·8)
Sanitation	26·5%	–0·2 (–3·6 to 3·2)	–0·8 (–4·5 to 2·9)
Handwashing	26·1%	0·4 (–3·2 to 3·9)	0·5 (–3·6 to 4·5)

*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, mother's height, mothers education level, number of children <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 3: **Diarrhoea prevalence from 1 and 2 years (combined) after intervention**

Discussion

In the WASH Benefits cluster-randomised controlled trial, we found no effect of any interventions (improved

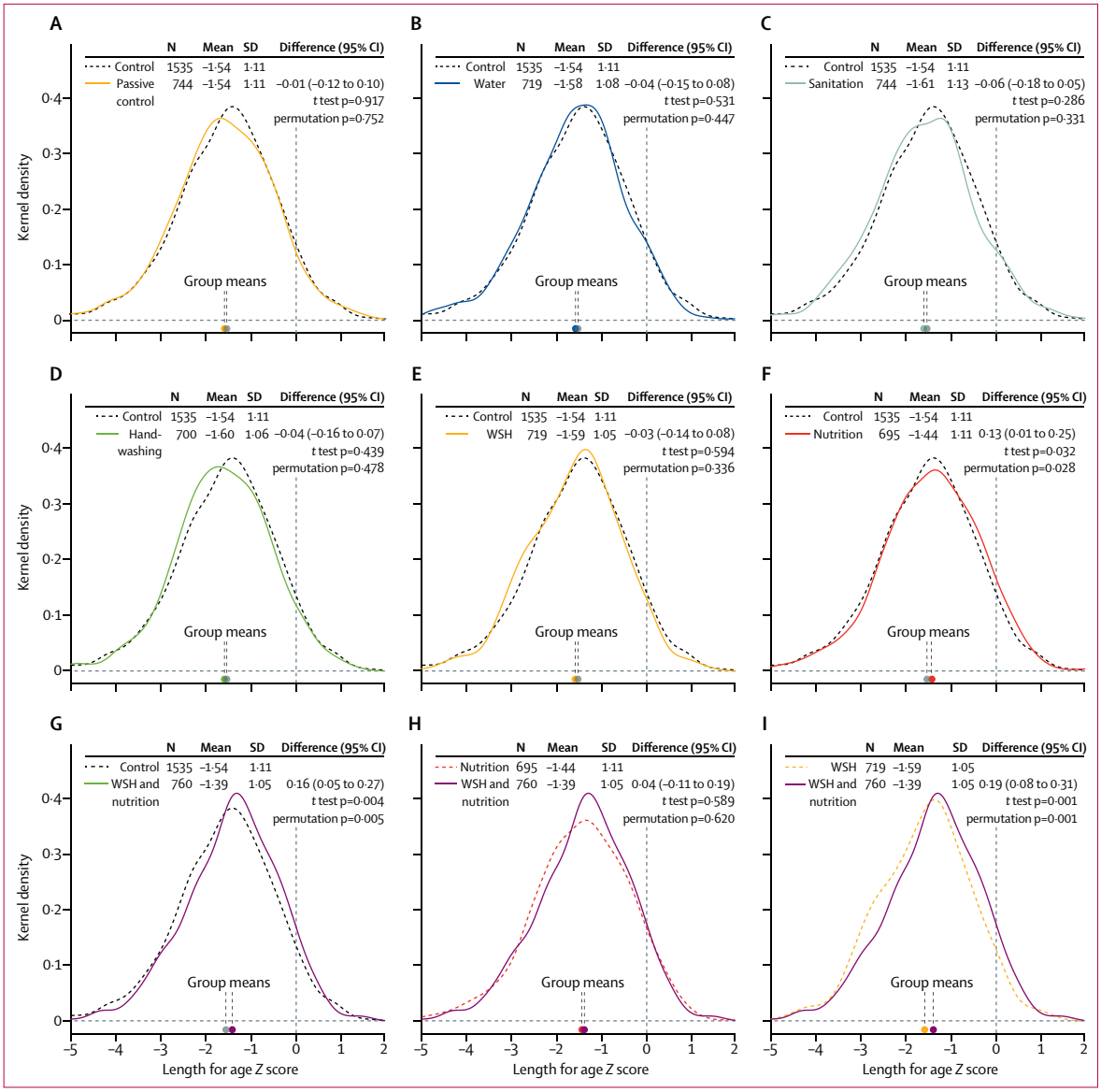


Figure 3: **Intervention effects on length-for-age Z scores in 6583 children after 2 years of intervention**
Kernel density plots show the distribution of length-for-age *Z* scores; dashed lines are the comparison group distribution and solid lines are the active comparator distribution. (A) Passive control *vs* active control. (B) Water *vs* active control. (C) Sanitation vs active control. (D) Handwashing *vs* active control. (E) WSH *vs* active control. (F) Nutrition *vs* active control. (G) WSH and nutrition *vs* active control. (H) WSH and nutrition *vs* nutrition. (I) WSH and nutrition *vs* WSH. p values for *t* test are for differences in group means from zero; permutation p values test the null hypothesis of no difference between groups using a Wilcoxon signed-rank test statistic. WSH=water, sanitation, and handwashing.

water quality, safe sanitation, handwashing, nutrition, or combinations of the interventions) on caregiver-reported diarrhoea prevalence during the first 2 years of life, and improvements in growth were only observed in groups including the nutrition intervention (maternal, infant, and young child feeding counselling and LNS distribution). With a large sample size and high-quality anthropometric measurements, this trial was powered to detect small effects in diarrhoea prevalence and length-for-age *Z* score had they been present. Lower adherence to the water and handwashing interventions by the end of the 2 years of intervention does not seem to be the only explanation for the absence of benefits: there were also no reductions in diarrhoea or improvements in growth in children in the water, handwashing, sanitation, or combined water, sanitation, and handwashing groups even in the first year (a typical measurement point in previous trials), when community-based promoters were most active and adherence was higher, whereas almost all of the growth benefits in the nutrition group and combined water, sanitation, handwashing, and nutrition group were already manifest in the first year. Adherence

	N	Mean (SD)	Difference *vs* active control (95% CI)	Difference *vs* nutrition (95% CI)	Difference *vs* water, sanitation, and handwashing (95% CI)
Weight-for-age *Z* score					
Active control	1548	−0·72 (1·01)	..	..	..
Passive control	721	−0·76 (0·97)	−0·04 (−0·13 to 0·05)	..	..
Water	727	−0·73 (1·00)	0·00 (−0·10 to 0·10)	..	..
Sanitation	747	−0·80 (1·05)	−0·07 (−0·19 to 0·04)	..	..
Handwashing	706	−0·77 (1·01)	−0·05 (−0·15 to 0·05)	..	..
Water, sanitation, and handwashing	725	−0·77 (0·98)	−0·02 (−0·12 to 0·08)	..	..
Nutrition	698	−0·65 (0·98)	0·11 (0·00 to 0·21)	..	..
Water, sanitation, handwashing, and nutrition	765	−0·60 (0·96)	0·14 (0·04 to 0·25)	0·04 (−0·07 to 0·15)	0·17 (0·05 to 0·30)
Weight-for-length *Z* score					
Active control	1536	0·11 (0·94)	..	..	..
Passive control	717	0·08 (0·92)	−0·04 (−0·13 to 0·05)	..	..
Water	719	0·14 (0·95)	0·04 (−0·06 to 0·13)	..	..
Sanitation	740	0·05 (0·97)	−0·05 (−0·14 to 0·05)	..	..
Handwashing	700	0·09 (0·93)	−0·02 (−0·11 to 0·06)	..	..
Water, sanitation, and handwashing	714	0·08 (0·92)	−0·02 (−0·10 to 0·07)	..	..
Nutrition	695	0·14 (0·92)	0·04 (−0·05 to 0·14)	..	..
Water, sanitation, handwashing, and nutrition	762	0·18 (0·90)	0·09 (0·00 to 0·19)	0·04 (−0·05 to 0·13)	0·12 (0·00 to 0·23)
Head circumference-for-age *Z* score					
Active control	1545	−0·27 (1·02)	..	..	..
Passive control	719	−0·27 (1·05)	0·00 (−0·10 to 0·10)	..	..
Water	727	−0·27 (1·03)	0·02 (−0·08 to 0·12)	..	..
Sanitation	745	−0·27 (1·04)	0·01 (−0·09 to 0·11)	..	..
Handwashing	705	−0·29 (0·99)	0·00 (−0·10 to 0·10)	..	..
Water, sanitation, and handwashing	729	−0·30 (0·96)	−0·03 (−0·12 to 0·06)	..	..
Nutrition	695	−0·23 (0·99)	0·05 (−0·05 to 0·15)	..	..
Water, sanitation, handwashing, and nutrition	763	−0·22 (0·99)	0·05 (−0·04 to 0·15)	−0·02 (−0·14 to 0·10)	0·08 (−0·05 to 0·20)

Median child age at 2-year follow-up was 2·05 years (IQR 1·93–2·16). All three secondary outcomes were prespecified.

Table 4: **Child growth *Z* scores at 2-year follow-up**

to the interventions was comparable to or better than what a government or large non-governmental organisation might hope to achieve at scale (appendix p 22), with increases in adherence indicators of 30 percentage points or higher in all intervention groups relative to the control in the first year.

These findings contrast with several systematic reviews[13–15] that have found significant protective benefits of water, sanitation, and hygiene interventions (including handwashing) on diarrhoea in efficacy trials, although most of these studies were shorter and had higher adherence. Results from other trials[16–18] also showed no effect of improved sanitation on diarrhoea, although differences in contexts and interventions complicate comparisons between these trials. Our trial differed from previous trials in that the intervention shifted households from unimproved sanitation (rather than open defecation) to improved sanitation. Additionally, the prevalence of diarrhoea in this study population was high, consistent with prevalence in 12–23-month-old infants measured in the 2014 Kenya Demographic and Health Survey.[19]

A systematic review and meta-analysis[20] of the effects of water quality and supply, sanitation, and hygiene interventions to improve growth identified only five randomised controlled trials of water or handwashing interventions, which did not suggest strong effects on growth, perhaps in part because the interventions lasted only 9–12 months. Since then, five more randomised trials of sanitation interventions have generated mixed evidence on child growth effects: two trials done in India and one in Indonesia had low adherence and no effect, and two done in settings with high rates of open defecation in India and Mali showed improvements in length-for-age *Z* score of 0·18–0·40 in children younger than 5 years.[16–18,20–22] The sanitation intervention in our trial was aligned with the focus on improved latrines initiated under the Millennium Development Goals, and the Sustainable Development Goals' recognition that children's faeces also need to be safely disposed of. This trial and its companion trial[9] in Bangladesh suggest that a compound-level approach to upgrading existing latrines and safely disposing of children's faeces is not sufficient

	n/N (%)	Difference *vs* active control (95% CI)	Difference *vs* nutrition (95% CI)	Difference *vs* water, sanitation, and handwashing (95% CI)
Stunting*				
Active control	483/1535 (31%)	..	..	..
Passive control	223/716 (31%)	−1·7 (−5·9 to 2·5)	..	..
Water	233/719 (32%)	0·1 (−4·2 to 4·3)	..	..
Sanitation	255/739 (35%)	2·3 (−2·0 to 6·6)	..	..
Handwashing	235/700 (34%)	0·8 (−3·5 to 5·1)	..	..
Water, sanitation, and handwashing	236/719 (33%)	1·3 (−3·0 to 5·6)	..	..
Nutrition	201/695 (29%)	−3·2 (−7·5 to 1·1)	..	..
Water, sanitation, handwashing, and nutrition	203/760 (27%)	−5·4 (−9·4 to −1·4)	−2·3 (−7·1 to 2·5)	−5·8 (−10·6 to −1·0)
Severe stunting†				
Active control	143/1535 (9%)	..	..	..
Passive control	62/716 (9%)	−0·8 (−3·3 to 1·8)	..	..
Water	69/719 (10%)	−0·5 (−3·2 to 2·2)	..	..
Sanitation	77/739 (10%)	1·0 (−1·8 to 3·7)	..	..
Handwashing	59/700 (8%)	−1·1 (−3·7 to 1·5)	..	..
Water, sanitation, and handwashing	65/719 (9%)	0·2 (−2·4 to 2·8)	..	..
Nutrition	55/695 (8%)	−1·6 (−4·2 to 1·0)	..	..
Water, sanitation, handwashing, and nutrition	55/760 (7%)	−2·7 (−5·1 to −0·2)	−0·9 (−3·7 to 2·0)	−2·7 (−5·6 to 0·2)
Wasting†				
Active control	22/1536 (1%)	..	..	..
Passive control	10/717 (1%)	0·0 (−1·1 to 1·1)	..	..
Water	9/719 (1%)	−0·2 (−1·3 to 0·8)	..	..
Sanitation	19/740 (3%)	1·1 (−0·3 to 2·4)	..	..
Handwashing	6/700 (1%)	−0·5 (−1·5 to 0·4)	..	..
Water, sanitation, and handwashing	10/714 (1%)	0·2 (−0·9 to 1·2)	..	..
Nutrition	8/695 (1%)	−0·3 (−1·3 to 0·8)	..	..
Water, sanitation, handwashing, and nutrition	11/762 (1%)	−0·1 (−1·2 to 1·0)	0·2 (−1·0 to 1·4)	0·0 (−1·2 to 1·1)
Underweight†				
Active control	148/1548 (10%)	..	..	..
Passive control	70/721 (10%)	−0·4 (−3·0 to 2·2)	..	..
Water	76/727 (10%)	−0·1 (−2·8 to 2·7)	..	..
Sanitation	87/747 (12%)	1·6 (−1·2 to 4·4)	..	..
Handwashing	71/706 (10%)	0·5 (−2·2 to 3·3)	..	..
Water, sanitation, and handwashing	72/725 (10%)	0·5 (−2·3 to 3·2)	..	..
Nutrition	59/698 (8%)	−1·2 (−3·9 to 1·5)	..	..
Water, sanitation, handwashing, and nutrition	52/765 (7%)	−3·0 (−5·4 to −0·6)	−1·8 (−4·7 to 1·1)	−3·3 (−6·2 to −0·5)

Median child age at 2-year follow-up was 2·05 years (IQR 1·93–2·16). *Prespecified secondary outcome. †Prespecified tertiary outcome.

Table 5: **Proportion of children stunted, severely stunted, wasted, and underweight at 2-year follow-up**

to improve child growth, and neither are water and handwashing interventions.

Conversely, counselling and LNS provided in the nutrition group improved length-for-age *Z* score by year 2. Compared with randomised controlled trials of LNS during complementary feeding, our finding of length-for-age *Z* score improvements of 0·13–0·16 in the nutrition groups falls in the middle of the spectrum between four trials: one from Malawi[23] that reported no effect on length-for-age *Z* score, one from Haiti[24] and one from Bangladesh[25] that reported an effect on length-for-age *Z* score comparable to this study, and one from Burkina Faso[26] that reported a larger effect on length-for-age *Z* score. Thus, there appears to be consistent evidence that LNS distribution together with some promotion of improved infant and young child feeding can reduce growth faltering, although this approach falls far short of eliminating the problem. Interventions will likely need to address the complex set of underlying determinants of growth faltering, including prenatal or preconception factors. Future analyses will explore changes in feeding practices that resulted from the intervention.

Although there were more improvements in anthropometric measures in the combined water, sanitation, handwashing, and nutrition group versus active control

than in the nutrition versus active control group, the differences were of little clinical or statistical significance. We conclude that combining nutrition with water, sanitation, and handwashing did not provide additional growth benefits beyond nutrition alone. Although the effect of water, sanitation, handwashing, and nutrition on mortality was not significant, the lower mortality in that group is consistent with the statistically significant effect of water, sanitation, handwashing, and nutrition on mortality in the Bangladesh trial.[9] Pending analyses will evaluate potential differences in effects on other child health outcomes.

It is possible that the water, sanitation, and handwashing interventions delivered in this trial did not sufficiently address important transmission routes for enteric pathogens.[11] Although the sanitation intervention included a sani-scoop and messages about preventing children from being exposed to domestic animal faeces, the emphasis was mostly on behaviours related to human faeces and might not have protected children from zoonotic pathogens.[27] Although chlorination of water has the advantage of providing residual protection against recontamination, it is not effective against protozoa such as *Giardia lamblia* and *Cryptosporidium* spp, the latter of which was identified as one of the most common causes of moderate-to-severe diarrhoea in children 0–23 months in a neighbouring part of Kenya.[28] Other limitations of this trial include the inability to mask the interventions; the absence of observable indicators of actual behaviour for the handwashing, sanitation, and nutrition interventions; lower adherence to the water and hygiene interventions during the second year of the trial than in the first year; and the use of a compound-level sanitation intervention, as opposed to community-level. Because masking was not possible, we focused on objective, observable indicators whenever possible rather than self-reported behaviours, recognising that the availability of a latrine or handwashing station stocked with water and soap does not necessarily imply that the materials were used. Despite an intensive design process that drew heavily on best practices in behaviour change, incorporation of lessons learned from the pilot randomised controlled trial, thorough verification of availability of the intervention materials, and periodic monitoring of indicators of recommended behaviours, adherence to the water and handwashing interventions appeared to reduce sharply in the last months of the trial. The waning intensity of promotion activities after a reduction in the stipend given to the health promoters could at least partly explain the drop in adherence. Finally, by contrast with water, handwashing, and nutrition interventions that directly benefit households that adhere to the intervention, a sanitation intervention in only a subset of compounds might not be sufficient to protect against exposure to faecal contamination in the environment that originates from other compounds in the community. We decided, however, to deliver compound-level interventions based on evidence that child exposure to enteric pathogens during the first 2 years of life occurs predominantly within the household compound.[29] Because environmental contamination and disease transmission pathways could be different in densely populated contexts, similar studies in urban areas would complement this rural trial.

Additional outcome measures collected in this trial will help to elucidate potential mechanisms for the observed effects, including indicators of environmental contamination, environmental enteric dysfunction, anaemia, enteric parasite infection, and child development. Molecular measurement of infections in the laboratory with stored stool specimens collected as part of this trial offer an opportunity for unbiased indicators of pathogen burden.

More intensive promotion and higher adherence could have resulted in larger effects than those reported, but our findings are relevant for large-scale programmes that struggle to achieve adherence rates as high as those of efficacy studies. The potential for water, sanitation, hygiene, and nutrition interventions to reduce diarrhoea and improve growth might be highly context-dependent. In our rural setting, water was plentiful but rarely available on premises, susceptible to contamination at the source and in storage, and rarely treated despite introduction of a nearly-universal filter distribution programme;[30] unimproved latrine coverage was high and there was a culture of using sanitation facilities for defecation by human beings, but there was probably persistent exposure to animal faeces; handwashing was not a common practice; breastfeeding was common, but exclusive breastfeeding was not, and most people had enough food, but not a diverse diet; diarrhoea prevalence was high; and many children had low length-for-age *Z* score, but not weight-for-length *Z* score. Our findings call into question the ability of large-scale water, sanitation, and handwashing interventions to reduce diarrhoea or improve growth. Our results suggest that integrated water, sanitation, and handwashing and nutrition programmes are no more effective than nutrition programmes at reducing diarrhoea or improving growth, and that nutritional interventions that include counselling and LNS can modestly reduce growth faltering, but fall short of eliminating it, even when LNS adherence is high.

Contributors

CN and CPS contributed equally to the manuscript. CN drafted the research protocol and manuscript with input from all listed coauthors and oversaw all aspects of the trial. CPS led the nutrition intervention and protocols for anthropometry data collection. KGD contributed to the nutrition intervention and interpretation of results. PK assisted with oversight of anthropometry data collection. CN, AJP, HND, TC, and PKR developed the water, sanitation, and handwashing intervention. CN, CPS, AJP, HND, and GN oversaw piloting and subsequent study implementation, contributed to refinements in interventions and measurements, and responded to threats to validity. CN, CPS, BFA, CDA, JB-C, AEH, AM, and JMC Jr developed the analytical approach, did the statistical analysis, and constructed the tables and figures. AL, SPL, and JMC Jr advised on harmonising the trials between Kenya

and Bangladesh and SMN helped adapt the trial to the Kenyan context. CN, CPS, AJP, BFA, JB-C, LCHF, AL, SPL, SMN, and JMC Jr secured funding for the trial. All authors have read, contributed to, and approved the final version of the manuscript.

Declaration of interests

All authors received funding for either salary or consulting fees through a grant from the Bill & Melinda Gates Foundation for this study.

Acknowledgments

We thank the study participants and promoters who participated in the trial, the fieldworkers who delivered the interventions and collected the data for the study, and the managers who ensured that everything ran smoothly. This research was financially supported in part by Global Development grant OPPGD759 from the Bill & Melinda Gates Foundation to the University of California, Berkeley, CA, USA, and grant AID-OAA-F-13-00040 from United States Agency for International Development (USAID) to Innovations for Poverty Action. This manuscript was made possible by the generous support of the American people through the USAID. The contents are the responsibility of the authors and do not necessarily reflect the views of USAID or the US Government.

For a **list of managers** see http://washbenefits.net

References

1 United National Children's Fund, WHO, World Bank Group. Levels and trends in child malnutrition. 2016. http://www.who.int/nutgrowthdb/jme_brochure2016.pdf?ua=1 (accessed March 18, 2017).

2 Sudfeld CR, McCoy DC, Danaei G, et al. Linear growth and child development in low- and middle-income countries: a meta-analysis. *Pediatrics* 2015; **135:** e1266–75.

3 Prendergast AJ, Humphrey JH. The stunting syndrome in developing countries. *Paediatr Int Child Health* 2014; **34:** 250–65.

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

5 Black RE, Victora CG, Walker SP, et al. Maternal and child undernutrition and overweight in low-income and middle-income countries. *Lancet* 2013; **382:** 427–51.

6 Scrimshaw NS, Taylor CE, Gordon JE. Interactions of nutrition and infection. *Monogr Ser World Health Organ* 1968; **57:** 3–329.

7 GBD 2015 Mortality and Causes of Death Collaborators. 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.

8 Richard SA, Black RE, Gilman RH, et al. Diarrhoea in early childhood: short-term association with weight and long-term association with length. *Am J Epidemiol* 2013; **178:** 1129–38.

9 Luby SP, Rahman M, Arnold BF, et al. Effects of water quality, sanitation, handwashing and nutritional interventions on diarrhoea and child growth in rural Bangladesh: a cluster randomised trial. *Lancet Glob Health* 2018; published online Jan 29. http://dx.doi.org/10.1016/S2214-109X(17)30490-4.

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 Christensen G, Dentz HN, Pickering AJ, et al. Pilot cluster randomized controlled trials to evaluate adoption of water, sanitation, and hygiene interventions and their combination in rural western Kenya. *Am J Trop Med Hyg* 2015; **92:** 437–47.

12 Balzer L, van der Laan M, Petersen M. Adaptive pre-specification in randomized trials with and without pair-matching. *Stat Med* 2016; **35:** 4528–45.

13 Wolf J, Prüss-Ustün A, Cumming O, et al. Assessing the impact of drinking water and sanitation on diarrhoeal disease in low- and middle-income settings: systematic review and meta-regression. *Trop Med Int Health* 2014; **19:** 928–42.

14 Clasen TF, Alexander KT, Sinclair D, et al. Interventions to improve water quality for preventing diarrhoea. *Cochrane Database Syst Rev* 2015; **10:** CD004794.

15 Ejemot-Nwadiaro RI, Ehiri JE, Arikpo D, Meremikwu MM, Critchley JA. Hand washing promotion for preventing diarrhoea. *Cochrane Database Syst Rev* 2015; **9:** CD004265.

16 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.

17 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.

18 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.

19 Kenya National Bureau of Statistics, Ministry of Health, National AIDS Control Council, Kenya Medical Research Institute, National Council for Population and Development, ICF International. Kenya Demographic and Health Survey 2014. Rockville, MD, USA: ICF International, 2015.

20 Dangour AD, Watson L, Cumming O, et al. Interventions to improve water quality and supply, sanitation and hygiene practices, and their effects on the nutritional status of children. *Cochrane Database Syst Rev* 2013; **8:** CD009382.

21 Cameron L, Shah M, Olivia S. Impact evaluation of a large-scale rural sanitation project in Indonesia. World Bank Policy Research Working Paper 6360. Washington, DC: World Bank, 2013.

22 Hammer J, Spears D. Village sanitation and children's human capital: evidence from a randomized experiment by the Maharashtra government. World Bank Policy Research Working Paper 6580. Washington, DC: World Bank, 2013.

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 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.

25 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.

26 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.

27 Harris A, Pickering AJ, Harris M, et al. Ruminants contribute fecal contamination to the urban household environment in Dhaka, Bangladesh. *Environ Sci Technol* 2016; **50:** 4642–49.

28 Kotloff KL, Nataro JP, Blackwelder WC, et al. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *Lancet* 2013; **382:** 209–22.

29 Ngure F, Reid BM, Humphrey JH, Mbuya MN, Pelto G, Stolzfus RJ. Water, sanitation, and hygiene (WASH), environmental enteropathy, nutrition, and early child development: making the links. *Ann NY Acad Sci* 2014; **1308:** 118–28.

30 Pickering AJ, Arnold BF, Dentz HN, Colford JM, Null C. Climate and health co-benefits in low-income countries: a case study of carbon-financed water filters in Kenya and a call for independent monitoring. *Environ Health Perspect* 2017; **125:** 278–83.

MICHELLE RUIZ: Hello this is Michelle Ruiz, instructional designer at the school of public health. Welcome to the second episode of “Epidemiology Case Studies” In the first episode, Doctor Jade Benjamin- Chung interviewed Doctor Ben Arnold about the surfer health study.
Here is the second episode about the Wash benefits study.

JADE BENJAMIN-CHUNG: Today, we're going to be talking about the WASH Benefits trials, which were conducted in Bangladesh and in Kenya and recently wrapped up within the last year or so. And I'm going to be interviewing Jack Colford. And I just wanted to start by mentioning that we've actually both worked on this trial quite a bit but I'm going to pretend like I don't know anything about WASH Benefits and just ask Jack a lot of questions. So let's get started. So can you tell me about the hypotheses that WASH Benefits was designed to test?

JACK COLFORD: Sure. So WASH Benefits was really trying to ask the question about whether we could improve growth in children. And we thought that there might be lots of different things that affect growth in children in developing country settings, but trying to tease out which ones we could intervene on and make a difference for was really the central hypothesis of WASH Benefits.

JADE BENJAMIN-CHUNG: And how did you become interested in that as a research question?

JACK COLFORD: Well my whole career for the past 25 years has involved many different types of drinking water, sanitation, hygiene-type studies. And the WASH Benefits study-- and by the way, WASH stands for water, sanitation, and hygiene, and benefits, obviously, we're looking for the benefits of WASH interventions-- the idea here was that we could bring together multiple interventions at the same time, or kind of this idea can you chew and spit gum at the same time, do more than one thing at a time, intervene through a couple of different mechanisms and have a synergistic effect?

JADE BENJAMIN-CHUNG: So you're a so-called WASH person, not a nutrition person traditionally.

JACK COLFORD: Correct.

JADE BENJAMIN-CHUNG: So tell me how WASH can be related to child growth, which is something that most people might not expect.

JACK COLFORD: Got it. Got it. Well, so certainly there's no difficulty in concluding that nutrition is associated with child growth. We're trying to test whether WASH has an important impact on child growth. So under this idea of trying to bring multiple interventions together, it just struck us as very sensible to bring together nutrition with WASH, which wasn't really being done much. And with the support of the Gates Foundation, that was able to enlarge the study to do this, we were able to add nutrition to certain arms of the study to test the individual effective of nutrition, compare it to the combined effect of WASH plus nutrition, and compare that to the effect of WASH interventions alone.

JADE BENJAMIN-CHUNG: And so you've referred to these different kinds of interventions. Can you tell me in a little more detail how you selected those interventions for this trial?

JACK COLFORD: Of course. So there had been decades of research-- often observational research-- on individual types of intervention. So for example, should you treat the water with chlorine or with filters? How should you get people to wash their hands more often? What kind of latrines should be built to help people use the latrines, and so on, and so forth.

So from multiple decades of formational research on different specific interventions, we then worked for a couple of years before the trial even began to figure out in the two settings we were working in-- Kenya and Bangladesh-- which of these particular interventions would be most accepted by the people and easiest to use. And if you do an intervention and people don't comply with it, it's not really an intervention. So the idea was to find in the WASH interventions that people in these cultural settings would understand, use, and stick with.

JADE BENJAMIN-CHUNG: So you've talked about these interventions. Let's zero in on one, so let's talk about sanitation for a moment. Tell me how improving the sanitation of a child's home could improve their growth. What the biological theory of change for how that would work?

JACK COLFORD: So the idea here is that fecal matter-- in particular, human fecal matter-- is the most dangerous substance in terms of transmission of infectious disease pathogens to children who might be exposed to it from contaminated food, or water, or even from eating soil. Children all around the world will put soil in their mouth as they're playing on the ground and so forth. So the idea is, will the use of sanitation intervention-- and that's not just a latrine, but it also involved providing tools for cleaning up fecal matter from around the courtyard and around the house.

So would all these sanitation interventions as a sanitation package help to reduce the burden of fecal ingestion the children would have for organisms that might cause diarrhea. And subsequently, the idea was that the diarrhea would lead to poor growth. So interrupting the organisms that caused diarrhea, that caused poor growth, that's the whole chain of thought there of the theory of change.

JADE BENJAMIN-CHUNG: So you've mentioned child diarrhea and child growth. Those are the primary outcomes of the trial. What were some of the other secondary and tertiary outcomes that were measured.

JACK COLFORD: Quite a number of other outcomes were part of the study. For instance, we were measuring the occurrence of intestinal parasitic infections. We were measuring anemia. We were measuring many other aspects of growth beyond length. So length is the primary

measurement, but head circumference and other elements of growth were part of that too.

In addition to intestinal parasites, we were measuring bacterial infections and other gastrointestinal conditions children might develop. We have future studies plan that will look at serologic changes-- changes in the antibodies of the blood of the children that was collected and stored. So quite a host of different outcomes. And a lot of these things are banked and will be done in future years.

JADE BENJAMIN-CHUNG: The trials were conducted in Bangladesh and Kenya. How were those locations selected?

JACK COLFORD: So one of the hidden facts of doing epidemiology is a lot of things have to come together to work right to make a study work, so it's almost like intersecting Venn diagrams of having study personnel in a country that can do a project, having experience working in that country, having permission to work in their country, having a funder be interested in working in that country. So when those four or five diagrams all come together, there's usually only an overlap in a very few number of countries.

We originally were going to do the study in four different countries but then the 2008 recession hit. And even the Gates Foundation was affected by the global recession.

JADE BENJAMIN-CHUNG: That's the funder of the trials.

JACK COLFORD: The Gates Foundation was the funder of the trials. So they asked if we could do the study at a very large scale but do it in two countries rather than four, so we reduced the number to two. So all those things coming together led us to Bangladesh and Kenya.

JADE BENJAMIN-CHUNG: And how different are those two settings with respect to water and sanitation and then child growth.

JACK COLFORD: So the reason to do it in more than one country was specifically to do it in settings that differed quite a bit. And for instance, one of the things that's quite different is the water availability. In Bangladesh, there's lots of water-- a high water table, lots of water available for use. In Kenya, there's not. In Bangladesh, the density of the villages, people are much closer to each other than in Kenya, for instance, where they're spread apart. So there are a number of different axes on which the two sites differ, because we wanted to see did the interventions have a different impact in different settings.

And I should say, each trial was designed to stand independently. Each trial was a test of the

hypotheses on its own. The data didn't need to be combined in order to do the analyzes.

JADE BENJAMIN-CHUNG: So the trial was a randomized trial with cluster randomization and it also used a factorial design. Tell me how you came to that particular design for this research question.

JACK COLFORD: Well, one of the elements of working with water and sanitation is that there's the strong potential that neighboring people are affected by an intervention that you might receive yourself. So if your household has a new latrine, that could have an impact on a neighboring household. If your household has a new water supply, that could, a, be difficult to deliver just to you, but also have a health impact on neighboring people living nearby.

So we call all of that a spillover effect. So one of the issues when you study interventions that have potential spillover effects, for statistical soundness, we need to have the elements that we're studying-- the individuals or the families-- be separated enough from each other that the spillover effects don't drown out what's going on. And it's very difficult with water sanitation hygiene interventions to work just in a village with people receiving different interventions. So it's very common in the field to do these types of studies where villages are the level of the intervention in order to prevent this spillover contact.

So then we pick villages that are far enough apart from each other-- and the villages are the cluster here, you referred to a cluster randomized trial-- the villages are the cluster. So the clusters are far enough apart from each other that they are not affecting each other either. And then it's, obviously, a randomized trial because we're trying to reduce confounding and all the things we've studied in the class about why we do randomization.

JADE BENJAMIN-CHUNG: And also, the factorial design.

JACK COLFORD: And the factorial design is an efficiency tool. So factorial means we're combining some interventions together and also studying them individually. So for example, we have--

JADE BENJAMIN-CHUNG: So the trial also used a factorial design. Tell us about that.

JACK COLFORD: So factorial design is a very efficient way to study multiple interventions alone and combined in the same trial. So for example, one of our arms had water sanitation and hygiene interventions all delivered to the participants, whereas other arms had only water, or only sanitation, or only hygiene interventions delivered.

So that obviously gives us the chance to see whether the combination provides more benefit

than the individual intervention alone. So factorial design basically is bringing a study design together in which combined interventions are studied both in combination and separately for this kind of efficiency rather than having to do separate trials.

JADE BENJAMIN-CHUNG: So the hypothesis then was that the combination of these interventions would actually have a synergistic effect on growth and diarrhea.

JACK COLFORD: Correct. We wanted to measure whether there was a stronger benefit to the combination than from the individual arms. Although we also worried that delivering the combination arms could be harder to do than delivering the single intervention arms because there's just more logistics involved in bringing three interventions together than one at a time.

JADE BENJAMIN-CHUNG: And how did you plan to measure that. How did you do plan to assess whether the interventions were delievered as intended.

JACK COLFORD: So we had a number of different compliance measures to record for each type of intervention, whether people were receiving the intervention as intended. So for example, with chlorination of water we can measure free chlorine in the water. With the latrines we can measure the usage of latrines both as reported or as observed.

With hand-washing we can measure, when we come back on unannounced spot visits, whether there is soap present at the hand-washing station that we hit established. With nutrition we could count up how many of the nutrition packets were still present at each visit. And again, it's best that the subjects not know that you're doing this, otherwise they might change their behavior to better comply with what you hope they do.

JADE BENJAMIN-CHUNG: So this sounds like a really huge undertaking. Can you tell me a little bit about who was involved, roughly how many people, and where they were located.

JACK COLFORD: Certainly. Yeah. So, of course, we had two countries. So everything's replicated in the two different countries. There's large teams of investigators both in the country with local teams, and then there's a number of different institutions involved with subject matter expertise in the study. So UC Davis was very involved with nutrition. The University of Buffalo was very involved with hand-washing.

Johns Hopkins was very involved with behavioral change theory and how to introduce these interventions that require a lot of behavioral intervention. Stanford was involved with leading the trial in Bangladesh. So just a huge, huge array of co-investigators sort of working closely

with Berkeley and trying to make this all come together.

JADE BENJAMIN-CHUNG: Excellent. Well thanks for telling me about the study design.

Thank you.

JADE BENJAMIN-CHUNG: When we come back, we'll talk about potential sources of bias and confounding, the analysis of the trials, and what the trials found.

JACK COLFORD: Thanks.

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

Clair Null, Christine P Stewart, Amy J Pickering, Holly N Dentz, Benjamin F Arnold, Charles D Arnold, Jade Benjamin-Chung, Thomas Clasen, Kathryn G Dewey, Lia C H Fernald, Alan E Hubbard, Patricia Kariger, Audrie Lin, Stephen P Luby, Andrew Mertens, Sammy M Njenga, Geoffrey Nyambane, Pavani K Ram, John M Colford Jr

Summary

Background Poor nutrition and exposure to faecal contamination are associated with diarrhoea and growth faltering, both of which have long-term consequences for child health. We aimed to assess whether water, sanitation, handwashing, and nutrition interventions reduced diarrhoea or growth faltering.

Methods The WASH Benefits cluster-randomised trial enrolled pregnant women from villages in rural Kenya and evaluated outcomes at 1 year and 2 years of follow-up. Geographically-adjacent clusters were block-randomised to active control (household visits to measure mid-upper-arm circumference), passive control (data collection only), or compound-level interventions including household visits to promote target behaviours: drinking chlorinated water (water); safe sanitation consisting of disposing faeces in an improved latrine (sanitation); handwashing with soap (handwashing); combined water, sanitation, and handwashing; counselling on appropriate maternal, infant, and young child feeding plus small-quantity lipid-based nutrient supplements from 6–24 months (nutrition); and combined water, sanitation, handwashing, and nutrition. Primary outcomes were caregiver-reported diarrhoea in the past 7 days and length-for-age *Z* score at year 2 in index children born to the enrolled pregnant women. Masking was not possible for data collection, but analyses were masked. Analysis was by intention to treat. This trial is registered with ClinicalTrials.gov, number NCT01704105.

Findings Between Nov 27, 2012, and May 21, 2014, 8246 women in 702 clusters were enrolled and randomly assigned an intervention or control group. 1919 women were assigned to the active control group; 938 to passive control; 904 to water; 892 to sanitation; 917 to handwashing; 912 to combined water, sanitation, and handwashing; 843 to nutrition; and 921 to combined water, sanitation, handwashing, and nutrition. Data on diarrhoea at year 1 or year 2 were available for 6494 children and data on length-for-age *Z* score in year 2 were available for 6583 children (86% of living children were measured at year 2). Adherence indicators for sanitation, handwashing, and nutrition were more than 70% at year 1, handwashing fell to less than 25% at year 2, and for water was less than 45% at year 1 and less than 25% at year 2; combined groups were comparable to single groups. None of the interventions reduced diarrhoea prevalence compared with the active control. Compared with active control (length-for-age *Z* score –1·54) children in nutrition and combined water, sanitation, handwashing, and nutrition were taller by year 2 (mean difference 0·13 [95% CI 0·01–0·25] in the nutrition group; 0·16 [0·05–0·27] in the combined water, sanitation, handwashing, and nutrition group). The individual water, sanitation, and handwashing groups, and combined water, sanitation, and handwashing group had no effect on linear growth.

Interpretation Behaviour change messaging combined with technologically simple interventions such as water treatment, household sanitation upgrades from unimproved to improved latrines, and handwashing stations did not reduce childhood diarrhoea or improve growth, even when adherence was at least as high as has been achieved by other programmes. Counselling and supplementation in the nutrition group and combined water, sanitation, handwashing, and nutrition interventions led to small growth benefits, but there was no advantage to integrating water, sanitation, and handwashing with nutrition. The interventions might have been more efficacious with higher adherence or in an environment with lower baseline sanitation coverage, especially in this context of high diarrhoea prevalence.

Funding Bill & Melinda Gates Foundation, United States Agency for International Development.

Lancet Glob Health 2018; 6: e316–29

Published Online January 29, 2018 http://dx.doi.org/10.1016/S2214-109X(18)30005-6

See Comment page e236

See Articles page e302

Innovations for Poverty Action, Kakamega, Kenya (C Null PhD, H N Dentz MPH, G Nyambane MA); **Center for International Policy Research and Evaluation, Mathematica Policy Research, Washington, DC, USA** (C Null); **Rollins School of Public Health, Emory University, Atlanta, GA, USA** (C Null, Prof T Clasen PhD); **Department of Nutrition, University of California, Davis, CA, USA** (C P Stewart PhD, H N Dentz, C D Arnold MS, Prof K Dewey PhD); **Department of Civil and Environmental Engineering** (A J Pickering PhD), **and Department of Infectious Diseases and Geographic Medicine** (S P Luby MD), **Stanford University, Stanford, CA, USA; Department of Civil and Environmental Engineering, Tufts University, Medford, MA, USA** (A J Pickering); **Division of Epidemiology** (B F Arnold PhD, J Benjamin-Chung PhD, A Lin PhD, A Mertens MS, Prof J M Colford Jr MD), **Division of Community Health Sciences** (Prof L C H Fernald PhD, P Kariger PhD), **and Division of Biostatistics** (A E Hubbard PhD) **School of Public Health, University of California, Berkeley, CA, USA; Eastern and Southern Africa Centre of International Parasite Control, Kenya Medical Research Institute, Nairobi, Kenya** (S M Njenga PhD); **Department of Epidemiology and Environmental Health, School of Public Health and Health**

Introduction

An estimated 156 million children worldwide suffer from stunting (linear growth faltering) and are unlikely to reach their full potential as adults.[1] Linear growth faltering is the most apparent sign of chronic undernutrition and is the physical manifestation of combined physiological and developmental insults. Early-life stunting leads to poor cognitive development in childhood, reduced

Professions, University at Buffalo, Buffalo, NY, USA
(P K Ram MD)

Correspondence to:
Dr Clair Null, Center for International Policy Research and Evaluation, Mathematica Policy Research, Washington, DC 20002, USA
cnull@mathematica-mpr.com

Research in context

Evidence before this study

Malnutrition and enteric infection are thought to act together to impair child health and survival, yet there is limited evidence of low cost, scalable interventions effective at breaking this cycle. A 2008 meta-analysis by Dewey and Adu-Afarwuah found that interventions offering nutrient supplementation or counselling on complementary feeding could result in modest improvements to child growth. Another meta-analysis by Waddington and Snilsveit in 2009 showed that water treatment or handwashing could prevent diarrhoea, but there had not been any randomised trials of the effect of sanitation on diarrhoea. During this study, five other randomised trials of the effects of sanitation on diarrhoea and growth were published, but three were limited by low adherence. Whether combining water, sanitation, handwashing, or nutrition interventions could result in added benefits for health and growth was not known.

Added value of this study

This trial is one of the first to provide experimental evidence on whether individual and combined water, sanitation, or handwashing interventions improve growth; combined water, sanitation, and handwashing interventions are more effective at reducing diarrhoea and growth faltering than any intervention alone; and nutrition counselling and supplementation are more effective when combined with improved water, sanitation, and handwashing. This is the first rigorous evaluation of upgrading from unimproved to improved latrines in sub-Saharan Africa. None of the interventions reduced diarrhoea, and only the interventions that included nutrition counselling and nutrient supplementation improved growth.

Implications of all the available evidence

Our results on growth effects are consistent with those from previous research on the combination of nutrition counselling and nutrient supplementation, finding modest effects on linear growth. It is possible that more intensive promotion and higher adherence would have resulted in larger effects, especially in this context of high diarrhoea prevalence, but few programmes are likely to be able to afford sustaining a more ambitious behaviour change programme than was included in this trial. In a context where most households already had an unimproved sanitation facility, provision of technologically simple interventions including chlorination for household treatment of drinking water, improved pit latrines, and handwashing stations—standard for most WASH programmes in rural areas of low-income countries—might not be sufficient to improve growth. By contrast with previous studies, this trial provided evidence that technologically simple water, sanitation, and handwashing interventions with adherence rates at least as high as most programmes achieve might not reduce childhood diarrhoea in all situations.

economic productivity in adulthood, and increased risk of morbidity and mortality.[2,3] Because nutrient supplementation and counselling interventions for maternal, infant, and young child feeding have been only marginally successful at preventing growth faltering, exposure to faecal contamination in the environment has recently been hypothesised to lead to environmental enteric dysfunction, which features chronic immune stimulation and impaired nutrient absorption, thereby constraining a growth response to improved nutrition.[4] In addition to the detrimental effects on growth and development, undernutrition was estimated to cause 45% of all child deaths in 2011, and it has long been recognised that undernutrition is an important determinant of susceptibility to infectious disease.[5,6] Diarrhoea is the second leading cause of death in children aged 1–59 months, contributing to almost 500 000 deaths in children younger than 5 years in 2015.[7] Frequent diarrhoea is also associated with linear growth faltering.[8] If there is a pathway independent of symptomatic diarrhoea linking environmental contamination to growth faltering, the benefits of improving water safety, sanitation, and handwashing could be underestimated because studies have generally focused on diarrhoea. It is unclear whether combined water, sanitation, handwashing, and nutritional interventions reduce diarrhoea or improve growth more than single interventions.

See Online for appendix

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[9] in Bangladesh evaluated the same objectives.

Methods

Study design

The Kenya WASH Benefits study was a cluster-randomised trial done in rural villages in Bungoma, Kakamega, and Vihiga counties in Kenya's western region (appendix p 11). We used a cluster design to facilitate the logistics of the behaviour change component of the interventions and minimise contamination between intervention and comparison households. We hypothesised that the interventions would improve the health of the index child in each household. We optimised the trial design to measure group-level differences in primary outcomes by including a large number of clusters, each comprising relatively few children (12 on average) with infrequent measurement. Each measurement round lasted roughly 1 year and was balanced across treatment

groups and geography to minimise seasonal or geographic confounding when comparing outcomes across groups.

With active and passive control groups and six intervention groups (water; sanitation; handwashing; combined water, sanitation, and handwashing; nutrition; and combined water, sanitation, handwashing, and nutrition), the design enabled 11 comparisons of each intervention group with the active control; combined water, sanitation, and handwashing with each intervention alone; and combined water, sanitation, handwashing, and nutrition with nutrition alone, and combined water, sanitation, and handwashing. A double-sized active control group was used to increase power because there were six separate intervention comparisons against control.[10] Households in the active control and all intervention groups were visited by community-based health promoters monthly to measure the child's mid-upper arm circumference. Health promoters did not visit households in passive control clusters. Measurement of outcomes, as well as water, sanitation, handwashing, and nutrition characteristics were measured in the passive control group at the same times as in other groups. The study design and rationale have been published previously.[10]

The study protocol was approved by the Committee for the Protection of Human Subjects at the University of California, Berkeley (protocol number 2011-09-3654), the institutional review board at Stanford University (IRB-23310), and the scientific and ethics review unit at the Kenya Medical Research Institute (protocol number SSC-2271). Under direction of the study investigators, Innovations for Poverty Action (IPA) was responsible for intervention delivery and data collection.

Participants

Villages were eligible for selection into the study if they were rural, most of the population relied on communal water sources and had unimproved sanitation facilities, and there were no other ongoing water, sanitation, handwashing, or nutrition programmes. Participants were identified through a complete census of eligible villages. Within selected villages, women were eligible to participate if they reported that they were in their second or third trimester of pregnancy, planned to continue to live at their current residence for the next 2 years, and could speak Kiswahili, Luhya, or English well enough to respond to an interviewer administered survey. IPA staff formed clusters from one to three neighbouring villages to have six or more pregnant women per cluster after the enrolment survey. Outcomes were assessed in the children born from these pregnancies (index children), including twins. Although the study area is one of the areas with the highest HIV prevalence in Kenya, according to the 2012 Kenya AIDS Indicator Survey, the prevalence in women aged 15–64 years in the study area was below 6% (that survey did not include testing of children). Because there would not have been sufficient sample size to allow for subgroup analysis by HIV status, no attempt was made to identify participants who were HIV positive. Participants gave written informed consent before enrolment.

Randomisation and masking

Clusters were randomly allocated to treatment using a random number generator with reproducible seed at the University of California, Berkeley. Groups of nine geographically-adjacent clusters were block-randomised into a double-sized active control; passive control; water; sanitation; handwashing; water, sanitation, and handwashing; nutrition; or water, sanitation, handwashing, and nutrition. Allocation by cluster identification number was communicated directly to the field team; investigators remained blinded to treatment assignments. Blinding of participants was not possible. Participants were informed of their treatment assignment after baseline data collection and might have known the treatment assignment of nearby villages. The health promoters and staff who delivered the interventions were not involved in data collection, but the data collection team could have inferred treatment status if they saw intervention materials in study communities.

Procedures

The interventions were designed to maximise adherence to behaviours that could protect children from exposure to pathogens in their environment and improve diet quality. Formative research in the study area concluded that the health benefits of target behaviours were already well understood, but this knowledge was not sufficient to lead to action. As such, the behaviour change strategy and intervention materials were selected to create enabling environments, build supportive social norms, and target emotional drivers of decision making. The messages and delivery modes for the behaviour change strategy drew from existing information, education, and communication materials from organisations such as WHO, the Kenyan Government, UNICEF, and the Alive and Thrive network, and extensive previous qualitative work on the drivers of handwashing behaviours. Monthly visit modules were developed and pilot-tested to provide behavioural recommendations to mothers and other caregivers using key thematic constructs of convenience, nurturing care, and aspiration. We did a pilot randomised controlled trial[11] to test the feasibility and acceptability of all the interventions and to collect data that allowed us to optimise the ratio of community-based promoters to study participants. To identify and correct systematic problems with adherence, staff confirmed that intervention materials were delivered to all study participants at the outset of the trial, and collected monitoring data on availability of intervention materials and recommended behaviours during unannounced visits to a random sample of at least 20% of participants in intervention groups 2, 6, 10, and 19 months after the interventions began.

Community-based promoters for intervention and active control groups were nominated by study mothers and other mothers of children younger than 3 years in the community. A second promoter was added if there were more than ten participants (single groups) or more than eight participants (combined groups) in the cluster, giving a total of 1031 promoters. Promoters attended 2 days (active control), 6 days (single groups), or 7 days (combined groups) of initial training led by study staff on how to measure mid-upper-arm circumference, communication skills, intervention-specific behaviour change messages and intervention materials, and the information they were expected to report to IPA. Refresher trainings were done 6, 12, and 18 months after the initial training. At 2, 4, 9, 15, and 21 months, study staff met with promoters in their clusters to observe visits and offer supportive supervision. Study staff called promoters monthly to collect information on their activities, intervention adherence in the households they visited, referrals to health centres, and births or deaths of study children. Promoters received a branded T-shirt, a mobile phone, job aids and intervention materials, and compensation of approximately US$15 per month for the first 6 months when they had more intensive engagement with the study participants, and $9 per month thereafter (the prevailing daily wage for unskilled labour in the study area is $1–2). Promoters were instructed to visit all participants in their cluster monthly and measure the child's arm circumference or the pregnant mother's abdomen.

In intervention groups, promoters engaged study participants and other compound members through interactive activities such as guided discussions using visual aids, song, and storytelling; resupplied consumable intervention materials; encouraged consistent practice of targeted behaviours; and helped troubleshoot barriers to adherence, including problems with intervention hardware and behavioural barriers. Promoters were provided with detailed plans for every visit, including key messages, scripts for discussing visual aids, and instructions for activities that emphasised the learning objectives. Visits lasted about 10 min in the active control group and 45–60 min in intervention groups during the first year when the key messages were conveyed. In the second year, promoters reinforced messages to maintain habits. All groups used messages on themes of nurture, aspiration, and self-efficacy, particularly in the context of a new birth. Interventions used convenience and social norms to encourage target behaviours.

For **intervention-specific training materials** see https://osf.io/fs23x

In the three intervention groups that included water, promoters advocated treatment of drinking water with sodium hypochlorite. Chlorine dispensers for convenient water treatment at the point of collection were installed at an average of five communal water sources in the cluster and refilled as needed. Every 6 months, households in study compounds were given a 1 L bottle of chlorine for point-of-use water treatment in case households collected rainwater or used a source without a dispenser. Promoters used chlorine test strips during their regular visits to determine if the household was using chlorine, and negative results stimulated conversation about addressing barriers to chlorination.

In the three intervention groups that included sanitation, promoters advocated using latrines for defecation and safe disposal of children's and animals' faeces into a latrine. Existing unimproved latrines in study households were upgraded to improved latrines by installing a plastic slab, which also had a tight-fitting lid over the hole. New latrines were constructed for study households that did not have a latrine or whose latrine was unlikely to last for 2 years. All households in study compounds received a sani-scoop with a paddle as a dedicated faeces-removal tool. Finally, all households with children younger than 3 years in study compounds received plastic potties to facilitate toilet training and transfer of child faeces to the latrine.

In the three intervention groups that included handwashing, promoters advocated handwashing with soap before handling food and after defecation (including assisting a child). Study compounds were given two permanent, water-frugal handwashing stations intended to be installed near the food preparation area and the latrine. Handwashing stations were constructed of painted metal, with two foot-pedal-operated jerry-cans that dispensed a light flow of rinse water and soapy water. Promoters added chunks of bar soap to the soapy water container quarterly.

In the two intervention groups that included nutrition, a set of ten age-targeted modules were developed to enable promoters to advocate for best practices in maternal, infant, and young child feeding: recommendations for dietary diversity during pregnancy and lactation, early initiation of breastfeeding, exclusive breastfeeding until 6 months, introduction of appropriate and diverse complementary foods at 6 months, and continued breastfeeding through 24 months. Facilitators and barriers to behaviour change were elicited using formative research and health promoter guides were developed to address common barriers and questions. Study mothers with children between 6–24 months were provided with two 10 g sachets per day of a small quantity of lipid-based nutrient supplement (LNS; Nutriset; Malauny, France) that could be mixed into the child's food. LNS provided 118 kcal per day and 12 essential vitamins and ten minerals. Promoters explained that LNS was not to replace breastfeeding or complementary foods.

Promoters and intervention materials were introduced at community meetings roughly 6 weeks after enrolment. All interventions were delivered within 3 months of enrolment (appendix p 1). LNS was introduced to each child when they turned 6 months old. All handwashing stations and latrines were inspected within a month of construction, and a subset of households was periodically visited to observe group-specific indicators of intervention adherence. These data alerted study investigators to any

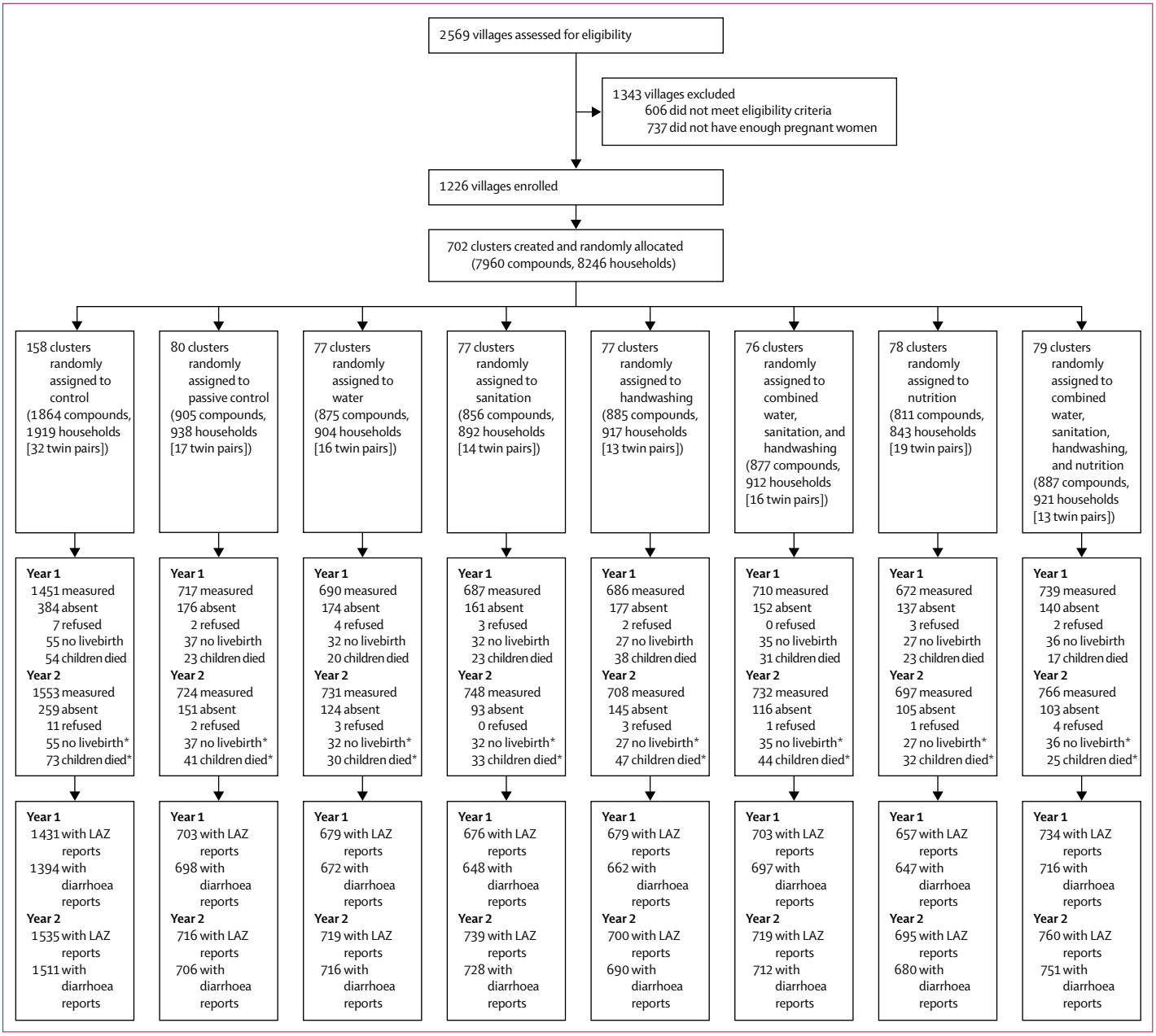


Figure 1: **Trial profile and analysis populations for primary outcomes**
LAZ=length-for-age *Z* scores. *Stillbirth and child death counts are cumulative.

issues with intervention implementation so they could be addressed consistently across all clusters and groups.

The enrolment survey included baseline demographics; assets; water, sanitation, and handwashing infrastructure; and target behaviours. Follow-up at 1 year and 2 years after intervention delivery consisted of an unannounced visit to study compounds to observe objective indicators of target behaviours (in all groups other than the passive control) and, on the following day, growth and health outcome measurements at a central location in the cluster (eg, a church or school).

Children identified as possibly malnourished (mid-upper-arm circumference <11·5 cm), either by the promoter during routine visits or by study staff during follow-up measurements, were referred to health facilities for treatment.

Outcomes

Adherence to the interventions was assessed using objective, observable indicators where possible (appendix pp 2, 3). 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. All child deaths reported by the health promoters were confirmed by a staff nurse who visited households. All outcomes were prespecified. Primary outcomes were caregiver-reported diarrhoea 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 and tertiary outcomes reported in this paper are 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; prevalence of stunting (length-for-age *Z* score less than –2), severe stunting (length-for-age *Z* score less than –3), wasting (weight-for-length *Z* score less than –2), and underweight (weight-for-age *Z* score less than –2); and all-cause mortality. We excluded children from *Z*-score analyses if their measurements were outside biologically plausible ranges following WHO recommendations. More details on exclusion criteria, measurement protocols, and outcome definitions are in the appendix (p 1).

Statistical analyses

Sample size calculations for the two primary outcomes were based on a minimum detectable effect of 0·15 in length-for-age *Z* score (intracluster correlation of 0·02 in our pilot study) and a relative risk of diarrhoea of 0·7 or smaller (assuming a 7-day prevalence of 12% in the active control group based on a pilot study to inform this trial) for a comparison of any intervention with the double-sized control group, assuming 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.[10,11] Sample size calculations indicated 80 clusters per group, each with ten children.

For more on the **updates to the analysis plan** see https://osf.io/7urqa/

Two biostatisticians, blinded to treatment assignment, independently replicated the analyses following the prespecified analysis plan with minor updates.[10] We analysed participants according to their randomised assignment (intention to treat), regardless of adherence to the intervention, using the active control group as the comparator. We used paired *t* tests for unadjusted length-for-age *Z* score comparisons and the Mantel-Haenszel prevalence ratio and difference for unadjusted diarrhoea and stunting comparisons, with randomisation block defining matched pairs or stratification. In secondary analyses, we estimated prevalence ratios and differences, adjusting for baseline covariates using targeted maximum likelihood estimation.[12] 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, which tested whether primary outcomes were the same in control households with more versus fewer households receiving interventions within a 2 km radius. In an analysis that was not prespecified, we tested for intervention effects on diarrhoea using only year 1 data.

The trial is registered at ClinicalTrials.gov, number NCT01704105. IPA convened a data and safety monitoring board.

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 the data in the study and had final responsibility for the decision to submit for publication.

Results

2569 villages were assessed for eligibility, of which 606 were excluded on the basis of village-level characteristics (primarily not meeting the study's rural criteria). 1226 villages were grouped into 702 clusters that had six or more pregnant women (figure 1). Between Nov 27, 2012, and May 21, 2014, 8246 pregnant women were enrolled in the study. 281 women did not have a livebirth and 140 women delivered twins. After at least three attempts to measure each child, 6659 (86%) of 7780 surviving children were measured at year 2, with diarrhoea reports for 6494 children and length-for-age *Z* score measures for 6583 children. Children were aged 2–18 months (median 12 months) at 1-year follow-up (January, 2014, to June, 2015) and aged 16–31 months (median 25 months) at 2-year follow-up (February, 2015, to July, 2016), but 11 184 (87%) of 12 841 children were in the target age ranges of 9–15 months at year 1 and 21–27 months at year 2 (appendix p 12).

Household characteristics were similar across groups at enrolment (table 1). Roughly three-quarters of participants collected drinking water from an improved source, but had to walk at least 10 min on average to the source. Over 80% of households owned a latrine, but less than 20% had access to an improved latrine. Less than 15% of households had soap available at a handwashing location. The prevalence of moderate-to-severe household hunger was 12% or lower.

Around 75% of households were visited by their promoter within the past month at year 1, but frequency of contact fell by year 2, with 40% or fewer households reporting a visit in the past month in each group (monitoring data suggest that most households were still visited at least every other month during the second year of the trial; see details in the appendix p 2, and table 2). Slightly less than half of households had detectable free chlorine in stored drinking water in the water group. Around 40% of drinking water samples tested in the water, sanitation, handwashing, and nutrition group had detectable free chlorine at year 1, which fell to around 20% by year 2. A high proportion of households (75%) had improved latrine access, which remained stable in year 1 and year 2 in households in the sanitation groups, increasing by more than 50% compared with the active control group. Reported safe disposal of children's faeces into a latrine fell by roughly half in all

	Active control (N=1919)	Passive control (N=938)	Water (N=904)	Sanitation (N=892)	Handwashing (N=917)	Water, sanitation, and handwashing (N=912)	Nutrition (N=843)	Water, sanitation, handwashing, and nutrition (N=921)
Maternal								
Age (years)	26 (6)	26 (7)	26 (6)	26 (7)	26 (6)	26 (6)	26 (6)	26 (6)
Completed at least primary education	916 (48%)	441 (47%)	447 (50%)	430 (48%)	402 (44%)	430 (47%)	409 (49%)	438 (48%)
Height (cm)	160 (6)	160 (7)	160 (6)	160 (6)	160 (6)	160 (6)	160 (7)	160 (7)
Study child is firstborn	490 (26%)	237 (25%)	205 (23%)	222 (25%)	208 (23%)	191 (21%)	206 (24%)	225 (25%)
Paternal								
Completed at least primary education	1098 (62%)	521 (60%)	532 (64%)	482 (58%)	500 (59%)	521 (61%)	491 (64%)	526 (62%)
Works in agriculture	749 (41%)	376 (43%)	378 (44%)	362 (43%)	363 (42%)	374 (43%)	343 (43%)	372 (43%)
Household								
Number of households per compound	2 (1)	2 (1)	2 (1)	2 (1)	2 (1)	2 (1)	2 (1)	2 (1)
Number of people per compound	8 (5)	8 (6)	8 (6)	8 (5)	8 (6)	8 (5)	8 (7)	8 (5)
Number of children <18 years in the household	3 (2)	3 (2)	3 (2)	3 (2)	3 (4)	3 (2)	3 (2)	3 (4)
Has electricity	122 (6%)	51 (5%)	60 (7%)	73 (8%)	67 (7%)	64 (7%)	58 (7%)	67 (7%)
Has a cement floor	107 (6%)	50 (5%)	71 (8%)	48 (5%)	41 (4%)	50 (5%)	48 (6%)	55 (6%)
Has an iron roof	1302 (68%)	600 (64%)	610 (68%)	587 (66%)	581 (63%)	574 (63%)	580 (69%)	615 (67%)
Owns a mobile phone	1526 (80%)	742 (79%)	705 (78%)	690 (77%)	722 (79%)	722 (79%)	685 (81%)	730 (79%)
Owns a motorcycle	185 (10%)	75 (8%)	81 (9%)	72 (8%)	91 (10%)	72 (8%)	81 (10%)	71 (8%)
Drinking water								
Primary drinking water source is improved*	1446 (76%)	699 (75%)	679 (75%)	675 (76%)	708 (78%)	624 (69%)	603 (72%)	697 (76%)
One-way walking time to primary water source (min)	11 (12)	12 (16)	12 (30)	10 (10)	11 (13)	11 (13)	11 (12)	11 (12)
Reported treating stored water	196 (13%)	92 (12%)	81 (11%)	94 (13%)	96 (13%)	97 (13%)	79 (12%)	106 (14%)
Sanitation								
Always or usually use primary toilet for defecation								
Men	1778 (95%)	867 (95%)	828 (94%)	810 (94%)	845 (95%)	851 (95%)	785 (95%)	854 (95%)
Women	1822 (96%)	898 (96%)	868 (96%)	840 (94%)	871 (96%)	877 (96%)	812 (96%)	872 (95%)
Daily defecating in the open								
Children aged 3 to <8 years	145 (12%)	87 (14%)	74 (13%)	68 (13%)	81 (14%)	75 (13%)	82 (15%)	75 (12%)
Children aged 0 to <3 years	789 (78%)	378 (77%)	376 (80%)	370 (75%)	358 (76%)	394 (77%)	363 (79%)	388 (78%)
Latrine								
Own any latrine	1561 (82%)	774 (83%)	750 (83%)	722 (81%)	756 (83%)	754 (83%)	701 (83%)	764 (83%)
Access to improved latrine	309 (17%)	153 (17%)	150 (18%)	131 (16%)	157 (19%)	153 (18%)	119 (15%)	143 (16%)
Human faeces observed in the compound	163 (9%)	79 (8%)	66 (7%)	72 (8%)	84 (9%)	73 (8%)	73 (9%)	87 (9%)
Handwashing location								
Has water within 2 m of handwashing location	487 (25%)	236 (25%)	242 (27%)	245 (28%)	245 (27%)	251 (28%)	228 (27%)	249 (27%)
Has soap within 2 m of handwashing location	164 (9%)	94 (10%)	91 (10%)	75 (8%)	83 (9%)	115 (13%)	90 (11%)	87 (9%)
Food security								
Prevalence of moderate-to-severe household hunger†	203 (11%)	113 (12%)	106 (12%)	91 (10%)	92 (10%)	101 (11%)	98 (12%)	104 (11%)

Data are n (%) or mean (SD). Percentages were calculated from smaller denominators than those shown at the top of the table for all variables because of missing values. *Defined by WHO UNICEF Joint Monitoring Program's definition for an improved water source. †Assessed by the Household Food Insecurity Access Scale.

Table 1: **Baseline characteristics by intervention group**

groups between year 1 and year 2, although the practice remained over twice as likely in the groups that included sanitation compared with other groups at year 1 and year 2. More than 75% of households in the intervention groups that included handwashing had water and soap available at a handwashing location at year 1, but this indicator also fell to about 20% by year 2. Adherence to LNS recommendations was high (≥95%) at year 1 and year 2, with children consuming a few more LNS sachets per month on average than would be expected at year 2. Across all indicators, adherence was comparable between the water, sanitation, and handwashing group and the water, sanitation, handwashing, and nutrition group compared with single intervention groups.

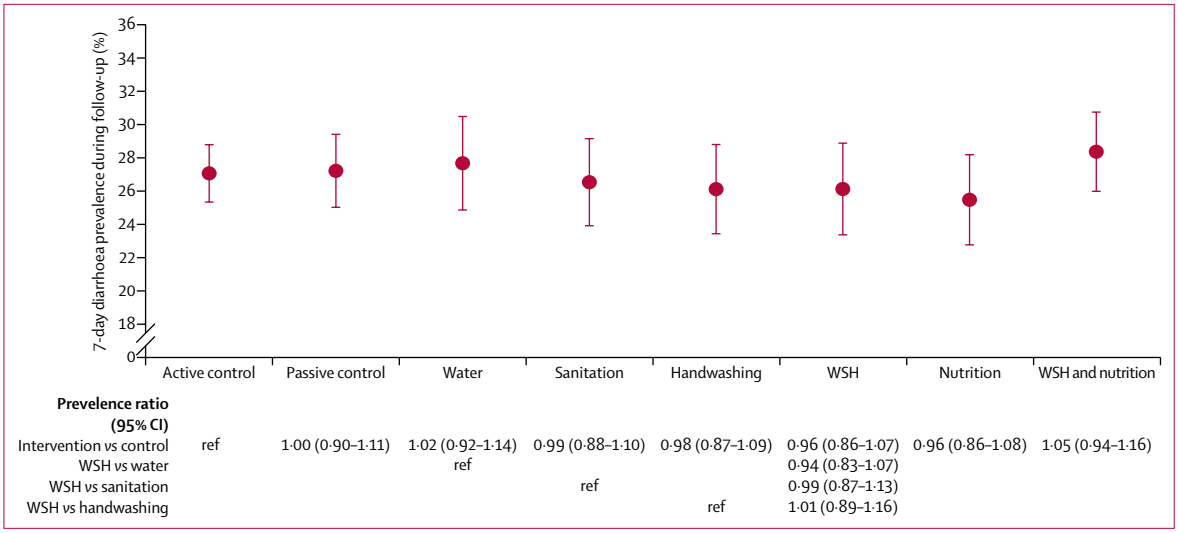
	Active Control (N=1919)	Passive Control (N=938)	Water (N=904)	Sanitation (N=892)	Handwashing (N=917)	Water, sanitation, and handwashing (N=912)	Nutrition (N=843)	Water, sanitation, handwashing, and nutrition (N=921)
Number of compounds assessed								
Enrolment	1913/1919 (100%)	936/938 (100%)	902/904 (100%)	890/892 (100%)	914/917 (100%)	912/912 (100%)	843/843 (100%)	918/921 (100%)
Year 1	1043/1919 (54%)	..	477/904 (53%)	473/892 (53%)	501/917 (55%)	536/912 (59%)	454/843(54%)	493/921 (54%)
Year 2	1458/1919 (76%)	..	696/904 (77%)	712/892 (80%)	690/917 (75%)	675/912 (74%)	650/843 (77%)	735/921 (100%)
Visited by promoter in past month								
Enrolment	..	..	..	..	..	..	..	..
Year 1	666/980 (68%)	..	338/445 (76%)	333/445 (75%)	333/480 (69%)	386/512 (75%)	344/433 (79%)	388/474 (82%)
Year 2	492/1412 (35%)	..	255/680 (37%)	278/692 (40%)	228/678 (34%)	241/649 (37%)	251/635 (40%)	259/710 (36%)
Stored drinking water has detectable free chlorine								
Enrolment	44/1529 (3%)	24/736 (3%)	20/720 (3%)	20/715 (3%)	30/743 (4%)	29/711 (4%)	14/661 (2%)	26/729 (4%)
Year 1	25/847 (3%)	..	151/385 (39%)	18/367 (5%)	20/417 (5%)	180/424 (42%)	9/392 (2%)	156/367 (43%)
Year 2	38/1365 (3%)	..	144/637 (23%)	17/641 (3%)	16/648 (2%)	112/598 (19%)	15/614 (2%)	128/652 (20%)
Access to improved latrine								
Enrolment	309/1788 (17%)	153/878 (17%)	150/844 (18%)	131/836 (16%)	157/847 (19%)	153/867 (18%)	119/794 (15%)	143/872 (16%)
Year 1	178/993 (18%)	..	74/461 (16%)	409/458 (89%)	65/486 (13%)	472/526 (90%)	63/424 (15%)	425/477 (89%)
Year 2	271/1381 (20%)	..	128/664 (19%)	534/683 (78%)	119/654 (18%)	529/644 (82%)	99/613 (16%)	561/706 (79%)
Child faeces safely disposed of								
Enrolment	114/721 (16%)	51/323 (16%)	53/310 (17%)	67/347 (19%)	54/319 (17%)	65/369 (18%)	33/310 (11%)	56/353 (16%)
Year 1	338/903 (37%)	..	158/424 (37%)	317/412 (77%)	157/431 (36%)	326/463 (70%)	155/391 (40%)	287/432 (66%)
Year 2	136/1320 (10%)	..	52/625 (8%)	240/643 (37%)	62/616 (10%)	205/597 (34%)	52/578 (9%)	219/657 (33%)
Handwashing location has water and soap								
Enrolment	96/1913 (5%)	58/936 (6%)	56/902 (6%)	42/890 (5%)	52/914 (6%)	64/912 (7%)	57/843 (7%)	53/918 (6%)
Year 1	124/1043 (12%)	..	53/477 (11%)	49/473 (10%)	381/501 (76%)	416/536 (78%)	61/454 (13%)	381/493 (77%)
Year 2	127/1458 (9%)	..	49/696 (7%)	57/712 (8%)	159/690 (23%)	130/675 (19%)	76/650 (12%)	152/735 (21%)
LNS sachets consumed (% expected)*								
Enrolment	..	..	..	..	..	..	..	..
Year 1	..	..	..	..	..	..	5264/5558 (95%)	5583/5838 (96%)
Year 2	..	..	..	..	..	..	3577/3136 (114%)	4028/3458 (116%)

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 reported proportion of 14 sachets consumed in the past week in index children aged 6–24 months.

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

Diarrhoea prevalence over the past 7 days (combining data from year 1 and year 2) was 27·1% in children in the active control group (figure 2, table 3). The intracluster correlation for diarrhoea was 0·012. Compared with the active control group, the diarrhoea prevalence ratios across all groups were not significantly different from one and differences were not significantly different from zero (figure 2, table 3). Diarrhoea prevalence was the same in the combined water, sanitation, and handwashing group and the individual water, sanitation, and handwashing groups. Although adherence to the water and handwashing interventions was higher in year 1 than in year 2, in an analysis that was not prespecified, diarrhoea prevalence was not significantly lower in any of the intervention groups at year 1 (appendix p 12). The high diarrhoea prevalence was fairly stable over 2 years of follow-up and there were no apparent seasonal trends (appendix p 13). Although we had prespecified a sensitivity analysis by age group of child at year 2, we did not complete this analysis because sample sizes in the age group strata were smaller than expected.

By year 2, when children were between 16 and 31 months old (median 25 months), mean length-for-age *Z* score in children in the active control group was –1·54 (SD 1·11; figure 3). The intracluster correlation for length-for-age *Z* score was 0·037. Compared with the active control group, only nutrition and combined water, sanitation, handwashing, and nutrition had higher length-for-age *Z* score (mean difference in score 0·13 [95% CI 0·01–0·25] for nutrition; 0·16 [0·05–0·27] for combined water, sanitation, handwashing, and nutrition; figure 3). Children in the combined water, sanitation, handwashing, and nutrition group were not significantly taller than children in the nutrition group (mean difference 0·04 [95% CI –0·11 to 0·19]; figure 3). Most length-for-age *Z* score gains in these two groups were already apparent by year 1 (0·11 [–0·01 to 0·22] for nutrition; 0·12 [0·01–0·22] for combined water, sanitation, handwashing, and nutrition; appendix p 14).



Prevelence ratio (95% CI)							
Intervention *vs* control	ref	1·00 (0·90–1·11)	1·02 (0·92–1·14)	0·99 (0·88–1·10)	0·98 (0·87–1·09)	0·96 (0·86–1·07)	0·96 (0·86–1·08)	1·05 (0·94–1·16)
WSH *vs* water			ref			0·94 (0·83–1·07)		
WSH *vs* sanitation				ref		0·99 (0·87–1·13)		
WSH *vs* handwashing					ref	1·01 (0·89–1·16)		

Figure 2: **Intervention effects on diarrhoea prevalence 1 and 2 years after intervention**
Data are mean (95% CI). ref=reference. WSH=water, sanitation, and handwashing.

Mean weight-for-age *Z* score at year 2 was higher in children in the nutrition and combined water, sanitation, handwashing, and nutrition groups than the mean of –0·72 (SD 1·01) in the active control group (table 4). Children in the active control group were close to WHO standards for weight-for-length *Z* score; however, weight-for-length *Z* score at year 2 was higher in the combined water, sanitation, handwashing, and nutrition group (table 4). There were no differences in mean head circumference for age *Z* score at year 2 between children in any of the intervention groups and those in the active control group. Results were similar at year 1, with the exception that differences in mean weight-for-length *Z* score between the active control and two groups with the nutrition intervention appear to have been numerically larger at year 1 (appendix p 15).

Compared with the active control group, a smaller proportion of children in the combined water, sanitation, handwashing, and nutrition group were stunted (too short for their age; –5·4 percentage points [95% CI –9·4 to –1·4]), severely stunted (–2·7 percentage points [–5·1 to –0·2]), or underweight (–3·0 percentage points [–5·4 to –0·6]; table 5); no other groups appeared to affect these outcomes. Notably, there were no significant differences between the combined water, sanitation, handwashing, and nutrition and nutrition groups for any growth outcomes. 1% of active control children were wasted and the proportions were similar across all groups.

Differences in growth outcomes between the active control and intervention groups were similar in magnitude and precision when estimated using adjusted models (appendix pp 16–19). We found no evidence of between-cluster spillover effects (appendix p 20).

The cumulative incidence of all-cause mortality was 3·9% in the active control and ranged from 5·3% in the handwashing group to 2·8% in the combined water, sanitation, handwashing, and nutrition group; none of the differences between intervention groups and the active control were statistically significant at α=0·05 (figure 1, appendix p 21).

	Mean* prevalence	Unadjusted† prevalence difference (95% CI)	Adjusted‡ prevalence difference (95% CI)
Intervention *vs* active control			
Active control	27·1%	..	..
Passive control	27·2%	–0·0 (–2·9 to 2·9)	–0·4 (–3·3 to 2·4)
Water	27·7%	0·7 (–2·3 to 3·6)	0·4 (–3·2 to 4·0)
Sanitation	26·5%	–0·3 (–3·3 to 2·6)	–0·3 (–3·2 to 2·6)
Handwashing	26·1%	–0·6 (–3·5 to 2·3)	–1·1 (–4·0 to 1·8)
Water, sanitation, and handwashing	26·1%	–1·2 (–4·1 to 1·7)	–1·1 (–4·3 to 2·0)
Nutrition	25·5%	–1·0 (–4·0 to 2·0)	–0·6 (–4·0 to 2·7)
Water, sanitation, handwashing, and nutrition	28·4%	1·2 (–1·7 to 4·1)	0·7 (–2·4 to 3·7)
Water, sanitation, and handwashing *vs* single groups			
Water, sanitation, and handwashing	26·1%	..	..
Water	27·7%	–1·6 (–5·1 to 1·9)	–2·1 (–6·0 to 1·8)
Sanitation	26·5%	–0·2 (–3·6 to 3·2)	–0·8 (–4·5 to 2·9)
Handwashing	26·1%	0·4 (–3·2 to 3·9)	0·5 (–3·6 to 4·5)

*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, mother's height, mothers education level, number of children <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 3: **Diarrhoea prevalence from 1 and 2 years (combined) after intervention**

Discussion

In the WASH Benefits cluster-randomised controlled trial, we found no effect of any interventions (improved

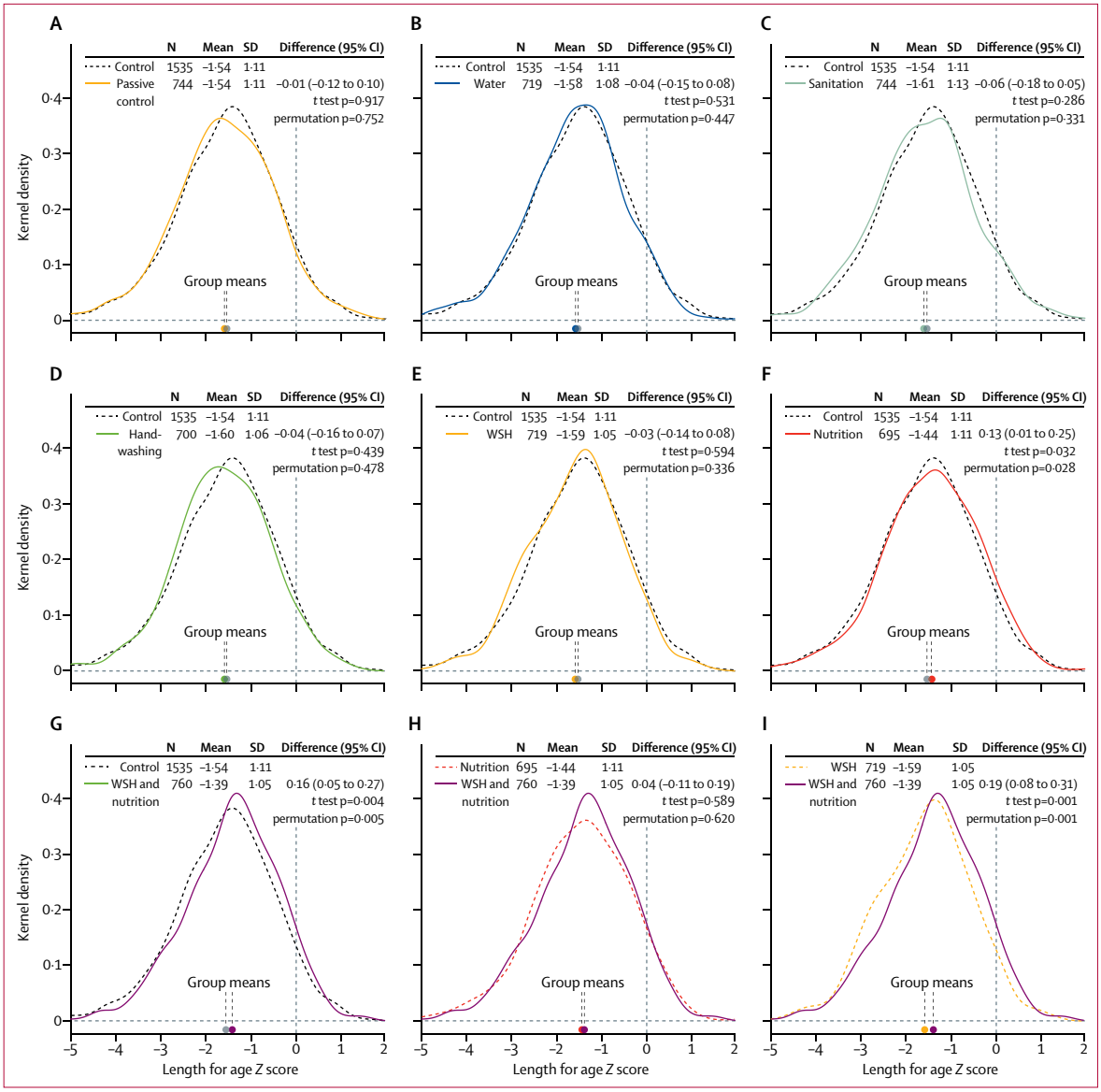


Figure 3: **Intervention effects on length-for-age Z scores in 6583 children after 2 years of intervention**
Kernel density plots show the distribution of length-for-age *Z* scores; dashed lines are the comparison group distribution and solid lines are the active comparator distribution. (A) Passive control *vs* active control. (B) Water *vs* active control. (C) Sanitation vs active control. (D) Handwashing *vs* active control. (E) WSH *vs* active control. (F) Nutrition *vs* active control. (G) WSH and nutrition *vs* active control. (H) WSH and nutrition *vs* nutrition. (I) WSH and nutrition *vs* WSH. p values for *t* test are for differences in group means from zero; permutation p values test the null hypothesis of no difference between groups using a Wilcoxon signed-rank test statistic. WSH=water, sanitation, and handwashing.

water quality, safe sanitation, handwashing, nutrition, or combinations of the interventions) on caregiver-reported diarrhoea prevalence during the first 2 years of life, and improvements in growth were only observed in groups including the nutrition intervention (maternal, infant, and young child feeding counselling and LNS distribution). With a large sample size and high-quality anthropometric measurements, this trial was powered to detect small effects in diarrhoea prevalence and length-for-age *Z* score had they been present. Lower adherence to the water and handwashing interventions by the end of the 2 years of intervention does not seem to be the only explanation for the absence of benefits: there were also no reductions in diarrhoea or improvements in growth in children in the water, handwashing, sanitation, or combined water, sanitation, and handwashing groups even in the first year (a typical measurement point in previous trials), when community-based promoters were most active and adherence was higher, whereas almost all of the growth benefits in the nutrition group and combined water, sanitation, handwashing, and nutrition group were already manifest in the first year. Adherence

	N	Mean (SD)	Difference *vs* active control (95% CI)	Difference *vs* nutrition (95% CI)	Difference *vs* water, sanitation, and handwashing (95% CI)
Weight-for-age *Z* score					
Active control	1548	−0·72 (1·01)	..	..	..
Passive control	721	−0·76 (0·97)	−0·04 (−0·13 to 0·05)	..	..
Water	727	−0·73 (1·00)	0·00 (−0·10 to 0·10)	..	..
Sanitation	747	−0·80 (1·05)	−0·07 (−0·19 to 0·04)	..	..
Handwashing	706	−0·77 (1·01)	−0·05 (−0·15 to 0·05)	..	..
Water, sanitation, and handwashing	725	−0·77 (0·98)	−0·02 (−0·12 to 0·08)	..	..
Nutrition	698	−0·65 (0·98)	0·11 (0·00 to 0·21)	..	..
Water, sanitation, handwashing, and nutrition	765	−0·60 (0·96)	0·14 (0·04 to 0·25)	0·04 (−0·07 to 0·15)	0·17 (0·05 to 0·30)
Weight-for-length *Z* score					
Active control	1536	0·11 (0·94)	..	..	..
Passive control	717	0·08 (0·92)	−0·04 (−0·13 to 0·05)	..	..
Water	719	0·14 (0·95)	0·04 (−0·06 to 0·13)	..	..
Sanitation	740	0·05 (0·97)	−0·05 (−0·14 to 0·05)	..	..
Handwashing	700	0·09 (0·93)	−0·02 (−0·11 to 0·06)	..	..
Water, sanitation, and handwashing	714	0·08 (0·92)	−0·02 (−0·10 to 0·07)	..	..
Nutrition	695	0·14 (0·92)	0·04 (−0·05 to 0·14)	..	..
Water, sanitation, handwashing, and nutrition	762	0·18 (0·90)	0·09 (0·00 to 0·19)	0·04 (−0·05 to 0·13)	0·12 (0·00 to 0·23)
Head circumference-for-age *Z* score					
Active control	1545	−0·27 (1·02)	..	..	..
Passive control	719	−0·27 (1·05)	0·00 (−0·10 to 0·10)	..	..
Water	727	−0·27 (1·03)	0·02 (−0·08 to 0·12)	..	..
Sanitation	745	−0·27 (1·04)	0·01 (−0·09 to 0·11)	..	..
Handwashing	705	−0·29 (0·99)	0·00 (−0·10 to 0·10)	..	..
Water, sanitation, and handwashing	729	−0·30 (0·96)	−0·03 (−0·12 to 0·06)	..	..
Nutrition	695	−0·23 (0·99)	0·05 (−0·05 to 0·15)	..	..
Water, sanitation, handwashing, and nutrition	763	−0·22 (0·99)	0·05 (−0·04 to 0·15)	−0·02 (−0·14 to 0·10)	0·08 (−0·05 to 0·20)

Median child age at 2-year follow-up was 2·05 years (IQR 1·93–2·16). All three secondary outcomes were prespecified.

Table 4: **Child growth *Z* scores at 2-year follow-up**

to the interventions was comparable to or better than what a government or large non-governmental organisation might hope to achieve at scale (appendix p 22), with increases in adherence indicators of 30 percentage points or higher in all intervention groups relative to the control in the first year.

These findings contrast with several systematic reviews[13–15] that have found significant protective benefits of water, sanitation, and hygiene interventions (including handwashing) on diarrhoea in efficacy trials, although most of these studies were shorter and had higher adherence. Results from other trials[16–18] also showed no effect of improved sanitation on diarrhoea, although differences in contexts and interventions complicate comparisons between these trials. Our trial differed from previous trials in that the intervention shifted households from unimproved sanitation (rather than open defecation) to improved sanitation. Additionally, the prevalence of diarrhoea in this study population was high, consistent with prevalence in 12–23-month-old infants measured in the 2014 Kenya Demographic and Health Survey.[19]

A systematic review and meta-analysis[20] of the effects of water quality and supply, sanitation, and hygiene interventions to improve growth identified only five randomised controlled trials of water or handwashing interventions, which did not suggest strong effects on growth, perhaps in part because the interventions lasted only 9–12 months. Since then, five more randomised trials of sanitation interventions have generated mixed evidence on child growth effects: two trials done in India and one in Indonesia had low adherence and no effect, and two done in settings with high rates of open defecation in India and Mali showed improvements in length-for-age *Z* score of 0·18–0·40 in children younger than 5 years.[16–18,20–22] The sanitation intervention in our trial was aligned with the focus on improved latrines initiated under the Millennium Development Goals, and the Sustainable Development Goals' recognition that children's faeces also need to be safely disposed of. This trial and its companion trial[9] in Bangladesh suggest that a compound-level approach to upgrading existing latrines and safely disposing of children's faeces is not sufficient

	n/N (%)	Difference *vs* active control (95% CI)	Difference *vs* nutrition (95% CI)	Difference *vs* water, sanitation, and handwashing (95% CI)
Stunting*				
Active control	483/1535 (31%)	..	..	..
Passive control	223/716 (31%)	−1·7 (−5·9 to 2·5)	..	..
Water	233/719 (32%)	0·1 (−4·2 to 4·3)	..	..
Sanitation	255/739 (35%)	2·3 (−2·0 to 6·6)	..	..
Handwashing	235/700 (34%)	0·8 (−3·5 to 5·1)	..	..
Water, sanitation, and handwashing	236/719 (33%)	1·3 (−3·0 to 5·6)	..	..
Nutrition	201/695 (29%)	−3·2 (−7·5 to 1·1)	..	..
Water, sanitation, handwashing, and nutrition	203/760 (27%)	−5·4 (−9·4 to −1·4)	−2·3 (−7·1 to 2·5)	−5·8 (−10·6 to −1·0)
Severe stunting†				
Active control	143/1535 (9%)	..	..	..
Passive control	62/716 (9%)	−0·8 (−3·3 to 1·8)	..	..
Water	69/719 (10%)	−0·5 (−3·2 to 2·2)	..	..
Sanitation	77/739 (10%)	1·0 (−1·8 to 3·7)	..	..
Handwashing	59/700 (8%)	−1·1 (−3·7 to 1·5)	..	..
Water, sanitation, and handwashing	65/719 (9%)	0·2 (−2·4 to 2·8)	..	..
Nutrition	55/695 (8%)	−1·6 (−4·2 to 1·0)	..	..
Water, sanitation, handwashing, and nutrition	55/760 (7%)	−2·7 (−5·1 to −0·2)	−0·9 (−3·7 to 2·0)	−2·7 (−5·6 to 0·2)
Wasting†				
Active control	22/1536 (1%)	..	..	..
Passive control	10/717 (1%)	0·0 (−1·1 to 1·1)	..	..
Water	9/719 (1%)	−0·2 (−1·3 to 0·8)	..	..
Sanitation	19/740 (3%)	1·1 (−0·3 to 2·4)	..	..
Handwashing	6/700 (1%)	−0·5 (−1·5 to 0·4)	..	..
Water, sanitation, and handwashing	10/714 (1%)	0·2 (−0·9 to 1·2)	..	..
Nutrition	8/695 (1%)	−0·3 (−1·3 to 0·8)	..	..
Water, sanitation, handwashing, and nutrition	11/762 (1%)	−0·1 (−1·2 to 1·0)	0·2 (−1·0 to 1·4)	0·0 (−1·2 to 1·1)
Underweight†				
Active control	148/1548 (10%)	..	..	..
Passive control	70/721 (10%)	−0·4 (−3·0 to 2·2)	..	..
Water	76/727 (10%)	−0·1 (−2·8 to 2·7)	..	..
Sanitation	87/747 (12%)	1·6 (−1·2 to 4·4)	..	..
Handwashing	71/706 (10%)	0·5 (−2·2 to 3·3)	..	..
Water, sanitation, and handwashing	72/725 (10%)	0·5 (−2·3 to 3·2)	..	..
Nutrition	59/698 (8%)	−1·2 (−3·9 to 1·5)	..	..
Water, sanitation, handwashing, and nutrition	52/765 (7%)	−3·0 (−5·4 to −0·6)	−1·8 (−4·7 to 1·1)	−3·3 (−6·2 to −0·5)

Median child age at 2-year follow-up was 2·05 years (IQR 1·93–2·16). *Prespecified secondary outcome. †Prespecified tertiary outcome.

Table 5: **Proportion of children stunted, severely stunted, wasted, and underweight at 2-year follow-up**

to improve child growth, and neither are water and handwashing interventions.

Conversely, counselling and LNS provided in the nutrition group improved length-for-age *Z* score by year 2. Compared with randomised controlled trials of LNS during complementary feeding, our finding of length-for-age *Z* score improvements of 0·13–0·16 in the nutrition groups falls in the middle of the spectrum between four trials: one from Malawi[23] that reported no effect on length-for-age *Z* score, one from Haiti[24] and one from Bangladesh[25] that reported an effect on length-for-age *Z* score comparable to this study, and one from Burkina Faso[26] that reported a larger effect on length-for-age *Z* score. Thus, there appears to be consistent evidence that LNS distribution together with some promotion of improved infant and young child feeding can reduce growth faltering, although this approach falls far short of eliminating the problem. Interventions will likely need to address the complex set of underlying determinants of growth faltering, including prenatal or preconception factors. Future analyses will explore changes in feeding practices that resulted from the intervention.

Although there were more improvements in anthropometric measures in the combined water, sanitation, handwashing, and nutrition group versus active control

than in the nutrition versus active control group, the differences were of little clinical or statistical significance. We conclude that combining nutrition with water, sanitation, and handwashing did not provide additional growth benefits beyond nutrition alone. Although the effect of water, sanitation, handwashing, and nutrition on mortality was not significant, the lower mortality in that group is consistent with the statistically significant effect of water, sanitation, handwashing, and nutrition on mortality in the Bangladesh trial.[9] Pending analyses will evaluate potential differences in effects on other child health outcomes.

It is possible that the water, sanitation, and handwashing interventions delivered in this trial did not sufficiently address important transmission routes for enteric pathogens.[11] Although the sanitation intervention included a sani-scoop and messages about preventing children from being exposed to domestic animal faeces, the emphasis was mostly on behaviours related to human faeces and might not have protected children from zoonotic pathogens.[27] Although chlorination of water has the advantage of providing residual protection against recontamination, it is not effective against protozoa such as *Giardia lamblia* and *Cryptosporidium* spp, the latter of which was identified as one of the most common causes of moderate-to-severe diarrhoea in children 0–23 months in a neighbouring part of Kenya.[28] Other limitations of this trial include the inability to mask the interventions; the absence of observable indicators of actual behaviour for the handwashing, sanitation, and nutrition interventions; lower adherence to the water and hygiene interventions during the second year of the trial than in the first year; and the use of a compound-level sanitation intervention, as opposed to community-level. Because masking was not possible, we focused on objective, observable indicators whenever possible rather than self-reported behaviours, recognising that the availability of a latrine or handwashing station stocked with water and soap does not necessarily imply that the materials were used. Despite an intensive design process that drew heavily on best practices in behaviour change, incorporation of lessons learned from the pilot randomised controlled trial, thorough verification of availability of the intervention materials, and periodic monitoring of indicators of recommended behaviours, adherence to the water and handwashing interventions appeared to reduce sharply in the last months of the trial. The waning intensity of promotion activities after a reduction in the stipend given to the health promoters could at least partly explain the drop in adherence. Finally, by contrast with water, handwashing, and nutrition interventions that directly benefit households that adhere to the intervention, a sanitation intervention in only a subset of compounds might not be sufficient to protect against exposure to faecal contamination in the environment that originates from other compounds in the community. We decided, however, to deliver compound-level interventions based on evidence that child exposure to enteric pathogens during the first 2 years of life occurs predominantly within the household compound.[29] Because environmental contamination and disease transmission pathways could be different in densely populated contexts, similar studies in urban areas would complement this rural trial.

Additional outcome measures collected in this trial will help to elucidate potential mechanisms for the observed effects, including indicators of environmental contamination, environmental enteric dysfunction, anaemia, enteric parasite infection, and child development. Molecular measurement of infections in the laboratory with stored stool specimens collected as part of this trial offer an opportunity for unbiased indicators of pathogen burden.

More intensive promotion and higher adherence could have resulted in larger effects than those reported, but our findings are relevant for large-scale programmes that struggle to achieve adherence rates as high as those of efficacy studies. The potential for water, sanitation, hygiene, and nutrition interventions to reduce diarrhoea and improve growth might be highly context-dependent. In our rural setting, water was plentiful but rarely available on premises, susceptible to contamination at the source and in storage, and rarely treated despite introduction of a nearly-universal filter distribution programme;[30] unimproved latrine coverage was high and there was a culture of using sanitation facilities for defecation by human beings, but there was probably persistent exposure to animal faeces; handwashing was not a common practice; breastfeeding was common, but exclusive breastfeeding was not, and most people had enough food, but not a diverse diet; diarrhoea prevalence was high; and many children had low length-for-age *Z* score, but not weight-for-length *Z* score. Our findings call into question the ability of large-scale water, sanitation, and handwashing interventions to reduce diarrhoea or improve growth. Our results suggest that integrated water, sanitation, and handwashing and nutrition programmes are no more effective than nutrition programmes at reducing diarrhoea or improving growth, and that nutritional interventions that include counselling and LNS can modestly reduce growth faltering, but fall short of eliminating it, even when LNS adherence is high.

Contributors

CN and CPS contributed equally to the manuscript. CN drafted the research protocol and manuscript with input from all listed coauthors and oversaw all aspects of the trial. CPS led the nutrition intervention and protocols for anthropometry data collection. KGD contributed to the nutrition intervention and interpretation of results. PK assisted with oversight of anthropometry data collection. CN, AJP, HND, TC, and PKR developed the water, sanitation, and handwashing intervention. CN, CPS, AJP, HND, and GN oversaw piloting and subsequent study implementation, contributed to refinements in interventions and measurements, and responded to threats to validity. CN, CPS, BFA, CDA, JB-C, AEH, AM, and JMC Jr developed the analytical approach, did the statistical analysis, and constructed the tables and figures. AL, SPL, and JMC Jr advised on harmonising the trials between Kenya

and Bangladesh and SMN helped adapt the trial to the Kenyan context. CN, CPS, AJP, BFA, JB-C, LCHF, AL, SPL, SMN, and JMC Jr secured funding for the trial. All authors have read, contributed to, and approved the final version of the manuscript.

Declaration of interests

All authors received funding for either salary or consulting fees through a grant from the Bill & Melinda Gates Foundation for this study.

Acknowledgments

We thank the study participants and promoters who participated in the trial, the fieldworkers who delivered the interventions and collected the data for the study, and the managers who ensured that everything ran smoothly. This research was financially supported in part by Global Development grant OPPGD759 from the Bill & Melinda Gates Foundation to the University of California, Berkeley, CA, USA, and grant AID-OAA-F-13-00040 from United States Agency for International Development (USAID) to Innovations for Poverty Action. This manuscript was made possible by the generous support of the American people through the USAID. The contents are the responsibility of the authors and do not necessarily reflect the views of USAID or the US Government.

For a **list of managers** see http://washbenefits.net

References

1 United National Children's Fund, WHO, World Bank Group. Levels and trends in child malnutrition. 2016. http://www.who.int/nutgrowthdb/jme_brochure2016.pdf?ua=1 (accessed March 18, 2017).
2 Sudfeld CR, McCoy DC, Danaei G, et al. Linear growth and child development in low- and middle-income countries: a meta-analysis. *Pediatrics* 2015; **135:** e1266–75.
3 Prendergast AJ, Humphrey JH. The stunting syndrome in developing countries. *Paediatr Int Child Health* 2014; **34:** 250–65.
4 Humphrey JH. Child undernutrition, tropical enteropathy, toilets, and handwashing. *Lancet* 2009; **374:** 1032–35.
5 Black RE, Victora CG, Walker SP, et al. Maternal and child undernutrition and overweight in low-income and middle-income countries. *Lancet* 2013; **382:** 427–51.
6 Scrimshaw NS, Taylor CE, Gordon JE. Interactions of nutrition and infection. *Monogr Ser World Health Organ* 1968; **57:** 3–329.
7 GBD 2015 Mortality and Causes of Death Collaborators. 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.
8 Richard SA, Black RE, Gilman RH, et al. Diarrhoea in early childhood: short-term association with weight and long-term association with length. *Am J Epidemiol* 2013; **178:** 1129–38.
9 Luby SP, Rahman M, Arnold BF, et al. Effects of water quality, sanitation, handwashing and nutritional interventions on diarrhoea and child growth in rural Bangladesh: a cluster randomised trial. *Lancet Glob Health* 2018; published online Jan 29. http://dx.doi.org/10.1016/S2214-109X(17)30490-4.
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 Christensen G, Dentz HN, Pickering AJ, et al. Pilot cluster randomized controlled trials to evaluate adoption of water, sanitation, and hygiene interventions and their combination in rural western Kenya. *Am J Trop Med Hyg* 2015; **92:** 437–47.
12 Balzer L, van der Laan M, Petersen M. Adaptive pre-specification in randomized trials with and without pair-matching. *Stat Med* 2016; **35:** 4528–45.
13 Wolf J, Prüss-Ustün A, Cumming O, et al. Assessing the impact of drinking water and sanitation on diarrhoeal disease in low- and middle-income settings: systematic review and meta-regression. *Trop Med Int Health* 2014; **19:** 928–42.
14 Clasen TF, Alexander KT, Sinclair D, et al. Interventions to improve water quality for preventing diarrhoea. *Cochrane Database Syst Rev* 2015; **10:** CD004794.
15 Ejemot-Nwadiaro RI, Ehiri JE, Arikpo D, Meremikwu MM, Critchley JA. Hand washing promotion for preventing diarrhoea. *Cochrane Database Syst Rev* 2015; **9:** CD004265.
16 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.
17 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.
18 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.
19 Kenya National Bureau of Statistics, Ministry of Health, National AIDS Control Council, Kenya Medical Research Institute, National Council for Population and Development, ICF International. Kenya Demographic and Health Survey 2014. Rockville, MD, USA: ICF International, 2015.
20 Dangour AD, Watson L, Cumming O, et al. Interventions to improve water quality and supply, sanitation and hygiene practices, and their effects on the nutritional status of children. *Cochrane Database Syst Rev* 2013; **8:** CD009382.
21 Cameron L, Shah M, Olivia S. Impact evaluation of a large-scale rural sanitation project in Indonesia. World Bank Policy Research Working Paper 6360. Washington, DC: World Bank, 2013.
22 Hammer J, Spears D. Village sanitation and children's human capital: evidence from a randomized experiment by the Maharashtra government. World Bank Policy Research Working Paper 6580. Washington, DC: World Bank, 2013.
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 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.
25 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.
26 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.
27 Harris A, Pickering AJ, Harris M, et al. Ruminants contribute fecal contamination to the urban household environment in Dhaka, Bangladesh. *Environ Sci Technol* 2016; **50:** 4642–49.
28 Kotloff KL, Nataro JP, Blackwelder WC, et al. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *Lancet* 2013; **382:** 209–22.
29 Ngure F, Reid BM, Humphrey JH, Mbuya MN, Pelto G, Stolzfus RJ. Water, sanitation, and hygiene (WASH), environmental enteropathy, nutrition, and early child development: making the links. *Ann NY Acad Sci* 2014; **1308:** 118–28.
30 Pickering AJ, Arnold BF, Dentz HN, Colford JM, Null C. Climate and health co-benefits in low-income countries: a case study of carbon-financed water filters in Kenya and a call for independent monitoring. *Environ Health Perspect* 2017; **125:** 278–83.

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.